Website:www.aigovernancereview.com

The report editor can be reached at

globalaigovernance@gmail.com

We welcome any comments on this report and any communication related to AI governance.

SHANGHAI INSTITUTE FOR SCIENCE OF SCIENCE

weisite：www.siss.sh.cn

mailbox：siss@siss.sh.cn

AI GOVERNANCE IN 2020
A YEAR IN REVIEW

OBSERVATIONS OF 52 GLOBAL EXPERTS

June, 2021

Shanghai Institute for Science of Science

Preprint

WHEN PEOPLE PULL TOGETHER,
NOTHING IS TOO HEAVY
TO BE LIFTED.

—— *BAO PUZI*

眾力并，
則萬鈞不足舉也。

——《抱朴子》

TABLE OF CONTENTS

FOREWORD

Although full of challenges, the year of 2020 was marvelous for Artificial Intelligence (AI).

The outbreak and spread of the COVID-19 pandemic significantly impacted the economic and social development of countries around the world. Policymakers and researchers around the world had to hurriedly put aside all their plans and spared no effort to deal with this disruptive new issue. As we see, AI governance is a sub-topic under this major issue.

In fact, over the past year when all work seemed in stagnation, research on AI governance went the opposite way; it drew more attention and stimulated deeper discussions as a result of the thorny issues brought about by the pandemic. For example, the application of digital tracking technology raised extensive discussions in many countries. In China, “Health Code” (a digital tracking technology) has been widely applied and praised as a powerful tool for pandemic prevention and control. In contrast, some countries hold a complicated and even negative attitude towards digital tracking technology even at the height of the pandemic.

This, at the very least, reminds us that AI governance, though with considerable controversy, concerns human destiny.

The year of 2020 was also quite important for Shanghai Institute for Science of Science as it marked our 40th anniversary. We had planned a series of celebration activities but all were cancelled, or just held on a small scale due to the pandemic. The Governance Forum of World AI Conference (Shanghai) 2020, organized by SISS, also had to change its organizational form to a combination of online and offline activities.

Over the 40-year development of Shanghai Institute for Science of Science, we have always believed that learning from international experience can effectively help our country to develop science and technology policies, and sometimes can even realize twice the results with half the work. We have been, in a variety of ways, constantly introducing international research results about the law of scientific development, the relation between science and economy, the social impacts of science and related policies. The knowledge introduced has made a notable contribution to the development of science and technology policies of China. In recent years, we have also been paying attention to the progress of countries in the world in terms of AI governance, hoping to learn, experience, and enhance understanding with such an effort.

Of course, learning should not be one-way, but mutual. In face of such challenges as the COVID-19 pandemic, and AI’s influence on all human beings means everyone’s thoughts matter. Mutual learning is not only about developing countries learning experience from developed countries, but also about developed countries learning about the thoughts of developing countries. The mutual understanding based on mutual learning is critical to the eventual establishment of an effective global consensus.

Last year, we worked with global experts for the first time to jointly compile the *AI Governance in 2019: A Year in Review*. We received extensive support from the experts we invited, which made us realize that such work could be more meaningful than we had anticipated. After the release of the report, we surprisingly found that more cooperation efforts had been formed based on the report. We believe this report itself has become a platform promoting mutual learning of all parties concerned.

We hope that the *AI Governance in 2020: A Year in Review* can give further play to the role as a platform for global exchanges, and record the extraordinary thoughts and actions in this extraordinary year worth remembering.

SHI Qian
Director of Shanghai Institute for Science of Science

EDITOR-IN-CHIEF: SHI QIAN

SHI Qian is the director of Shanghai Institute for Science of Science (SISS). Before joining SISS, Professor SHI was the vice president of the Shanghai Academy of Sciences & Technology and concurrently the vice president of the Shanghai Institute of Industrial Technology. He has been long engaged in the general planning for science and technology development, research project management, innovation platform building, and services for innovation and entrepreneurship. Professor SHI participated in the formulation of a number of national industrial development plans and the implementation of major national science and technology projects, where he presided over several soft science research projects, such as “Research on Shanghai’s Medium and Long-Term (2021—2035) Developmental Strategy of Science and Technology” from the Shanghai Municipal Government. Professor SHI obtained the Shanghai Special Award for Scientific and Technological Progress in 2016. He is also the director of the Technology Foresight Committee of the Chinese Association for Science of Science and S&T Policy, and the deputy director of the Expert Advisory Committee of the National New-Generation AI Innovation and Development Pilot Zone in Shanghai.

INTRODUCTION

Last year, we compiled our first annual report on AI governance. The purpose was to identify critical progress from numerous AI governance studies. We were pleasantly surprised by the enthusiastic responses to our invitations, resulting in 50 expert contributions to our report. The positive feedback from various individuals and organizations on the final publication encouraged us to continue the initiative in the future. Notable contributions include the recommendation of the Montreal AI Ethics Institute and a letter from the senior advisor at the Office of the President of the United Nations General Assembly. We hope that this report can improve understanding of - and help to bridge - different viewpoints on the challenges and opportunities of AI governance. That is the reason we compiled the report this year.

2020 will leave a deep mark in human history, as the outbreak and spread of the COVID-19 pandemic strained the economic and social development of the whole world. We once even expected that little progress on global AI governance would be made in 2020. However, there was still significant interest in continuing the annual report. Compared with last year, the number of participating authors (and institutions) turned out to be a little higher this year, as 52 experts (from 47 institutions) provided contributions.

As some authors have worked on this report for two consecutive years, they have been able to build on their work from the first year. Take OpenAI as an example: while its release plan for GPT-2 in 2019 sparked some controversies, the new release plan they proposed at the launch of GPT-3 in 2020 seems better received. The European Union is another good example: following its AI Ethical Framework released in 2019, it issued a White Paper in 2020, proposing corresponding regulatory rules.

The unusual situation created by the pandemic has also resulted in serious reflection on AI and its governance. Being compelled to reflect on AI may provide us with new ideas for future exploration. Seán Ó hÉigeartaigh from the University of Cambridge has been delving into whether AI deserves its hype or whether attention should be focused on the basics of the problem, like investments in public health during the pandemic. Other experts question whether the numerous existing AI governance studies can be effectively translated into policies for dealing with the pandemic.

By deciding to put together this global observation report this year, we were also able to invite experts and institutions who were not involved in the previous year. After the previous report was released, it was pointed out that the voice of the Global South, especially Latin America and Africa, had been neglected. As a result, an effort was made to include experts from Latin America and Africa, to reflect the concerns and the work done in AI governance there.

Brief introductions to the opinions of the participating experts are as follows.

Technical Communities

Technical experts have a prominent place in AI governance. This year, Turing Award winner Prof. John Hopcroft continues to offer his opinions as a scientist. He has identified 7 issues in various areas and highlighted the importance of “oversight”.

Bart Selman, President of AAAI, reminded us of an important issue: AI technology often operates in a manner that is quite foreign to us. This makes good governance dependent on close collaboration between AI researchers and policymakers.

GONG Ke, President of the World Federation of Engineering Organizations (WFEO), explained how WFEO is proactively promoting the efforts related to the ethical governance of AI. He also highlighted the concept of the green development of AI.

The work from OpenMined is fascinating this year; they introduced a new privacy-enhancing technology following secure computation, federated learning and differential privacy: structured transparency.

Prof. ZHANG Jianwei from the University of Hamburg described the role of AI against the backdrop of the pandemic.

Scientists are also considering the political and cultural issues raised by technologies. Prof. Alex Pentland from MIT detailed an emerging issue: how to govern digital platforms interoperating across sovereign and institutional borders.

Brian Christian, a bestselling author of books about science and humanities, including “The Alignment Problem”, raised a topic fundamental to society: how to achieve “alignment” within and between organizations.

Roman V. Yampolskiy, who specializes in AI safety, brought up the concern that AI might be ungovernable.

The Industrial Community

As a large international consulting firm, PwC has the opportunity to observe the AI dynamics of the international industrial community. Anand S. Rao, a partner at the firm, mentioned an interesting phenomenon: AI remains popular with the industrial community against the backdrop of the pandemic, with the global venture capital funding for AI rising continuously in 2020. However, very few companies had fully embedded and automated AI risk management and controls in place.

Abhishek Gupta, Founder of the Montreal AI Ethics Institute, arrived at deep insights on the reason why AI governance principles cannot be effectively implemented, based on his corporate background. Therefore, he proposed the “patterns of practice” feasible for practitioners.

Irene Solaiman from OpenAI, an organization with global influence, explained the release plan of GPT-3. Having learned from the launch of GPT-2, OpenAI adopted the new method of releasing GPT-3 through an API, which can be accessed by approved users with an API key.

YANG Fan, Co-founder of SenseTime, announced that as a world-leading AI company, SenseTime is making real efforts at “sustainable development”.

As an observer of the societal impacts of AI, Danil Kerimi voiced the opinion that AI promotes the rewriting of the social contract.

Steven Hoffman, venture investor in the Silicon Valley, discussed specific problems about data. In his opinion, commoditizing the data might not be the optimal solution, and we should focus on curbing abuses.

Omar Costilla-Reyes, an expert in smart medicine, elaborated on how social institutions can adapt to the development of AI, and how to adopt new certification methods in medicine.

The Interdisciplinary Research Community

Allan Dafoe and Alexis Carlier from the University of Oxford mentioned the implementation mechanisms for AI governance. For example, a leading national AI conference now requires that all paper submissions include a responsibility statement.

Jared Brown from the Future of Life Institute touched on specific issues: what the AI risks are, and how to identify, assess and manage them.

Sociologists Petra Ahrweiler and Martin Neumann mentioned that the formulation of regulations on AI governance requires an inventory of knowledge corpora about human values which are to be implemented with AI technologies.

Jurist Malavika Jayaram emphasized that the movement to decolonize data should include efforts aimed at letting every region work hard to preserve the sovereignty and autonomy of data that does not fit neatly into Western parameters.

Nathalie Smuha, another jurist, asserted that governance principles should be translated into enforceable legislation.

Wendell Wallach, a well-known expert in the ethics of science and technology, evaluated the use of AI in the unique scenario created by the pandemic, as well as in the complex geopolitical situation. In his opinion, global cooperation is essential.

Marie-Therese Png, who is active in encouraging the discourse of developing countries in global AI governance, worries that the development of AI may bring about a new round of colonization.

Logician Markus Knauff offered some opinions on the future development of AI from the perspective of cognitive psychology.

For developing countries, both the awareness of AI governance and their regulatory capacity should be taken into consideration. Urvashi Aneja highlighted how low levels of regulatory and institutional capacity pose further challenges to the suitability of risk-based approaches.

International Organizations

Irakli Beridze, Head of the Centre for AI and Robotics of the United Nations Interregional Crime and Justice Research Institute (UNICRI), showcased a gratifying improvement in the operationalization of AI governance: the UNICRI Centre for AI and Robotics, together with INTERPOL's Innovation Centre, have undertaken to develop an operationally oriented toolkit for the responsible use of AI by law enforcement. This toolkit is intended to support and guide the design, development and deployment of AI in a responsible manner.

Danit Gal, former Technology Advisor at the United Nations, who led work on AI in the implementation of the United Nations Secretary-General's Roadmap for Digital Cooperation, introduced the work she participated in. Her comprehensive vision urges global AI governance cooperation initiatives to equitably engage the Global South and underrepresented communities.

Seán Ó hÉigeartaigh from the University of Cambridge is part of the Global Partnership on AI's AI and Pandemics working group. His contribution argues that while the hype about AI is pervasive, it can provide limited solutions to the problem of the pandemic. We should instead focus on what is important - investments on public health. In addition, the debate about whether to use digital tracing technologies to deal with the pandemic has reminded us of the complexity of governance issues.

Cyrus Hodes, an observer of the international governance of AI, summarized the efforts made by representative organizations in the past year, when the pandemic shook the world. He reiterated the points mentioned by other contributors and emphasized that firm actions are critical following the formulation of principles.

Countries and Regions

Eugenio Vargas Garcia, Former Senior Adviser of the Office of the President of the United Nations General Assembly, recognizes that AI is important enough that international cooperation is required to jointly address related issues. He reviewed the relevant progress of UNESCO and the UN Secretary-General Roadmap for Digital Cooperation on AI governance. Eugenio Vargas Garcia has been calling for increased representation from the Global South in international discussions on AI governance in recent years.

Europe is an active promoter of AI governance. Eva Kaili, Member of the European Parliament, highlighted the efforts of Europe in determining the global leaders in AI governance, and emphasized the AI governance model of Europe where the regulator sets the principles and the market applies the principles by defining the standards of the product or service.

As the Coordinator of the European Commission's

High-Level Expert Group on Artificial Intelligence, Charlotte Stix analyzed how the European Union continued to methodically advance its AI governance framework in the year of the pandemic. She introduced the White Paper on AI: a European Approach to Excellence and Trust, published by the EU this year, which focused on the seven key requirements for trustworthy AI mentioned in the regulatory proposal.

Caroline Jeanmaire from UC Berkeley introduced the nine principles for AI design, development and application issued by President Trump near the end of his presidency, and presented the common concerns of America and Europe across the Atlantic: to ensure the control over AI risks.

Arisa Ema from Japan presented the joint statement of the 2nd French-German-Japanese AI Symposium held in 2020. While the first conference emphasized a human-centric approach, this year, in the face of the pandemic, the joint statement emphasized the importance of cooperation in addressing problems on a planetary scale.

As a strong power in information technology, India also made gratifying progress on AI governance in 2020. Raj Shekhar cited the discussions in India on all aspects of personal data protection and responsible AI.

Poon King Wang at the Lee Kuan Yew Centre for Innovative Cities at the Singapore University of Technology and Design (SUTD) proposed the “cross-sector GPS” initiative to help the Singaporean Government to address employment issues in the age of AI-an excellent case of transforming academic discussions into specific policy initiatives.

Working across the African continent, Victor Famubode unveiled the gratifying progress in awakening African governments to the ethical implications of AI.

Despite having a population of 600 million, Latin America has been under-represented in international discussions on AI. Thus, several experts from South America were invited this year, including Olga Cavalli from Argentina, Edson Prestes from Brazil, Constanza Gómez Mont from Columbia, Jean García Periche from Dominican Republic, and José A. Guridi Bustos from Chile. They discussed the AI governance progress of Latin America from their own point of view respectively. They share a common view that Latin America should have its own voice in global AI governance. They also agreed unanimously that Latin America, with its relatively low level of AI technology, R&D, and application, should not blindly adopt the governance models of developed countries.

Countries and Regions (China)

AI is developing rapidly in China. FU Ying, Chairperson for International Security and Strategy, Tsinghua University, talked about the impact of AI on international security, and noted that China is willing to dialogue and cooperate with all parties.

ZHAO Zhiyun, Director of the Advance Office of Development Planning for New-generation AI, MOST, illustrated how against the backdrop of the pandemic, China is steadily promoting the further implementation of the governance principles issued in 2019.

Professor SU Jun from Tsinghua University is taking the lead in promoting a comprehensive social experiment, not only to make an overall experimental evaluation of the social impacts of artificial intelligence, but most importantly to lay a theoretical foundation for China to “build an intelligent society with humanism”.

LI Xiuquan from Chinese Academy of Science and Technology for Development (CASTED) talked about the considerations on promoting AI governance through both technical and institutional innovation.

Professor WANG Guoyu from Fudan University introduced the AI governance efforts promoted by Chinese computer experts and philosophers through organizations such as China Computer Federation (CCF) in 2020.

The World Artificial Intelligence Conference organized by Shanghai has wide influence among the peers across the world. WANG Yingchun from Shanghai Institute for Science of Science presented the relevant information about WAIC 2020 - AI Governance Forum.

Based on the annual observations of 2019, it is evident that the global AI governance system is taking shape. The annual observations of 2020 show that this is ongoing but with the added feature of 2020: the reflection on the global AI governance system.

Based on the reports of more than 50 experts, it is clear that the industrial and policy research communities, as well as the international organizations, regions, and countries have made progress on AI governance in 2020. While it is possible to identify these emerging trends, it is impossible to present all of the progress made. The annual report summarized here is primarily intended to provide a springboard for further conversations and discussions.

EXECUTIVE EDITORS: LI HUI; BRIAN TSE (INVITED)

LI Hui is an associate professor at the Shanghai Institute for Science of Science. He regularly participates in the formulation of AI strategies for Shanghai as well as on a national level. He also frequently publishes his views on AI governance in major Chinese media such as *People's Daily, Guangming Daily* and *Wenhui Daily*. He has played a prominent role in organizing the Governance Forum of the World Artificial Intelligence Conference 2019, 2020 and 2021. He earned his PhD in history of science from Shanghai Jiao Tong University in 2011. His background led to his research interests on issues related to AI governance with a long-term perspective and global thinking.

Brian Tse focuses on researching and improving cooperation over AI safety, governance, and stability. Brian is a Policy Affiliate at the University of Oxford's Center for the Governance of AI, Coordinator at the Beijing AI Academy's AI4SDGs Cooperation Network, and Senior Advisor at the Partnership on AI. He has advised organizations such as Google DeepMind, OpenAI, Baidu, Tencent's WeBank, Carnegie Endowment for International Peace, and Tsinghua University's World Peace Forum. He is a member of the AI safety program committee at AAAI & IJCAI, IEEE P2894 XAI Working Group, and IEEE P2863 Organizational Governance of AI Working Group.

ACKNOWLEDGEMENT

We appreciate the recognition and support of all the authors involved. We appreciate the participation of 50 experts in *AI Governance in 2019: A Year in Review*, who enabled the "birth" of our report, and the participation of 52 experts in *AI Governance in 2020: A Year in Review*, who enabled the continuation of our report. We appreciate the joint efforts made by 80 experts from all over the world in the past two years, which have laid a foundation for the serial production of this report. It is our belief that the report will truly become an annual tradition for scholarly exchange on topics of AI governance.

We also appreciate all the work that the authors have done. Although the report is released under the name of Shanghai Institute for Science of Science, it is actually a result of the partnership among all authors. In addition, one of the features of this report is that the authors of the report are also its editors. The formal editing team made few modifications to the articles, meaning that the authors were able to check both the contents and the grammars and expressions of their contributions. Many of the authors are actually the organizers of the report, helping to invite new applicable authors for their engagement. For example, all the South American authors of the 2020 report were referred to us by Mr. Eugenio Vargas Garcia from the United Nations.

We would like to thank all our colleagues at Shanghai Institute for Science of Science for their support. Although the report is compiled by the team led by the Director of SISS in person, the leading team, the research management department, and the administrative department of Shanghai Institute for Science of Science also provided great and solid support in setup, funding, promotion of the project in the later stage, and other aspects.

We would like to thank our sponsors. The report would not be completed without the financial support from Shanghai Institute for Science of Science. After the publication of the 2019 report, we were honored to receive recognition and inquiries from some foundations. Among other things, we fortunately have received support from Jaan Tallinn, who provided funding for the production of the 2020 report.

We would also like to thank our volunteers. who were attracted by the reports last year to join us. Caroline Jeanmaire from University of California, Berkeley helped us polish the framework and the introduction of the report. Irene Solaiman from OpenAI gave us suggestions and helped us with some editing work. Herbert Chia (Sequoia Capital China) helped us contact with some contributors.

XU Nuo from Shanghai Institute for Science of Science, provided support in polishing the translation to aid better understanding for readers from different language environments. CHEN Yakun from the Chinese Academy of Social Sciences, and HU Xiaomeng from Hunan Normal University also contributed lots of advice and suggestions.

NIE Yunzhe from Columbia University in the City of New York contacted the experts as the project leader of the report, making many painstaking efforts.

We also extend our appreciation to three individuals, who have devoted their time and efforts to review, edit and polish the entire report prior to its publication: JI Caixuan, PhD student in Area Studies from the University of Oxford; Dr. ZHANG Kai, PhD in Materials from the University of Oxford; and LI Hailong, MSc in Food Safety and Management from the University of Birmingham.

We appreciate the support of all our partners, peers and friends from all over the world! We are so honored to work with them all for such a significant task.

Part I Technical Perspectives from World-class Scientists

Issues on AI Governance

By John E. Hopcroft

AI will deliver great benefits and be a significant component of our environment. Oversight is important to ensure both responsible use of AI by organizations and faster economic growth. A number of issues are:

1. Update the legal system

For example, who will be responsible when a driverless vehicle is in an accident? Is it the owner, the manufacturer, or the developer of the AI system? Updating the legal system is important for companies to make investments.

2. Clarify data ownership

When we do a web search, the search company saves all searches by IP address. From this database, one can extract who we are, where we shop, what items we purchase. Looking at all searches from a single IP address also provides information about our gender, our hobbies, and other personal information. Will we allow search companies to sell this information?

3. Fairness

The impact of AI needs to be regulated so that all classes of people benefit, not just the rich and powerful.

4. Eliminating bias

AI systems are trained using data and bias in the data will impact the AI system. For example, if high level jobs are primarily held by men, an AI system is likely to primarily recommend men for high level jobs.

5. Explanability

Many AI system are black boxes and provide answers to queries but no explanation for the answers. If a system turns down an individual for a job, the individual may insist on a reason for the denial. Was it based on his qualifications or some other factor?

6. Face and affect recognition

Face and affect recognition has many important applications but it should not be used for some matters such as racial or ethnic origin, personality traits, or mental health.

7. Social network communication

Individuals use companies like Facebook and Twitter to communicate information or disinformation to large numbers of users. This can have positive or negative impact on a nation. AI may be used to control what is communicated and will require AI governance.

These are a few of the issues that need to be thought about and regulated. Other issues concern safety in human AI interactions, security, ethical situations, etc. One also needs to consider at what level issues should be considered. Which issues are the prerogative of government and which should be developed by industry? The creation of an AI governance policy needs to include government, academy, and industry representatives.

ABOUT THE AUTHOR

John E. Hopcroft

John E. Hopcroft is the IBM Professor of Engineering and Applied Mathematics in Computer Science at Cornell University. From January 1994 until June 2001, he was the Joseph Silbert Dean of Engineering. After receiving both his M.S. (1962) and Ph.D. (1964) in electrical engineering from Stanford University, he spent three years on the faculty of Princeton University. He joined the Cornell faculty in 1967, was named professor in 1972 and the Joseph C. Ford Professor of Computer Science in 1985. He served as chairman of the Department of Computer Science from 1987 to 1992 and was the associate dean for college affairs in 1993. An undergraduate alumnus of Seattle University, Hopcroft was honored with a Doctor of Humanities Degree, Honoris Causa, in 1990.

Hopcroft's research centers on theoretical aspects of computing, especially analysis of algorithms, automata theory, and graph algorithms. He has coauthored four books on formal languages and algorithms with Jeffrey D. Ullman and Alfred V. Aho. His most recent work is on the study of information capture and access.

He was honored with the A. M. Turing Award in 1986. He is a member of the National Academy of Sciences (NAS), the National Academy of Engineering (NAE), a foreign member of the Chinese Academy of Sciences, and a fellow of the American Academy of Arts and Sciences (AAAS), the American Association for the Advancement of Science, the Institute of Electrical and Electronics Engineers (IEEE), and the Association of Computing Machinery (ACM). In 1992, he was appointed by President Bush to the National Science Board (NSB), which oversees the National Science Foundation (NSF), and served through May 1998. From 1995-98, Hopcroft served on the National Research Council's Commission on Physical Sciences, Mathematics, and Applications.

In addition to these appointments, Hopcroft serves as a member of the SIAM financial management committee, IIIT New Delhi advisory board, Microsoft's technical advisory board for research Asia, and the Engineering Advisory Board, Seattle University.

Understanding AI for Governance

By Bart Selman

It is exciting to see the recent significant level of activity around governance for AI. Good AI governance is vital for a proper transition of AI technology into our society. The goal is to keep human interests central and ensure that society truly benefits from AI technology.

In this statement, I would like to highlight a key challenge for developing proper AI governance. Good AI governance requires a deep understanding of both the opportunities as well as the limitations of AI methods. AI systems are starting to match or even exceed human performance on a range of cognitive tasks. However, it is essential to realize that these capabilities are often achieved in ways that are very different from how the human mind handles these tasks. A fundamental difficulty is that when we consider AI technologies, we tend to anthropomorphize the systems. In other words, we assume the systems perform cognitive tasks in a manner similar to ourselves.

Let me give an example. Machine translation based on deep learning approaches has made incredible advances in recent years. Current models provide reasonable translations between dozens of language pairs. However, rather counterintuitively, these translations are obtained without any real understanding of the text that is being translated. It is difficult for us even to imagine that one could translate between two languages without understanding either language. Nevertheless, this is what current language translation systems do — they operate in an almost alien way.

This “alien” mode of operation is also apparent in highly complex deep models trained for all types of data interpretation and decision-making tasks. It may seem reasonable to require the AI system to explain its decisions (i.e., “the right to an explanation”). However, researchers are discovering that attempts to provide explanations can easily lead to pseudo-explanations that may satisfy the human user but do not accurately reflect the system's internal decision-making process.

In general, we are seeing emerging AI systems that operate without the rich context of human knowledge and commonsense that we take for granted. So, for example, in case of a medical emergency, we might instruct our self-driving car to ignore some traffic rules to take us as quickly as possible to the nearest hospital. A human driver would realize that this should be done without putting others in danger. However, this is part of our commonsense but would need to be explicitly coded into a self-driving vehicle. To develop proper AI governance, we need to consider that AI technology often operates in a manner that is quite foreign to us. Good governance will therefore require close collaboration between AI researchers and policymakers.

ABOUT THE AUTHOR

By Bart Selman

Bart Selman is the Joseph C. Ford Professor of Engineering and Computer Science at Cornell University. Prof. Selman is the President of the Association for the Advancement of Artificial Intelligence (AAAI), the main international professional society for AI researchers and practitioners. He was also the co-Chair of a national study to determine a Roadmap for AI Research guiding US government research investments in AI. Prof. Selman was previously at AT&T Bell Laboratories. His research interests include artificial intelligence, computational sustainability, efficient reasoning procedures, machine learning, deep learning, deep reinforcement learning, planning, knowledge representation, and connections between computer science and statistical physics. He has (co-)authored over 150 publications, including six best paper awards and two classic paper awards. His papers have appeared in venues spanning *Nature*, *Science*, *Proc. Natl. Acad. of Sci.*, and a variety of conferences and journals in AI and Computer Science. He has received the Cornell Stephen Miles Excellence in Teaching Award, the Cornell Outstanding Educator Award, an NSF Career Award, and an Alfred P. Sloan Research Fellowship. He is a Fellow of the American Association for Artificial Intelligence (AAAI), a Fellow of the American Association for the Advancement of Science (AAAS), and a Fellow of the American Association for Computing Machinery (ACM). He is the recipient of the Inaugural IJCAI John McCarthy Research Award, the premier award for artificial intelligence researchers at mid-career stage.

Some Engineering Views to the AI Development and Governance

By GONG Ke

Artificial Intelligence (AI) has been now going out of scientific papers and technical labs into engineering applications and into people's daily lives. This is indeed the turning point in the historical development of AI. Engineering and engineers have become a crucial player of AI in its applications and further development as well as in its governance, to ensure AI for the good of all people and the planet.

AI is a new tool for human development, and the ultimate goal of its development and governance is to ensure it is good for sustainable development.

Toolmaking is an important feature of human being which is different from other animals. From simple stone tools and ironware to levers, pulleys, and all kinds of machinery and electrical power, to computer and modern information network, human being has used these tools to extend its physical and intelligent strength and to increase its capacity to survival and wellbeing. Hence, human history is often characterized by the progress of the human-made tools, such as Stone Age, Iron Age, Machine Age and Information Age, etc. In this regard, AI is nothing but a new tool for human development, which is bringing humankind into the Intelligent Age.

It must be pointed out that the purpose of human inventing and using tools is to solve problems in survival and development. At present, the foremost problem facing humankind is the issues of sustainable development. Therefore, AI should become a powerful tool for human sustainable development. This is also the ultimate goal of AI development and governance. In 2020, a notable development is that the role of AI in achieving the Sustainable Development Goals (SDGs) has been analyzed by researchers and paid attentions by many international organizations. In order to guide AI development and support its governance, more comprehensive and insightful researches are urgently required to explore the relationship between AI and the SDGs.

Engineering society should shoulder the responsibility and be an important part of the joint AI development and governance by multiple stakeholders.

The World Federation of Engineering Organizations (WFEO) realizes its responsibility for promoting AI for the good of people and its environment. An interdisciplinary working group with experts from industry and academia has been established by its standing technical committee of Engineering for Innovative Technologies. On March 4th, 2020, the first World Engineering Day for Sustainable Development , WFEO released its position paper on "Promoting Responsible Conduct of Big Data and AI Innovation and Application in Engineering" , in which 7 principles have been proposed, they are:

- Good for Humanity and Its Environment
- Inclusiveness, Fairness, Public Awareness and Empowerment
- Opening and Sharing while Respecting Privacy and Data Integrity
- Transparency
- Accountability
- Peace, Safety and Security
- Collaboration

It is found that many governments, NGOs, corporate, etc. have declared the principles of AI applications. Although these principles are in common to a big extent, they are still far from sufficient to reach a global consensus. Generally speaking, engineering has not yet played its important role in AI governance, and there is a shortage of crosstalk between different sectors and different societal groups. Global dialogues of multiples stakeholders are urgently needed.

The green development of AI should become a focus of AI development and its governance.

AI has an incredible potential to improve the productivity, quality, safety and efficiency of various engineering projects, so that AI is empowering and transforming every aspect of engineering.

It is exciting to see many applications of AI for increasing energy efficiency in various productions and households, and it is widely recognized that the power efficiency of AI products is far behind human workers. However, the green and lower carbon development of AI has not yet been emphasized in AI R&D and applications. Therefore, it must call for a full understanding of the rigid constraints of AI development, and paying more attention to the green development and application of AI, which should also become one of the focuses of AI governance.

Additionally, it is noticeable that although many research and development works have been carried out for AI safety and security. Compared to the antivirus software and services market some 20 years ago when PC started its wide applications, there is still a lack of products and services for AI safety and security.

ABOUT THE AUTHOR GONG Ke

Prof. GONG Ke is President of the World Federation of Engineering Organizations (WFEO), and Executive Director of the Chinese Institute of New Generation Artificial Intelligence Development Strategies.

Prof. GONG is an engineer in information communication technology by training. After obtaining the degree of Doctor of Technical Science from Technical University of Graz, Austria, he had worked in Tsinghua University from 1987-2006, where he served as the director of the Chinese National Research Center of Information Science and Technology, and the vice president of the University; later, he was appointed as President of Tianjin University (2006 - 2011), and President of Nankai University (2011 - 2018).

Prof. GONG has worked in WFEO since 2009, had chaired the Committees of Information and Communication, and then of Engineering for Innovative Technologies. He was elected as President and took the office in 2019.

Prof. GONG has been directing the Chinese Institute of New Generation Artificial Intelligence Development Strategies, since it cofounded in 2017 by Chinese Academy of Engineering and Tianjin Municipal Government.

From 2013 to 2017, Prof. GONG Ke was member of the Scientific Advisory Board to the UN Secretary-General Ban Ki-moon. He has also served at executive positions of China Association of Science and Technology, and etc.

How Can We Use Data and AI for Good, Without Also Enabling Misuse?

By Claudia Ghezzou Cuervas-Mons, Emma Bluemke, ZHOU Pengyuan and Andrew Trask

How can we gain beneficial use from data without collateral harm? As AI development progresses, this question is ever more important. The solution to this question is an aim we call “structured transparency”.

Countless modern activities and services demand access to sensitive and personal information in order to grant the benefits of the services. Such exchanges, at times, lead to unfavourable transactions for the user; data privacy neglect and harm from data misuse.

Nonetheless, data use is fundamental for the application and development of AI algorithms in society.

To address this challenge, many privacy-enhancing technologies (PETs) such as secure computation, federated learning and differential privacy have been developed to allow data use while ensuring privacy. However, no one technique solves data privacy issues outright.

However, when used in certain combinations, these techniques could allow data to be used only for the intended or approved purpose. These techniques could allow AI researchers and governance bodies to arrange a very precise technical or social arrangement of who can see what information.

In this article, we propose a useful framework for thinking about how emerging technologies can help you achieve your desired information flow. To achieve structured transparency - in other words, to allow only beneficial use from data while mitigating harms - one must consider what information is shared with whom, when, and how. To do so, we suggest considering the following five components: a framework for thinking about structured transparency, and importantly, which emerging technologies can help you achieve it.

1. Input privacy: enables utilising hidden information without revealing its contents.Technical input privacy tools come primarily from the field of cryptography - public-key cryptography, end-to-end encryption, secure multi-party computation, homomorphic encryption, functional encryption, garbled-circuits, oblivious RAM, federated learning, on-device analysis, and secure enclaves are several popular (and overlapping) techniques capable of providing input privacy.

2. Output privacy: enables reading and receiving information and prevents reverse engineering so that the input can be concealed. Technical output privacy tools (chiefly, differential privacy and related techniques) can provide strict upper bounds on the likelihood that a data point could be reverse-engineered.

3. Input verification: ensures the robustness and reliability of the data source. Most input verification techniques rely on combinations of public-key infrastructure (SSI, Key Transparency, etc.), cryptographic signatures, input privacy techniques with active security, and zero-knowledge proofs.

4. Output verification: ensures the computations taken on a given information flow are legitimate. The major limitation of output verification tools is that the verifier must examine the data in order to perform the verification.

5. Flow governance: guarantees that the aforementioned components of the information flow are met, thus ensuring structured transparency; in other words, that the intended flow of information is preserved throughout each of its components. Technically, flow governance could be best exemplified by Secure Multi-Party Computation (SMPC), where impartial parties overview the flow of arbitrary information. However the flow is also fundamentally driven by the incentives of the system, whether that be the incentives of the algorithm via the optimization metric (screen-time? diagnosis accuracy?) or the incentives of the parties involved.

As data governance and data privacy issues increase, we hope this framework provides clarity for how to apply emerging privacy technologies to your particular use case.

ABOUT THE AUTHOR

Open Mined:
Claudia Ghezzou Cuervas-Mons, Emma Bluemke (has chosen not to display her information), ZHOU Pengyuan and Andrew Trask

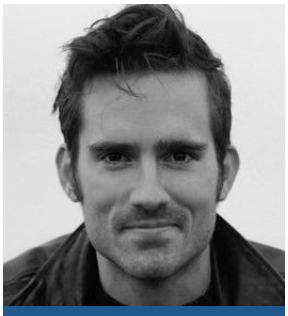
Claudia Ghezzou Cuervas-Mons

Claudia is currently working as a Clinical Data Coordinator in a CRO. Her background focused on Biochemistry (BSc) and Neuroscience (MSc), doing her Master's thesis on the automated quantification of ischaemic events on CT scans with Dr. Bentley at Imperial College London.

She has a personal interest on the way machine learning can advance medicine and science, how to apply these tools for optimised disease detection and treatment. Additionally, she is appealed to explore the societal implications of the advances in AI and how to ensure its impact is as much as possible a positive one.

ZHOU Pengyuan

Pengyuan is a postdoc researcher working with Prof. PAN Hui at the Department of Computer Science, University of Helsinki, and SyMLab in HKUST. In May 2020, Pengyuan got his Ph.D. from University of Helsinki supervised by Prof. Jussi Kangasharju in Collaborative Networking Lab. His research focuses on mobile edge computing and communication systems, connected vehicles and edge AI. The key motivation for his research is the heavy demand for computational and networking resources on the edge of the network.



Andrew Trask

Andrew Trask is a PhD student at the University of Oxford studying privacy preserving AI techniques and the leader of the OpenMined community — a group of over 11,000 members dedicated to lowering the barrier to entry for privacy enhancing technologies.

Human-Centered AI/Robotics Research and Development in the Post-Pandemic Era

By ZHANG Jianwei

In this extremely exceptional SARS-CoV-2 pandemic year, AI has to face the global anti-pandemic challenges besides its on-pace progress in adding more top-down approaches, cross-modal learning, and unsupervised architecture to supervised deep learning approaches guided by single-criterion optimization. Mostly in home-office mode and based on data/algorithms, clear advances can be observed both on the fundamental research and the novel application level, shaping the theory and practice of AI governance.

There is an urgent need for useful AI and robot systems to curb the rapid spread of the SARS-CoV-2 virus and to bring daily work and life back to normal. These requirements put the focus of AI governance on the precise modeling of the paths of the virus infection, the rapid development of real-time testing methods, effective vaccines and drugs, and on autonomous robot systems that can assist medical doctors/nurses and replace human workers in labor-intensive tasks such as delivery, harvesting, factory assembly, etc.

We have developed an AI approach to determine the likelihood of asymptomatic carriers of the SARS-CoV-2 virus by using interaction-based continuous learning and the inference of individual probability to rank the likelihood of contagion. Compared to traditional contact tracing methods, our approach significantly reduces the screening and the duration of quarantine required to search for the potential asymptomatic virus carriers by as much as 94%. At the same time, we are integrating human-robot interaction, multi-level learning and robot decision-making to develop an autonomous, safe and intelligent robot system for both nasopharyngeal and oropharyngeal swab sampling, which will protect medical staff. Collaborating with Pixelbiotech, a German start-up, we have applied AI to empower the imaging data analysis of a multiplex smFISH (single molecule Fluorescence In Situ Hybridization) probe, which reduces the testing time of SARS-CoV-2 virus to 15 minutes. Overall, we believe that AI and robotics are having a growing impact on the fight against the global pandemic by modelling the precise spreading process as well as by automatically and efficiently detecting the virus.

Based on the Sino-German interdisciplinary collaborative project TRR169 on “Cross-modal Learning: Adaptation, Prediction and Interaction”, we continue to research and develop human enhancement to facilitate physical human-robot collaboration. Furthermore, we are developing applications for human enhancement and support in collaboration with laboratories from neuroscience and psychology, such as neuro-computational representation and therapeutic application of neurological diseases. Such long-term human-centered research and development represents one of the most important AI vertical applications.

Within TRR169, we use the established collaboration between the partners from Hamburg and Beijing to achieve a higher level of understanding, modelling and implementing crossmodal systems, and of understanding and unifying the neural, cognitive and computational mechanisms of crossmodal learning. Significant progress in several areas and in particular in deep-learning (algorithms, software, and applications) encourages us to pursue a more integrated set of objectives as future-oriented, strong-AI research themes: novel

learning architectures/strategies, robustness, anticipation and prediction, generalization and transfer, and benchmarking. All in all, our mission is to develop transparent, interpretable brain-inspired AI approaches, which are a significant technological feature of AI governance.

ABOUT THE AUTHOR

ZHANG Jianwei

ZHANG Jianwei is professor and director of TAMS, Department of Informatics, University of Hamburg, Germany. He received both his Bachelor of Engineering (1986, with distinction) and Master of Engineering (1989) at the Department of Computer Science of Tsinghua University, Beijing, China, his PhD (1994) at the Institute of Real-Time Computer Systems and Robotics, Department of Computer Science, University of Karlsruhe, Germany. His research interests are cognitive sensor fusion, robust robot perception, intelligent robot architecture, multimodal human-robot interaction, dexterous robot manipulation, brain-inspired machine learning, etc. In these areas he has published over 400 journal and conference papers, technical reports and six books. He holds over 40 patents on intelligent components and systems. His research results have been applied in robotic systems for real-world scenarios such as medical assistant, rehabilitation, advanced 3C assembly, online quality monitoring of industrial processes, etc. ZHANG Jianwei is coordinator of the DFG/NSFC Transregional Collaborative Research Centre SFB/TRR169 “Crossmodal Learning” and several EU, German and industrial AI projects. He has received multiple best paper awards. He is the General Chairs of IEEE MFI 2012, IEEE/RSJ IROS 2015, and the International Symposium of Human-Centered Robotics and Systems 2018. He is life-long Academician of Academy of Sciences in Hamburg Germany.

AI and Data Governance for Digital Platforms

By Alex Pentland

Modernizing and digitizing governances of national, international, and commercial interactions to become more efficient, transparent, and inclusive is a key global priority, and dozens of efforts are already underway. However, current efforts are mostly piecemeal and incremental.

Governance of digital platforms has become unexpectedly urgent with the pilot deployment of nationally-backed digital platforms that provide a uniform framework for not only finance but trade and logistics, authentication, fraud detection and analytics (e.g., AI). China, for instance, is moving existing Silk Road investments onto Chinese digital systems that are dramatically more agile and cheaper than Western systems. Singapore has developed a similar digital trade and logistics infrastructure for investments within its Temasek Sovereign Wealth Fund, and Switzerland has recently deployed the Swiss Trust Chain (with help from MIT Connection Science program). Finally, most major economies have either deployed or are seriously considering deployment of national digital currencies. We have been involved in deployment of two such currencies, and soon will help launch the digital version of a major trading currency.

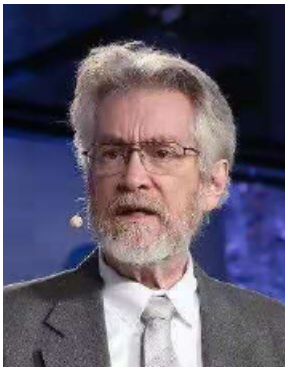
These systems are poised to integrate the majority of the world's trade into efficient, unified frameworks that seamlessly interoperate across sovereign and institutional borders. However, their accountability, inclusiveness and governance may not satisfy many nations. It is imperative that nations engage in the standards specification and deployment of these digital governance systems.

Perhaps the first challenge to be addressed by any new system for digital governance is repairing the world's tattered finances. If nations do not cooperate, we risk a "race to the bottom", and smaller nations will suffer the most. Moreover, unlike at the end of World War II, the deployment of these new digital trade platforms will provide nations with possibilities for beggaring their neighbors in ways that are far less visible than an official devaluation.

This suggests that a new "Bretton Woods" multilateral effort is required, with the goal of renovating multilateral institutions using the more efficient, secure, and inclusive digital platforms that are analogous to those developed by China, Singapore, and Switzerland. Unlike the World War II effort, such coordination must not only be centered around banking and finance, but must be intimately dependent on digital technical standards such as those created by the IEEE and the computational social science needed to measure and forecast interactions between finance, sustainability, and social factors.

ABOUT THE AUTHOR

Alex Pentland



Professor Alex “Sandy” Pentland directs the MIT Connection Science and Human Dynamics labs and previously helped create and direct the MIT Media Lab and the Media Lab Asia in India. He is one of the most-cited scientists in the world, and Forbes recently declared him one of the “7 most powerful data scientists in the world” along with Google founders and the Chief Technical Officer of the United States. He has received numerous awards and prizes such as the McKinsey Award from Harvard Business Review, the 40th Anniversary of the Internet from DARPA, and the Brandeis Award for work in privacy. He is a founding member of advisory boards for Google, AT&T, Nissan, and the UN Secretary General, a serial entrepreneur who has co-founded more than a dozen companies including social enterprises such as the Data Transparency Lab, the Harvard-ODI-MIT DataPop Alliance and the Institute for Data Driven Design. He is a member of the U.S. National Academy of Engineering and leader within the World Economic Forum.

Over the years Sandy has advised more than 60 PhD students. Almost half are now tenured faculty at leading institutions, with another one-quarter leading industry research groups and a final quarter founders of their own companies. Together Sandy and his students have pioneered computational social science, organizational engineering, wearable computing (Google Glass), image understanding, and modern biometrics. His most recent books are *Social Physics*, published by Penguin Press, and *Honest Signals*, published by MIT Press. Interesting experiences include dining with British Royalty and the President of India, staging fashion shows in Paris, Tokyo, and New York, and developing a method for counting beavers from space.

Alignment Was a Human Problem First, and Still Is

By Brian Christian

Over the past decade we saw how deep learning made it possible for networks to perform complex tasks like identifying a face without the need for any manual "feature engineering". Something similar is beginning to happen in this decade with values.

To use a social network like Twitter as an example, instead of manually identifying ratios of comments to shares to likes that constitute "healthy" engagement on the platform, employees can simply identify certain communities as being "healthy" or "unhealthy". The system can then *infer* which specific metrics predict a healthy community. The platform can then be optimized in a way that promotes "healthy" engagement – potentially without any human ever explicitly defining what metrics, exactly, constitute "healthy" engagement [1].

As with many advances in technological capability, this alleviates one problem while emphasizing another. We end the problem of metrics, but we are left with the problem of judgment. *Who*, for instance, decides what constitutes "healthy engagement"? Who provides oversight? The engineering problem of specifying one's priorities *numerically* gives way to the human question of governance.

Etymology is often a surprisingly helpful guide to the present. The use of "alignment" in AI first came from computer scientist Stuart Russell in 2014. But he borrowed the term from the fields of economics and management science, which have been speaking for decades about how values and interests are "aligned" within and between organizations.

As the "alignment problem" between a system's engineers and an AI system they create begins to be "solved", our attention must turn to the larger question of alignment in its original sense: between that team of engineers and their managers, between the managers and the executives, between the executives and the shareholders, between the company as a whole with its regulators and its users.

The etymology reminds us that "alignment" has always been a problem that exists between humans and other humans. As we enter into the age of AI, this will be truer than ever.

References

[1]Milli, S., Belli, L., & Hardt, M. (2021). From Optimizing Engagement to Measuring Value. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 714-722. https://doi.org/10.1145/3442188.3445933

ABOUT THE AUTHOR

Brian Christian

Brian Christian is a Visiting Scholar at the University of California, Berkeley and the bestselling author of three acclaimed books of nonfiction about the human implications of computer science: *The Most Human Human*, *Algorithms to Live By* (with Tom Griffiths), and *The Alignment Problem*. His books have been named a *New York Times* Editors' Choice, a *Wall Street Journal* bestseller, a *New Yorker* favorite book of the year, Amazon best science book of the year, and *MIT Technology Review* best book of the year, and have been translated into nineteen languages. Christian has lectured at Google, Facebook, Microsoft, the Santa Fe Institute and the London School of Economics, and holds degrees in philosophy, computer science, and poetry from Brown University and the University of Washington. He lives in San Francisco.

On Governability of AI

By Roman Yampolskiy

In order to make future AIs beneficial for all of humanity, AI governance initiatives attempt to make AI governable by the world's governments, international organization and multinational corporations collaborating on establishing a regulatory framework and industry standards. However, direct governance of AI is not meaningful, and what is implied by the term is governance of AI researchers and creators in terms of what products and services they are permitted to develop and how. Whatever it is possible to govern scientists and engineers working on AI depends on the difficulty of creating Artificial General Intelligence (AGI).

If computational resources and data collection efforts necessary to create AGI are comparable in cost and human capital to the Manhattan project conducted by USA to develop nuclear bomb technology, governments have a number of "carrots" and "sticks" they can use to guide researchers and to mold the future AI to their specifications. On the other hand, if it turns out that there is a much more efficient way to create the first AGI or a "seed" AI which can grow into a full-blown superintelligence, for example, by a teenager on a $1,000 laptop in their garage (an admittedly less likely, but nevertheless possible scenario), governments' attempts at regulation may be futile. We note that historical attempts at software governance (ex. spam, computer viruses, deep fakes) had only a very limited amount of success. With AGI as an independent agent, it may be ungovernable because traditional methods of assigning responsibility and punishment-based enforcement are not applicable to software.

Even presuming a, resource-heavy, favorable case for governance, we are still left with a number of established technical limits to AI predictability [1], explainability [2], and controllability [3]. It follows that AI governability, which requires, at least, those three capabilities for successful regulation is likewise only partially achievable, meaning smarter than human AI would be ungovernable by us in some important ways. Finally, even where AI governance is achievable, those in charge may be unwilling to take personal responsibility for AI's failures [4], or deliberate actions even if performed in the context of instituted governance framework. Consequently, a highly capable, creative and uncontrolled AGI may end up implicitly or even explicitly controlling some of the institutions and individuals, which we entrusted to govern such intelligent software.

References

[1]Yampolskiy, R. V. (2020c). Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent. Journal of Artificial Intelligence and Consciousness, 7(01), 109-118.

[2]Yampolskiy, R. V. (2020b). Unexplainability and Incomprehensibility of AI. Journal of Artificial Intelligence and Consciousness, 7(02), 277-291.

[3]Yampolskiy, R. V. (2020a). On Controllability of AI. arXiv preprint arXiv:2008.04071.

[4]Yampolskiy, R. V. (2019). Predicting future AI failures from historic examples. foresight.

ABOUT THE AUTHOR

Roman Yampolskiy

Dr. Roman V. Yampolskiy is a Tenured Associate Professor in the department of Computer Science and Engineering at the Speed School of Engineering, University of Louisville (UoL). He is the founding and current director of the Cyber Security Lab and an author of many books including *Artificial Superintelligence: a Futuristic Approach*. During his tenure at UofL, Dr. Yampolskiy has been recognized as: Distinguished Teaching Professor, Professor of the Year, Faculty Favorite, Top 4 Faculty, Leader in Engineering Education, Top 10 of Online College Professor of the Year, and Outstanding Early Career in Education award winner among many other honors and distinctions. Yampolskiy is a Senior member of IEEE and AGI; Member of Kentucky Academy of Science, and Research Associate of GCRI. Dr. Yampolskiy's main areas of interest are AI Safety and Cybersecurity. Dr. Yampolskiy is an author of over 100 publications including multiple journal articles and books. Dr. Yampolskiy has been an invited speaker at 100+ events including Swedish National Academy of Science, Supreme Court of Korea, Princeton University and many others.

Part II Responsible Leadership from the Industry

Operationalizing AI Ethics: Challenges and Opportunities

By Anand S. Rao

The spread of COVID-19 was swift and caught most governments, companies and citizens off-guard. Every aspect of life for almost every individual on this planet has been impacted by COVID-19. From direct impact (e.g., death, hospitalization, infection) to indirect impact (e.g., loss of job, working from home, mental health), the virus has affected almost everyone on this planet.

As a direct response to COVID-19 there has been a significant rise in the use of digital channels across all sectors in all countries. Riding on the back of this surge has been the increased adoption of advanced analytics, automation, and Artificial Intelligence (AI). We conducted a global survey of 1,018 executives in November, 2020, to understand how the crisis has impacted companies and their attitudes towards AI. We highlight two key aspects from this survey in this brief paper.

1. AI Investments are on the rise inspite of the economic crisis

In our survey, 44% of the respondents said that COVID-19 had a negative impact on their business and surprisingly an equal number said that COVID-19 had a positive impact on their business. Interestingly, the larger the companies (>$10bn in revenue) the more likely for them to have seen a significant positive impact from COVID-19. In addition, these larger companies, nearly 4 in 10, had invested more in AI development before the pandemic and were moving from experimental to operational use of AI. These companies, having seen their return on AI during the pandemic, were significantly more likely to increase their use of AI (38%), explore new use cases for AI (39%) and train more employees to use AI (35%). This is true not just of large companies, but also smaller companies that had heavily invested in AI prior to the pandemic. In addition, the global venture capital funding for AI rose to $71.9 billion in Q3, 2020, surpassing the previous record quarter in Q4, 2018 of $69 billion.

2. Managing and mitigating AI risks using Responsible AI is becoming critical

The spread of COVID-19 has seen an increased use of AI in applications such as facial recognition, contact tracing, monitoring of employee movement etc. Identifying, mitigating, and managing AI risks - ranging from bias, privacy, transparency, accountability, explainability, robustness, safety, and security - is one of the major challenges for companies as they deploy AI models and systems to address and manage through the iris. Only 12% of companies had fully embedded and automated

AI risk management and controls in place. Another 37% of respondents had strategies and policies to tackle AI risks, but has no automated solutions. For the fully embedded AI segment these numbers went up to 29% (with embedded and automated AI risk management) and 38% (with strategies and policies) respectively. Of all the different areas of AI risks - algorithmic bias is a central focus of nearly 36% of all respondents. Reliability, robustness, security, and data privacy are some of the other AI risks that feature prominently amongst the companies scaling their AI.

In summary, the increased adoption of AI, increased investments in AI, and the risks posed by AI offer an opportunity to adopt responsible AI practices to manage and mitigate these risks.

ABOUT THE AUTHOR

Anand S. Rao

Anand is a Partner in PwC Advisory, with over 32 years of experience in industry and in research. He leads PwC's Artificial Intelligence efforts globally and is the Innovation Lead for the Emerging Technology Group. He is responsible for a team of practitioners who work with C-level executives, advising them on a range of topics including global growth strategies, marketing, sales, distribution and digital strategies, behavioral economics and customer experience, risk management and statistical and computational analytics. His experience spans primarily financial services, insurance, telecommunications, and healthcare. He lived and worked at clients in Asia, Australia, Europe and Americas.

Anand is responsible for research and commercial relationships with academic institutions and start-ups focused on new and innovative big data and analytic techniques. With his PhD and research career in Artificial Intelligence and subsequent experience in management consulting, he brings a unique combination of business domain experience, statistical, and computational analytics expertise to generate unique insights into the theory and practice of “data science”.

Prior to his strategic consulting career, he has led and managed innovative artificial-intelligence based approaches for air-combat modeling, air-traffic management, customer service decision support, and telecommunications network management. He has also taught and supervised doctoral students at the University of Melbourne and served as the Program Director of the Center for Intelligent Decision Systems.

He has co-edited four books, published over fifty peer-reviewed papers in conferences and journals. He frequently speaks at technology and business forums on analytics, AI, behavioral economics, and innovation; served on the Organizing and Program Committees of International Journals, Conferences, and Workshops.

Patterns of Practice Will Be Fundamental to the Success of AI Governance

By Abhishek Gupta

AI governance has certainly gained steam in 2020 with a lot of calls to action that have leveraged expertise in both the legal and technical fields to propose frameworks to govern both the development and deployment of AI systems. There are a lot of commonalities in these initiatives, with most of them focusing on areas of transparency, accountability, bias, privacy, non-discrimination, and other generally agreed upon values from the over 100 sets of principles in AI ethics, with most having at least some component focused on AI governance.

There has been a noticeable movement from 2019 when AI governance was a topic of discussion where people talked about abstract ideas and 2020 saw much more of a push to actually put those ideas into practice. Yet, as much as we saw movement, there were still some shortcomings that hindered the deployment of these governance mechanisms. In particular, 2020 was a year where we saw hasty roll-outs of these systems in tracking face mask compliance [1], grading students [2], handing out unemployment benefits [3], and more. So, what could we have done better?

As I have detailed in my work titled Green Lighting ML: Confidentiality, Integrity, and Availability of Machine Learning Systems in Deployment [4] that I presented with my co-author Erick Galinkin at several conferences in 2020 including ICML, what we have seen is that there is little focus on the practical manifestation of these ideas. Specifically, there is a missing focus on the needs and patterns of practice of designers and developers on the ground who will have at least partial responsibility in operationalizing these ideas. This is not to say that government mandates and management of the organization is not going to play an important role in how AI systems are governed. Quite the contrary, it is in fact essential that we consider the measures I am going to recommend as a supplement to the others, especially as they will help to bolster the efficacy of any other organizational-scale mechanisms that are applied in AI governance.

From a practitioner's perspective, there are numerous challenges that one faces when they encounter abstract principles coupled with business pressures and deadlines to deliver products and services on time and with high quality. It is at these points that there is a breakdown in the actual operationalization of the AI governance mechanisms which needs to be fixed.

From my experience, the first method that helps to mitigate this issue is to strive to incorporate pieces of the governance requirements within existing workflows of designers and developers rather than jumping to create net new mechanisms. The benefit of doing so is that there is lower friction in the acceptance of these new requirements and they are also quicker to deploy and then gather evidence to see if they are effective or not. Armed with this evidence, one can make a stronger case for their incorporation at a wider level. Second, and perhaps the most important aspect of creating AI governance solutions is to include the practitioners in the process of developing these mechanisms. The requirement there is two-fold: one, you are able to surface the exact places where the AI governance solutions might fail when they are asked to be implemented in practice based on the experience of the practitioners and two, you also build trust with those practitioners so that they are not only aware of what will be asked of them but given that they are active contrib-

utors, they will have a strong sense of ownership and desire to see this succeed. Some of these insights are also discussed in my book *Actionable AI Ethics* (https://atg-abhishek.github.io/actionable-ai-ethics) that takes this very hands-on and practical approach to putting AI ethics into practice addressing some of the very challenges that I have highlighted here.

Thus, keeping in mind these patterns of practice will be crucial if we are to actually move forward in putting AI governance to work rather than spend another precious few months and years debating on the abstract ideas. The time for action is now and it starts by paying attention to how these systems are actually designed and developed in practice.

References

[1]https://www.theverge.com/2020/5/7/21250357/-france-masks-public-transport-mandatory-ai-surveillance-camera-software

[2]https://www.wired.co.uk/article/gcse-results-alevels-algorithm-explained

[3]https://www.usnews.com/news/best-states/articles/2020-02-14/ai-algorithms-intended-to-detect-welfare-fraud-often-punish-the-poor-instead

[4]Gupta, A., & Galinkin, E. (2020). Green Lighting ML: Confidentiality, Integrity, and Availability of Machine Learning Systems in Deployment. ArXiv:2007.04693 [Cs, Eess]. http://arxiv.org/abs/2007.04693

ABOUT THE AUTHOR

Abhishek Gupta

Abhishek Gupta is the founder of the Montreal AI Ethics Institute (MAIEI) and a Machine Learning Engineer at Microsoft where he serves on the CSE Responsible AI Board. He is the author of the forthcoming book titled *Actionable AI Ethics* that will be a practical and hands-on guide for operationalizing AI ethics.

He is a Visiting AI Ethics Researcher, Future of Work in the International Visitor Leadership Program with the U.S. Department of State, the Responsible AI Lead, Data Advisory Council for the Northwest Commission on Colleges and Universities, AI Advisory Board Member for Dawson College, Associate Member of the LF AI Foundation at The Linux Foundation, and a Faculty Associate in the Frankfurt Big Data Lab at Goethe University. Abhishek's research focuses on applied technical and policy methods to address ethical, safety and inclusivity concerns using AI in different domains. He has built the largest community-driven, public consultation group on AI ethics in the world that has made significant contributions to the development of many initiatives in the domain of responsible AI. More information about his work can be found at https://atg-abhishek.github.io.

Building on Lessons for Responsible Publication: Safely Deploying GPT-3

By Irene Solaiman

2020 combined unprecedented events in the world with precedented challenges in AI. The AI research community is grappling with best practices for responsible publication and safe deployment, but now largely coordinates remotely in work-from-home environments. Concurrently, existing concerns, like disinformation, show present-day consequences, like in pandemic response and political institutions.

AI systems, particularly generative models, have become increasingly powerful and demand safeguards. Generative models are trained on data such as text or images and seek to generate outputs that are similar to that data; for language models, that can mean predicting the next word in a sentence. Our concerns about risks are shared in industry, academia, and the public. Notably, Dr. Emily Bender and Dr. Timnit Gebru co-authored a paper raising concerns with powerful language models, highlighting concerns like embedded biases. Work by OpenAI and other researchers has shown harmful biases, potential for misuse and generating disinformation, and difficulty detecting synthetic text.

Recognizing these risks, OpenAI conducted an unusual responsible publication strategy in 2019. We released incrementally powerful versions of our language model GPT-2 in stages to research model characteristics before each release. Our larger, higher-performing language model, GPT-3, required further consideration. Like its predecessor, it has flexible capabilities, from text summarization, translation, and question-answering to even three-digit arithmetic. GPT-3 also has a strong “few-shot learning” ability, i.e. the ability to solve problems given few demonstrations. GPT-3 has higher misuse potential than GPT-2 and still shows discriminatory bias, necessitating its careful deployment.

In the interest of safety, we released GPT-3 through an API; OpenAI hosts the system, and approved users can access it with an API key. We provide an interface for users to experiment, develop new applications, or conduct research. This release was part of a company decision to safely productize the system to help fund our research, but includes an academic access program for researchers to help identify model characteristics, especially on key areas of bias, misuse, and detection.

An API functions as a means of safety, security and accessibility. We deployed GPT-3 without releasing the full system to the public, allowing us to easily respond to misuse and improve the system as we learn about it. We outlined usage guidelines, vet all use cases, and terminate any that cause harm or have insufficient safeguards. We host the system, which is expensive to run, alleviating cost pressures and making the system accessible for small businesses and organizations.

Since the system’s flexible capabilities make all possible use cases impossible to predict, we limit users and are broadening access over time. Researchers at Middlebury Institute, University of Washington, and the Allen Institute for AI helped us scope disinformation and bias risks respectively. Our internal research helps build our usage guidelines and model improvements.

Recent foundational work in publication norms proves

the need to continually invest as systems grow more powerful. AI researchers, including at the Partnership on AI and Microsoft, informed practices like broader impacts analyses and model documentation. Impacts are context-dependent and mitigation requires not just technical, but social science and sociotechnical research. This research must continue with diverse communities and adapt to both the fast-pace of AI advancements and unexpected challenges of unprecedented events.

ABOUT THE AUTHOR

Irene Solaiman

Irene Solaiman is a Policy Policy Manager at OpenAI. She leads policymaker engagement and conducts social impact and fairness analysis. Prior to OpenAI, she was a fellow at Harvard's Berkman Klein Center as part of the Assembly Student Fellowship researching the ethics and governance of AI. Irene holds a Master in Public Policy from the Harvard Kennedy School and a B.A. in International Relations from the University of Maryland where she was elected Phi Beta Kappa.

Artificial Intelligence Should Follow Sustainable Development Principles

By YANG Fan

As Artificial Intelligence(AI) grows more mature, its development also extends to other business areas, empowering all industrial sectors and exerting an unprecedented influence on the global economic system. The demand for AI talents has also changed from technical specialization to interdisciplinary thinking, or “AI + X”, along with the constantly deepening development of AI covering all sector and multiple fields. Meanwhile, discussions on the risks and ethical hazards associated with AI in the future are also increasing. We should not only consider AI from a mere technical perspective, but also attach importance to the social values that AI brings about in empowerment.

From the perspective of social attributes of AI, “sustainable development” is a must for building the AI ecology. The concept of “Sustainable Development of AI” derives from the United Nations’ (UN) seventeen world-transforming goals. In other words, achieving the sustainable development of AI may create social values apart from achieving economic goals, which is helpful to building a community with a shared future for mankind. In June 2020, the UN launched the Roadmap for Digital Cooperation, which outlines eight key areas for action, such as achieving universal connectivity and ensuring digital inclusion for all. This roadmap explains more clearly the value and implications of sustainable development in the era of AI and big data.

On June 22, 2020, SenseTime Intelligent Industry Institute (hereinafter referred to as “SenseTime”) and QingYuan Research Institute, Shanghai Jiao Tong University, jointly held the “AI Sustainable Development Forum 2030”, and released the White Paper on the Sustainable Development of AI. It is the first time for the Chinese AI industry to discuss the concept of, and planning for, sustainable development. The White Paper proposes core values of people orientation, shared benefits, integrative development and scientific research innovation, clarifies the negotiatory and inclusive principles of AI ethics (respect, open dialogues, and inclusive culture), the all-benefiting and altruistic principles of AI to benefit people (sharing benefits and building an inclusive society), the conscientious and self-disciplined principles of AI to empower industries (accountability, self-discipline, and safety), and the open and sharing principles to develop a trustworthy AI (open innovation to make technology more reliable), and proposes new concepts and ideas for AI governance.

During the promotion of the framework of sustainable development, SenseTime has practiced the principles of sustainable development with practical actions. In July 2019, SenseTime established the Committee for AI Governance and Ethics, and designed the framework for AI risk assessment and management. All projects must be reviewed and passed by the committee before being launched. SenseTime has also put a lot of efforts in segment industries. For example, on September 23, 2020, SenseTime launched its sub-brand, “SenseTime Education”, which narrows the AI digital gap with a series of products such as education platform and teacher training programs, to deliver more innovative talents for the AI field. This action represents the UN's sustainable goal of quality education. In the medical field, the SenseCare®, the AI-supported diagnosis and treatment system of SenseTime effectively contains the spread of COVID-19, while improving the efficiency of

health professionals, and contributes to the UN's sustainable goal of good health and well-being. Similar cases, which represent the important ideas and practices of SenseTime concerning the empowerment of industries with AI, are too numerous to be enumerated.

As the pioneer of computer vision, SenseTime will constantly play its role in the sustainable development and governance of AI, acquire an in-depth understanding of the new trends in global science and technology development, open up new development paths of ethics and governance, further promote the corporate concept of building "Responsible AI", and stick to the principles of sustainable development, to deeply intertwine social responsibilities with corporate development.

ABOUT THE AUTHOR YANG Fan

Mr. YANG is currently Vice President & Co-Founder and Chairman of AI Ethics & Governance Committee of SenseTime, in charge of the planning and development of SenseTime's ecosystem and integration with other industries. Before joining SenseTime, he worked at Microsoft Research Asia (MSRA) for many years. At MSRA, he was actively engaged in the research and development of new technologies, many of which are widely applied in Microsoft's flagship products worldwide. He has accumulated rich experience in planning and implementing the commercialization of cutting-edge technologies.

Since the foundation of SenseTime, YANG Fan has led the company to successfully practice the innovative business model of "AI+Industry" in industries such as smartphone, retailing, and smart city, and empowered the engineering and tech team of SenseTime to develop a comprehensive methodology that covers the full cycle from underlying fundamental architecture to technological industrialization. In industrial practice, he spearheaded the commercialization of AI in fintech and Internet entertainment, masterminded the overall layout of AI and its exponential growth in the smart city industry, and was the first to put forward the concept of a city-level video analysis center and put it to practice, leading SenseTime to become the pacesetter in this field. In terms of business promotion, he led the team in establishing major cooperation with strategic customers such as China Mobile, Huawei, Xiaomi, and Honda. It was also due to his decision and leadership that SenseTime successfully applied for the intelligent vision project of China's National Open Innovative Platform for Next-generation AI.

Mr. YANG is an engineering Ph.D. candidate of the Innovation Leadership Program at Tsinghua University, where he also received his Bachelor's and Master's degrees from the Department of Electronic Engineering. He holds 13 PCT patents, 2 Chinese invention patents, and 2 Chinese utility model patents.

He is currently also the Expert in the Industry Expert Consultant Library of Shenzhen Stock Exchange, the Expert in the AI Expert Library of MIIT, Member of the Overall Expert Group on AI Social Experimentation of MOST, Managing Director of the Chinese Association for Artificial Intelligence, Vice President of the Chinese Association of Young Scientists and Technologists, Executive Committee Member of China Committee of 100 of Digital Economy of the Chinese Institute of Electronics, and Committee Member of the VR/AR Technology and Industry Branch of the Chinese Institute of Electronics. He was awarded "Beijing Outstanding Young Talent" in 2016.

SociAl Contract for 21st Century

By Danil Kerimi

It is safe to say that 2020 has been nothing like the year we were expecting. Plans have been disrupted, travel postponed, and meetings conducted exclusively online. Inter- and intra-national, urban and rural, have and have-nots fault lines became clearer than ever before. As the first truly global digital pandemic, the present crisis is a perfect stress test for our policies, institutions and strategies. It forced us to throw away old assumptions, ask difficult questions and take nothing for granted.

Yet, the greatest advances in productivity growth normally follow periods of great disruption. While focusing on the short-term disruptions of the crisis and its mid-term consequences, we cannot lose track of long-term challenges facing humankind.

Urbanization, globalization and digitization were the megatrends that shaped societies in the last decades of 20th century and early 21st. Nowhere was this transformation felt more than in Asia.

2020 and its first digital global pandemic marked a new milestone in this evolution. Not only did the crisis force us to rethink the livability of many cities around the world, it marked a new so far qualitatively but soon quantitively visible new wave of globalization.

2020 also marked the official arrival of techlash in many parts of Asia. Algorithms enabled shifts in power dynamics were scrutinized by many governments and citizens around the world. The wonky AI policy debates escaped the halls of universities and think-tanks to run wild in the ministerial corridors, corporate boardrooms, street markets and private living rooms as more and more AI-enabled devices and services entered the lives of millions.

It is customary to differentiate between agricultural and industrial revolutions and their impact on our systems of production, consumption which in turn affect our living environment, economies and political systems. In the past centuries we have witnessed at least three agricultural and four industrials revolutions. With intersection of physical, cyber and bio worlds, breakthroughs in and enabled by AI, happening against the backdrop of growing population, warming planet, competing superpowers and growing consumer consciousness, we will be seeing the emergence of a new agri-aqua-industrial revolution. 2021 will be the year, when the social contract will be rewritten in front of our eyes. This time it will be AI-enabled and AI-enabling.

ABOUT THE AUTHOR

Danil Kerimi

Danil Kerimi is an experienced technology executive skilled in the nuanced creation and articulation of market sensitive agenda and messaging in complex multijurisdictional arenas. He is an inspirational leader who builds high-performing teams in digital transformation. His extensive experience in the developed, emerging and frontier markets helps Danil to deliver impactful projects in corporate strategy, business development, public and investor relations, ESG, tech and economic diplomacy.

As an expert in investor/internal/external/government relations, Danil is working with national, regional and municipal governments, large enterprises, start-up and investor communities on accelerating and securing digital transition around the world and applying newfound capabilities for solving biggest challenges in health, food and environmental ecosystems. Danil is most passionate about unlocking both the entrepreneurial and intrapreneurial potential of individuals and organizations to drive positive change in the world.

Who Should Own Our Data? Data Ownership & Policy

By Steven Hoffman

With AI's ability to process and make use of vast amounts of data, policymakers around the world are looking to limit the abuse of people's personal data. A growing number of lawmakers are proposing that individuals own and have exclusive property rights in the data they generate online. This sounds good in theory, but does it solve the fundamental problem of an asymmetrical marketplace, where large companies profit from this data in ways that may harm individuals and society?

To rectify this problem, many policymakers believe that that social media and other companies should pay people to license their data. This way, the people who generate the data can receive compensation and control how it's used. On the surface, this sounds like a great idea, but in practice, most people will wind up trading away their privacy rights for very little in the way of monetary compensation. The fact is that the vast majority of people won't read the fine print of these licensing terms, and even if they do, they aren't in a position to understand the repercussions. This means the idea that they are actually in control, will just be an illusion.

There is also the issue of injecting friction into the flow of information. The world's economy thrives on the rapid exchange of data. If we bog down the system with a lot of complex regulations, it stands to seriously impact all the businesses and consumers that rely on this data. Everything from logistics, healthcare and advertising to e-commerce, media, and public services will be impacted. The reason people's personal information is in such high demand is because it is incredibly valuable to society and has so many different uses.

Think about how many ways a person's name, address, birthdate, and other basic data are used. What does it mean to license and restrict this information? Do people need to sign licensing agreements with every app they use? And who actually owns this data? The fact is that data is information, and to treat it like a commodity fundamentally alters its nature and will impact our entire social system. Before we take this radical step, we need to assess the impact of this type of legislation. Right now, most of the public and private services we take for granted rely on the free flow of this data.

Instead, we should focus on the core problem, which is not whether people get paid a small licensing fee but regulating how this data is shared and used. Most people care about their privacy, and they want to make sure their data does not fall into the wrong hands and wind up negatively, impacting their lives or society.

We need to recognize that the problem isn't that people are not getting paid for their data. When someone signs up for Facebook, it is free precisely because people are agreeing to let the company use their data to make a profit. Paying people an additional licensing fee to grant Facebook the right to use it will not change this. The problem comes later, when companies do not place adequate controls on how this data is handled and managed.

Policymakers should focus on curbing the abuses, rather than commoditizing the data. Most countries already

have strict regulations on how people's medical data can be used and shared. We need similar regulations on how companies can use and share personal data. In the end, good legislation will need to find a balance between protecting people's personal privacy and enabling the free flow of information. Only in this way can we advance both individual and social interests, while not impeding economic growth and constraining the services we all depend upon.

ABOUT THE AUTHOR

Steven Hoffman

Steven Hoffman, or Captain Hoff as he is called in Silicon Valley, is the Chairman & CEO of Founders Space (FoundersSpace.com), one of the world's leading incubators and accelerators. He is also an angel investor, limited partner at August Capital, serial entrepreneur, and author of several award-winning books. These include *Make Elephants Fly* (MakeElephantsFly.com), *Surviving a Startup* (SurvivingAStartup.com), and *The Five Forces* (FiveForcesBook.com).

The Governance of AI in Digital Healthcare for a Post-Pandemic World Requires Multistakeholder Partnerships

By Omar Costilla-Reyes

We have experienced in the last few years an incredible increase in activity in the digital health space in the United States and the world. Startups, research, and investment in the space have never been this active. The accelerated speed of adoption has been facilitated by the ongoing global pandemic and the need to find remote and innovative solutions in healthcare.

The regulatory body that approves the use of digital health technologies for clinical applications is the Food and Drug Administration (FDA) in the United States. The first set of products that went through the FDA "de novo" certification model for digital health software includes "digital therapeutics" those are "evidence-based therapeutic interventions to patients that are driven by high-quality software programs to prevent, manage, or treat a broad spectrum of physical, mental, and behavioral conditions" as defined by the digital therapeutics alliance; a non-profit trade association of industry leaders and stakeholders in digital therapeutics. FDA-approved solutions include drug addiction and insomnia software, also the first video-game-based software solution for the treatment of attention deficit disorder in children.

While a clinical digital intervention offers great promise to alleviate major healthcare issues, its acceptance, integration with the healthcare system and policy reform have been difficult and are currently in the early stages. Some of the digital therapeutics model goes after a drug-like reimbursement model; not widely accepted by policymakers and insurers due to its digital nature.

The regulatory pathway for digital therapeutics is also difficult to follow for new players in the space. Clinical regulatory approval requires evidence-based products that can only be tested by running large, costly, and time-consuming clinical trials.

Current solutions in the digital therapeutics space focus mainly on digitalizing evidence-based healthcare solutions as interventions, for example, by digitalizing cognitive behavioral therapy programs for mental health.

The new generation of digital healthcare products envision an aim of incorporating artificial intelligence that learn and change with longitudinal data obtained from mobile phones and wearable data to improve the solution's performance and personalization over time. This is an early-stage area of research that requires partnerships in the research, governmental, and industrial sectors.

There is also progress in the research of digital healthcare. In the mental health space, for example, there are two major digital health initiatives: in the United States, the University of California Los Angeles has launched a depression grand challenge to research objective measurements of depressive symptoms with mobile devices, this in a large cohort of patients in partnership with Apple Inc. While in the European Union the radar central nervous system initiative aims to explore the potential of wearable and mobile devices to prevent and treat depression, multiple sclerosis, and epilepsy, also in partnership with Janssen Inc.

The FDA is currently working on defining how to legislate Software as a Medical Device (SaMD) that its core component consists of artificial intelligence that learns and adapts over time according to patient's needs. The FDA is running now a pre-certification model that constantly evaluates companies and products to obtain regulatory approval to offer SaMD AI solutions. The evaluation includes patient safety, product quality, clinical responsibility, cybersecurity, and proactive culture.

Countries around the world should use the United states SaMD certification model as inspiration to create its legislation, considering the differences in cultural, societal, and economical contexts.

Policymakers have the challenge to allow freedom of innovation in the digital healthcare space, but at the same time, guarantee patient safety and product efficacy. The opportunity for digital therapeutics to provide effective digital solutions in healthcare is unique and timely for a post-pandemic world, a place where the digital adoption of healthcare solutions has increased like never seen before.

To revolutionize global digital healthcare, it is required to align several scientific disciplines such as psychology and computer science, while at the same time, working closely with policymakers to create effective and safe solutions that offer an opportunity for a large positive impact on the lives of millions around the world.

ABOUT THE AUTHOR

Omar Costilla-Reyes

Omar Costilla-Reyes is a researcher at the Institute for Medical and Engineering Science (IMES) at the Massachusetts Institute of Technology (MIT). Omar is the current vice-president of the postdoctoral association at MIT. Omar's interests are in healthcare, artificial intelligence, education, and Latin America. In his current research, Omar focuses on designing the next generation of digital validated artificial intelligence solutions in healthcare. Omar obtained his Ph.D. from the University of Manchester in England. His Ph.D. thesis work focused on artificial intelligence in healthcare and biometrics. Omar is the founder and president of the Artificial Intelligence Latin American Summit. The goal of the summit is to incentivize leaders of the Latin American region in government, industry, and academia to invest and develop Artificial Intelligence for social benefit.

Part III Interdisciplinary Analyses from Professional Researchers

Emerging Institutions for AI Governance

By Allan Dafoe and Alexis Carlier

Much AI governance work involves preparation for a constitutional moment: an opportunity to create long-lasting, decision-shaping, institutions. Doing this well is a formidable task. It requires a fine balance. Institutions must constrain our actions today, yet preserve the freedom to adjust as circumstances change and our wisdom grows. Despite this difficulty, decisions must be made - and in 2020, multiple noteworthy institutions emerged. The research community has an opportunity to inform such decisions.

NeurIPS - one of the leading AI conferences - now requires that all paper submissions include a statement of the "potential broader impact of their work, including its ethical aspects and future societal consequences". This exciting institutional innovation could enhance the machine learning community's engagement and expertise with technology governance. At the Centre for the Governance of AI (GovAI), we produced an (unofficial) guide to writing the NeurIPS impact statement. In a Nature Machine Intelligence paper, *Institutionalizing Ethics in AI through Broader Impact Requirements*, we recommend measures to increase the chances that this innovation succeeds. GovAI affiliate Carolyn Ashurst ran a NeurIPS workshop examining how potential harmful impacts should affect the research community; GovAI Director Allan Dafoe, one of the co-authors of this article, discussed the challenges and opportunities of such a requirement on a plenary panel at EMNLP, a top natural language processing conference.

Facebook's independent Oversight board began its work. The board's decisions are meant to be binding, acting as a "Supreme Court" for Facebook's content moderation. This is a commendable initiative, representing a rare case of a technology company (voluntarily) exposing itself to external constraints so as to improve technology governance. Researchers can support this endeavor by developing expertise to evaluate and advise the board, while the wider community - the public, media, non-profits, Facebook employees - can help hold the board and company accountable to their promising intentions.

The Partnership on AI created the 'AI and Shared Prosperity Initiative', encouraging private AI actors to commit to an inclusive economic future. GovAI wrote the report which helped catalyze the initiative: The Windfall Clause: Distributing the Benefits of AI, which proposes a policy instrument to lessen AI-induced inequality. Lead author and former GovAI researcher Cullen O'Keefe (now at OpenAI) is a member of the initiative's research group, and GovAI affiliate Anton Korinek serves on the Steering Committee.

The aforementioned institutions are designed by humans. Bringing this about is one ambition (among many) of "Cooperative AI", a field recently outlined by Allan Dafoe, Thore Graepel, and other colleagues from DeepMind and elsewhere, in their paper *Open Problems in Cooperative AI* and associated *Nature* commentary and NeurIPS workshop. We see an opportunity for the field of AI to explicitly focus on cooperation; and since institutions are central to cooperation, a promising avenue for Cooperative AI research is institutional design. Humans could determine an institution's goals, while AI systems serve as design tools and infrastructure for building new innovative institutions.

We are excited to work together on these challenges with our colleagues in the AI governance community.

ABOUT THE AUTHOR

Allan Dafoe, Alexis Carlier

Allan Dafoe

Allan Dafoe is Associate Professor in the International Politics of AI and Director of the Centre for the Governance of AI at the Future of Humanity Institute, University of Oxford. His research examines the causes of great power war and the global politics surrounding transformative technologies, in particular concerning the risks from artificial intelligence. To help scientists better study these and other topics he also works on methods for causal inference and for promoting transparency.

Alexis Carlier

Alexis helps to manage the Centre for the Governance of AI. Previously, Alexis worked as a Governance Associate at the Consortium on the Landscape of AI Safety, collaborated on AI forecasting research with the Johns Hopkins Applied Physics Laboratory, and worked as a machine learning engineer at Sensyne Health.

Risk Management of AI Systems, But How?

By Jared T. Brown

The OECD's AI principles correctly state the potential risks from AI systems "... should be continually assessed and managed". But how should governments do that?

For a sense of scale of the governance challenge, consider some of the methods national governments have developed over decades to manage the risk of operating automobiles: regulations on who is allowed to operate automobiles and in what conditions (e.g., rules on licenses, intoxication levels, insurance requirements); the engineered design and safety features of automobiles (e.g., air bags, headlights); or continuous enforcement of rules on how to operate the automobile (e.g., traffic citations). And yet, despite being a technology well over 100 years old used across the globe, governments do not fully manage the negative consequences of automobiles, as accidents still result in over one million deaths annually.

And unlike automobile accidents, which have a relatively defined set of negative consequences, AI systems have a vast array of negative consequences that are much harder to identify and measure. As described in the European Union's White Paper of AI, these consequences are "both material (safety and health of individuals, including loss of life, damage to property) and immaterial (loss of privacy, limitations to the right of freedom of expression, human dignity, discrimination for instance in access to employment)". Further, risk management of AI systems will be especially difficult because there is an inherent "risk that AI's pursuit of its defined goals may diverge from the underlying or original human intent and cause unintended consequences — including those that negatively impact privacy, civil rights, civil liberties, confidentiality, security, and safety" (in the words of the official U.S. government regulatory guidelines for AI applications). As the Future of Life Institute (FLI) and our partners have captured in a multitude of formal and informal responses to government policy proposals issued throughout 2020 (futureoflife.org/policy-work), assessing and managing AI-related risk will be an extremely difficult and complex task.

While most governments will admirably try to manage the risk, they will not be able to do it properly without constant feedback from experts that understand the cutting-edge of increasingly capable, rapidly evolving AI systems. Thus, turning to 2021 and beyond, civil society organizations and AI experts will need to proactively engage with governments to help develop the policy equivalents of seat belt laws, speed limits, and traffic courts for managing the risk AI systems. In the United States, this will require robust engagement with the National Institute of Standards and Technology, which has been legally assigned the herculean task of developing a "voluntary risk management framework for trustworthy artificial intelligence systems". Elsewhere, in the European Union, experts will need to work with officials to further specify their proposal to conduct "conformity assessments for high-risk AI applications". Having largely taken the important step admitting that there is considerable risk to be managed, in 2021, we must now set to the task of developing the best ways of doing so.

ABOUT THE AUTHOR

Jared T. Brown

Jared Brown is the Senior Advisor for Government Affairs at the Future of Life Institute. He is also a Special Advisor for Government Affairs at the Global Catastrophic Risk Institute. He has spent his career working at the intersection of public policy and risk management, having previously served as an Analyst in Emergency Management and Homeland Security Policy at the U.S. Congressional Research Service and in homeland security at the U.S. Department of Transportation. He has earned a Master of Public Policy from Georgetown University and a B.S. in Social Psychology from the University of California, San Diego.

AI Governance for the People

By Petra Ahrweiler and Martin Neumann

AI governance is closely associated with the concept of "bureaucratic governance" famously described by Max Weber. According to Weber bureaucratic governance, in contrast to traditional or charismatic governance modes - legitimises political power by formal, legal, and rational reference mechanisms. Though Weber acknowledges the gains of bureaucratic governance as efficiency, objectivity, and rationality, he deplores the increased rationalisation in social life as "iron cage" that encages individuals in systemic control of a "polar night of icy darkness".

The frail relationship between subject and society was analysed as the main source for anomic social phenomena where the subject-society divide manifested itself most dramatically. Looking at crisis scenarios today, it is surprising to see the continuity of problems and drivers. Many contemporary crises are nurtured by the tension between subject and technology-based society. Because crisis phenomena became overtly complex, interconnected, and global, dealing with them would require concerted action across the globe.

At first sight, it seems obvious that algorithmic procedures are efficient, objective, and neutral, unbiased by human passion and follow the standards of formal rationality. Governance decisions, however, are about people: they are value decisions. Different societal value propositions concerning justice, equality, responsibility, human dignity and rights etc., are responsible for peculiarities of different governance systems across the globe. Social values are generated and enforced by beliefs, cultural orientations and deeply-rooted meaning structures of societies. Here it is where the "iron cage" becomes manifest in applying AI to governance: AI-based governance with its opaque algorithms does not yet reflect the complex value base of societal decision making in governance.

It is well-researched in the meantime that algorithmic decision making in governance can generate discrimination and unfairness, that it may lack accuracy and flexibility, and that it can lead to severe ethical challenges. AI-based governance is challenged by the same pitfalls of bureaucratic rationality leading to the modernisation crisis in nineteenth century Europe with regard to reflexive modernisation phenomena.

Thus, how should we escape the iron cage trap of bureaucratic governance? The answer might not be "less AI", but "better AI", namely AI governance for the people. Again, Max Weber can provide a starting point: for implementing cultural dynamics and the evolution of value propositions into AI-based governance, an option is to re-visit Max Weber's "Economic Ethics of the World Religions", extending the unit of analysis from religion to culture, and extending the focus of analysis from economics to society at large.

Using machine learning and NLP techniques can enable an inventory of knowledge corpora about specific value systems across the globe. The inventory of dynamically changing value systems can help to develop a set of indicators for measurable effects of meaning-making that can be identified by AI systems. In order to evaluate AI-based governance ex ante, distributed artificial intelligence systems and, in order to capture people's values, hermeneutics, can be further co-developed for integrating cultural sense-making and meaning structures into AI-based governance. With this, AI governance can become responsible, responsive, and deliberate as AI governance for the people.

ABOUT THE AUTHOR

Petra Ahrweiler, Martin Neumann

Petra Ahrweiler

Prof. Dr. Petra Ahrweiler leads TISSS Lab, the Technology and Innovation Sociology / Social Simulation Laboratory of Johannes Gutenberg University (JGU) at Mainz, Germany. At Free University Berlin, Germany, she received her PhD for a study on Artificial Intelligence, and got her habilitation at the University of Bielefeld, Germany, for a study on simulation in Science and Technology Studies.

Martin Neumann

Dr. Martin Neumann is research associate at the chair of Prof. Petra Ahrweiler. After a PhD. in history of science, he gained experience in agent-based social simulation with a focus on norms, crime, and conflict. Currently he is working in research projects on societal impacts of artificial intelligence, specifically in AI-based governance practices.

From Diversity to Decoloniality: A Critical Turn

By Malavika Jayaram

There is no shortage of frameworks for the governance of AI, yet their utility and legitimacy are not universally accepted. In addition to claims that the proliferation of principles and guidelines is about “ethics washing” rather than a commitment to real change, and that it furthers the case for industry self-regulation rather than the enforcement of legally binding rules, a more fundamental set of questions is emerging: about the values and norms that are embedded and reified in these frameworks, and - equally importantly - the ones that are not.

It has become increasingly common - fashionable even – to look to multiple sources of values, such as Shintoism or Ubuntu philosophy, in an attempt to make the AI governance discourse more “inclusive”. This trend, however well-intentioned, is often restricted to a cosmetic exercise, cherry-picking concepts and anecdotes without engaging more critically with the structures and histories that they embody, or with the dominant narratives and patterns of power that AI systems can internalize and amplify. These philosophies are often co-opted and remixed by scholars and actors with privilege, on stages and in spaces that either remain closed to the cultures and communities that they borrow from or appropriate, or that fetishize them to perform diversity.

A growing body of scholarship rejects the idea of “inclusion” as paternalistic and inequitable - Who is doing the including? Into what? On what terms? How is it entrenching and reproducing pre-existing power dynamics? - and offers decolonial theory as a lens to understand the residue of coloniser-colonised relationships that continues to shape and produce digital technologies including AI. Colonial histories were built on landgrabs, and on the exploitation and dispossession of material resources. The effects of colonialism endure even today, continuing the process of appropriation and extraction: the survival of these power dynamics and structures is referred to as coloniality. Scholars have conceptualized Data Colonialism as a digital emanation of these colonial practices, as human behaviour and interactions - represented by data - are mined and manipulated for material gain. The movement to decolonize data includes, among other things, efforts to preserve the sovereignty and autonomy of data that does not fit neatly into Western parameters, to have indigenous and marginalized communities control their information and their stories, and to decolonize research methodologies and practices.

To the extent that AI embeds and amplifies data colonialism at scale, it has a disproportionate impact on the production of meaning and the construction of reality, even of algorithmic “truth”. As such, it contains the capacity to diminish or erase other systems of knowledge production and sense-making. The turn to Decolonial AI is a conversation about resisting, undoing and providing alternatives to the dominant values, assumptions and biases that make AI systems oppressive, inequitable, and unsustainable. As scholars continue to identify and study sites of coloniality, such as predictive policing, welfare systems, and identity cards, which embed and perpetuate colonial histories and stereotypes about race, criminality, and poverty, the field of AI governance will expand from one that enshrines a narrow, homogenous idea of universal values, to one that supports a genuinely pluralistic set

of imaginaries and possibilities. This turn from diversity to decoloniality promises greater opportunities for a humane approach to living with, or without, AI.

ABOUT THE AUTHOR

Malavika Jayaram

Malavika is Practice Assistant Professor and Lee Kong Chian Fellow at SMU School of Law, Singapore. She is also a Faculty Associate at the Berkman Klein Center for Internet and Society at Harvard University (BKC), and the inaugural Executive Director of the Digital Asia Hub, an independent research think-tank incubated by BKC. A technology lawyer for over 15 years, she practised law at Allen & Overy, and was Vice President and Technology Counsel at Citigroup, both in London.

A graduate of the National Law School of India, Malavika has an LL.M. from Northwestern University, Chicago. She was a Visiting Scholar at the Annenberg School for Communication, University of Pennsylvania, and has had fellowships at the University of Sydney and the Institute for Technology & Society, Rio de Janeiro. She taught India's first course on information technology and law in 1997.

Malavika was previously a member of the Executive Committee of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, and an Associate Fellow with Chatham House (the Royal Institute of International Affairs), as part of its Asia-Pacific Programme. She was also a member of the High-level Expert Advisory Group to the OECD project, "Going Digital: Making the Transformation Work for Growth and Well-being".

Governing Artificial Intelligence: from Principles to Law

By Nathalie Smuha

If 2019 was the year that AI governance debates shifted "from principles to practice", 2020 marked the year that this shift took a new turn, namely "from principles to law". It became increasingly apparent that some of the risks posed by the development and use of AI are simply too substantial to be left to mere guidelines or to the self-regulatory goodwill of private actors. This realization was spurred not only by media coverage of problematic AI applications and by vocal civil society organizations, but also by private actors themselves, some of which openly began to ask for binding rules that provide legal certainty, while simultaneously enhancing citizens' trust. Consequently, policymakers started to examine more carefully the legal gaps impeding effective protection against the harmful effects of certain AI applications.

Thus, building on the work of its High-Level Expert Group on AI, the European Commission published a White Paper that maps a number of gaps in the EU legal order, and provides a blueprint for new regulation to be proposed in 2021. Furthermore, the Council of Europe's Ad Hoc Committee on AI (CAHAI) published a Feasibility Study that examines the potential elements of a legal framework for the development, design and application of AI, based on its standards in the field of human rights, democracy and the rule of law. It provides an overview of existing binding and non-binding instruments applicable to AI, lists their (dis)advantages and - on that basis - puts forward essential rights and obligations that could be included in a future binding instrument, such as an international convention.

Of course, it may well take years before these initiatives result in enforceable legislation. Moreover, they are still regional in scope, whereas many of the risks raised by AI require a global approach. In addition, further interdisciplinary research is needed to better identify, understand and mitigate AI's potential adverse effects. The work to secure an appropriate legal framework for AI that can protect citizens globally is thus far from over. However, the path on which the abovementioned policymakers embarked is encouraging, and can hopefully set an example for others to follow. Going forward, I would like to raise three points that should be kept in mind.

First, AI policies and regulations should embrace a holistic perspective. AI systems do not exist in isolation, but are part of a broader socio-technical environment. This environment encompasses the numerous actors and processes that are involved in the lifecycle of AI systems, as well as the data they use and the infrastructure they run on. Governing AI hence also requires appropriate measures to govern data flows and digital infrastructures.

Second, it must be ensured that the seats around the negotiation table for AI regulation are sufficiently diverse. Too often, the most marginalized individuals and communities are first in line to suffer the negative effects of harmful AI applications, thereby entrenching and deepening inequities and problematic power relationships. Therefore, not only those who have something to gain, but also those who have something to lose should be represented in AI governance debates and help shape AI's legal framework.

Finally, it is essential for policymakers to consider not only the individual harms that can arise from the use of AI systems, but also the collective and societal harms. When taking a long-term perspective, it becomes more visible how the delegation of human autonomy in not just one area but in ever more domains of our lives - accumulatively - risks shaking the foundations of our moral, democratic and societal infrastructures. Only by acknowledging this risk can we take steps to tackle it.

ABOUT THE AUTHOR

Nathalie Smuha

Nathalie Smuha is a researcher at the KU Leuven Faculty of Law and the Leuven.AI Institute, where she examines legal and ethical questions around Artificial Intelligence (AI) and other new technologies. Her research focuses particularly on the impact of AI on human rights and societal values. Nathalie regularly advises governments and international organizations on AI-related policy matters. She is involved in the Council of Europe's Ad Hoc Committee on Artificial Intelligence (CAHAI) as an independent expert, and a member of the OECD's Network of Experts on AI (ONE AI). She also worked at the European Commission (DG Connect) where she coordinated the work of the High-Level Expert Group on Artificial Intelligence and contributed to EU policy-making on AI. Nathalie is a qualified attorney at the New York Bar, and previously worked as a lawyer in an international law firm. She holds degrees in Law and Philosophy from the KU Leuven, and an LL.M. from the University of Chicago School of Law.

The COVID-19 Pandemic and the Geopolitics of AI Development

By Wendell Wallach

In addition to a tragic loss of lives and stressed healthcare systems, the 2020 pandemic forced the shutdown of vast segments of local, national, and international economic activity. And yet, as work, education, and play moved online, there was a dramatic acceleration in the growth of the digital economy, including investments in infrastructure and AI research. Heads of companies, such as the Microsoft Corporation, were surprised to see that goals projected out 3 - 5 years were being surpassed in the second half of the year.

During this period, work on AI ethics and governance progressed from the elucidation of principles for guiding AI research and systems design, to procedures and policies for the deployment of AI-based technologies. In China, the continuing rollout of a digital credit system coincided with a wider discussion of AI ethics focused primarily on privacy concerns and the abuse of personal data. In Europe and the U.S. civil liberty advocates led efforts to place limits on the use of facial recognition technologies. These initiatives were largely successful due to increasing awareness that the algorithms used in facial recognition systems are alarmingly high in yielding inaccurate results, and therefore should not be used for law enforcement and other critical activities.

During 2020, the bifurcation of the digital economy into two spheres of influence, championed by the Trump administration, proceeded with little evidence of abatement. The U.S. has been weakened as a geopolitical force, as once nascent authoritarian and anti-democratic tendencies coalesced into an active political force under Donald Trump. While the U.S. is deeply divided politically, one area in which the public generally concurs is its distrust of China. Indeed, the Biden administration has an opportunity to reset U.S. policy towards China, but it will be difficult to alter the present course in light of the continued U.S. public's distrust regarding Chinese intentions.

China's unique political system demonstrated an ability to effectively address the public health crisis posed by COVID-19. Simultaneously, the politicization of the pandemic in the U.S. led to horrific death tolls and stressed hospitals in a country that prides itself on being a leader in medical research and healthcare. This, in turn, it revealed weaknesses in American democratic institutions.

I have long pointed out the ways in which Americans do not understand China and the Chinese misunderstand the U.S. Furthermore, I have been underscoring the serious need for international cooperation in the development of emerging technologies and in climate policy, to avoid disastrous consequences that will seriously undermine prospects for current and future generations. While there has been an increase in countries giving minimal expression to the need for cooperation and a degree of coordination in the development of AI and other emerging technologies, there has also been little tangible progress. Countries in the West increasingly coalesce into governance alliances in which China and Russia are either excluded or invited to join, only if they abide by pre-established rules. There is active debate among many in the West as to whether sharing standards and AI policies with China is in their own country's best interest. Competing standards for infrastructure, particularly in the rollout

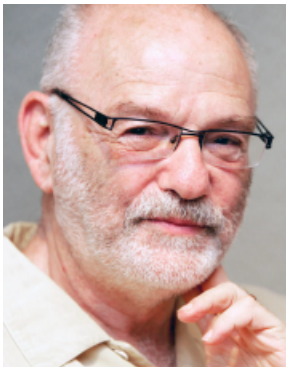
of 5G, will again sever the world into two spheres of influence with potentially ruinous outcomes.

The advent of the Biden administration still offers the possibility of a reset for U.S. relations with China, and cooperation on serious issues such as cybersecurity, biosecurity, AI-enabled weaponry, and geoengineering. But for this to happen, all factions will need to take active steps towards that goal in 2021.

Making conscious tradeoffs between near- and longer-term national benefits and geopolitical stability is always difficult, regardless of the circumstances. Nonetheless, in this case, as we look towards the future, I would argue, it is essential for international security to ensure pandemic recovery, to enable a collective response to global warming, and to mitigate undesirable societal impacts arising out of the accelerating digital transformation.

ABOUT THE AUTHOR

Wendell Wallach



Wendell Wallach is a scholar at Yale University's Interdisciplinary Center for Bioethics, where he chaired Technology and Ethics studies for eleven years. He is also a senior advisor to The Hastings Center and a senior fellow at the Carnegie Council for Ethics in International Affairs. His latest book, a primer on emerging technologies, is entitled, *A Dangerous Master: How to Keep Technology from Slipping beyond Our Control.* In addition, he co-authored (with Colin Allen) *Moral Machines: Teaching Robots Right From Wrong*. The eight volume *Library of Essays on the Ethics of Emerging Technologies* (edited by Wallach) was published by Routledge in Winter 2017. He received the World Technology Award for Ethics in 2014 and for Journalism and Media in 2015, as well as a Fulbright Research Chair at the University of Ottawa in 2015-2016. The World Economic Forum appointed Mr. Wallach co-chair of its Global Future Council on Technology, Values, and Policy for the 2016-2018 term, and he is presently a member of their AI Council. Wendell is the lead organizer for the 1st International Congress for the Governance of AI (ICGAI).

Mitigating Legacies of Inequality: Global South Participation in AI Governance

By Marie-Therese Png

In a historic year defined by the COVID-19 pandemic and global protests for racial justice, it is no surprise that the historic arc of structural inequality has been spotlighted in AI governance.

AI governance initiatives recognise their responsibilities in ensuring AI deployment and regulation do not "lock-in" intra and international inequalities. Accordingly, we see greater efforts towards heterogeneous representation in constructing guardrails and principles. For example, the UN Secretary-General's 2020 Roadmap for Digital Cooperation called for greater participation of Africa, South America and Central Asia in AI governance, to rebalance the discourse monopoly held by North America, Europe and China.

As Jasanoff and Hurlbut remind us, we must be aware of "who sits at the table, what questions and concerns are sidelined and what power asymmetries are shaping the terms of debate". Strategies for harms mitigation cannot be defined by those who benefit from AI systems, but must be defined by those who know and experience the costs.

For example, Dr. Eugenio Vargas Garcia identifies that though lethal autonomous weapons are likely to be deployed first in conflict zones in the developing world, many of these regions are completely absent in the regulatory discourse. Exporting harms of AI systems to marginalised populations or low/middle-income countries via beta-testing is well documented.

In order to convene effective stakeholder coalitions to mitigate harms, we must first identify structural barriers to meaningful political participation of Global South stakeholders. If state, and civil society, actors cannot act unilaterally to protect their interests, or forge contextualised governance, AI governance initiatives will enact what Professor Ruha Benjamin terms "techno-benevolence" - interventions that intend to address inequalities, but instead reproduce or deepen dependency and extractivism. Moreover, policies will be replicated across jurisdictions in ways that are incompatible with the goals and constraints of developing countries.

In DeepMind's 2020 Decolonial AI paper, my co-authors Dr. Shakir Mohamed, Dr. William Isaac, and I posit that we cannot understand present AI inequalities, or anticipate their futures, without looking at their historic trajectories. The first-mover advantages and exclusionary path dependencies we see today are, in part, living relics from our colonial histories. When we seek to increase Global South representation, we recognise that "Global South" describes a geography which emerged from legacies of colonialism.

The existence of coalitions such as the G77 and Non Aligned Movement, key in the decolonisation and independence movements in Africa, Asia, Latin America and other regions, affirm the continuities of colonialism in contemporary global inequality. Today, these coalitions represent two-thirds of UN membership, and 55% of the global population. They are a platform for Global South countries to articulate collective interests and promote South-South cooperation.

In 2020 Professor Ulises Mejias proposed a Non Aligned

Technology Movement, with a primary goal of transitioning from technologies that reinforce dependency dynamics, to technologies that support the self-determination of developing countries. Such ethos is a pre-requisite for meaningful Global South participation, robust discussions of risk mitigation, and preventing locked-in inequalities.

ABOUT THE AUTHOR

Marie-Therese Png

Marie-Therese Png is a PhD candidate at the Oxford Internet Institute, researching political decolonisation efforts in the governance of autonomous decision making systems. She was previously Technology Advisor to the UN Secretary General's High Level Panel on Digital Cooperation, working on technology policy domains including digital inclusion, lethal autonomous weapons, cybersecurity, and algorithmic racial discrimination with special focus on multi-stakeholder coalition building and advocating for geographic representation. Marie-Therese has worked with Google DeepMind on AI value alignment, co-authoring the peer-reviewed academic paper *Decolonial Theory as Socio-technical Foresight in Artificial Intelligence Research*, and is a member of the IEEE Ethically Aligned AI Classical Ethics Committee.

Her research includes case study work on facial recognition in Singapore's smart city ecosystem with a lens of digital human rights and non-Western perspectives in AI governance, and has worked across technology ethics and systemic harms at the MIT Media Lab and the Harvard AI Initiative. Marie-Therese was a co-organiser in the NeurIPS Resistance AI workshop, the MIT BioSummit biohacking movement, and the first AI Roundtable at the World Government Forum. Marie-Therese holds an undergraduate in Evolutionary Biology & Social Sciences from Oxford, and a Master's in Developmental Cognition and Intergroup Conflict from Harvard.

Artificial Intelligence Needs More Natural Intelligence

By Markus Knauff

When AI was born, it was mainly a joint enterprise of computer science and psychology. The goal was to build machines that think like humans. The interdisciplinary collaboration even resulted in a new academic discipline, cognitive science, which studies information processing in all kinds of biological and technical systems. Today, the huge commercial success of deep learning has pushed this collaboration into the background of AI. In fact, the field is becoming increasingly anti-psychological. Deep learning systems are inspired by human learning, but today they learn in a very different way than humans. As a result, they can be easily fooled and produce many errors. Nevertheless, they are successful because they can discover latent patterns of relevance by analysing large amounts of data generated by humans. However, the limitations of this approach are already on the horizon. One reason is the devastating inability of AI systems to reason and draw inferences. Consider the following inferences:

1. If it rains, the street gets wet.
   It rains.
   Therefore, the street gets wet.

2. The spy is in Berlin or in Paris, but not both.
   The spy is in Paris.
   Therefore, he is not in Berlin.

Solving such inference lies in the core of intelligence and is trivial for most human beings. Yet, if we ask Alexa, Siri, or other virtual assistant AI systems you get answers that are complete nonsense. It is hard to imagine AI making further progress if it is unable to solve such easy problems. Here are some results from cognitive psychology that can help to build future AI system that comes closer to human-level intelligence:

1. When people think about the world, they mentally simulate real, hypothetical, or imaginary situations.
2. Human reasoning is defeasible, i.e., people retract previously drawn conclusions in light of new evidence.
3. For humans, logical connectives such as if, then, all, some, none, etc., have different meanings than in classical logic.
4. Humans often use shortcuts and heuristics that lead to useful but logically invalid conclusions.
5. People can generate conclusions on their own rather than just evaluate them.
6. The discovery of an inconsistency often causes humans to abandon a previous belief. In this way, the cognitive system avoids the explosion of conclusions.
7. When multiple conclusions are possible, human reasoners prefer just one of them, but systematically ignore others. This is an instance of the principal of cognitive economy that guides many thinking, reasoning, and decision-making processes in human beings.

AI should incorporate such insights from cognitive psychology to develop systems that interact with humans and produce results that are understood and accepted by the user. Cognitive scientists have begun to implement computational reasoning systems that reflect these core principles of natural intelligence. AI should not lag behind these developments and become more psychological again to make technical systems more intelligent.

ABOUT THE AUTHOR

Markus Knauff

Markus Knauff is Professor (and formally Chair) of Experimental Psychology and Cognitive Science at the University of Giessen. Prior to his current position, he worked at the University of Freiburg, Princeton University, and the Max Planck Institute for Biological Cybernetics. He has been President of the German Cognitive Science Society, Associate Editor of Cognitive Science, and Chair of the 35th Annual Meeting of the Cognitive Science Society in Berlin. From 2011 to 2018, he was Director of the Priority Program News Frameworks of Rationality, in which the German Research Foundation (DFG) funded 15 research projects from psychology, AI, philosophy, and logic to understand the nature of human rationality. Knauff's research focuses on reasoning and rationality in natural and artificial cognitive systems. Methodologically he combines psychological experiments, computational modeling, and neuroimaging to understand the neural correlates of cognition and behavior. He worked on AI projects for many years, but then returned to human cognition as AI became increasingly anti-psychological. Now he sees the potential (and need) of re-psychologizing AI. His most recent publications include: *Space to Reason: A Spatial Theory of Human Thought* (2013) and *The Handbook of Rationality* (2021, together with the Philosopher W. Spohn), both with MIT Press.

Limits of Risk Based Frameworks in Developing Countries

By Urvashi Aneja

Many countries around the world, including India, are developing risk-based frameworks for the governance of AI. Risk based frameworks are perhaps appropriate to fuel innovation. But they are inappropriate for developing countries like India, where AI is viewed as a tool to address complex development and governance challenges. The stakes and trade-offs are different for developing countries because emerging technologies like AI are shaping development and state building trajectories. Low levels of regulatory and institutional capacity pose further challenges to the suitability of risk based approaches. Risk based approaches can create regulatory blind-spots with regard to disparate impacts for vulnerable populations and systemic risks. Assessing risk is not an objective exercise; it is deeply embedded in socio-cultural values and priorities. Risk based approaches also face methodological and epistemic challenges - even while some AI applications have a low risk, their cumulative effect could be large. While these concerns may be less paramount from the perspective of enabling innovation, they are certainly crucial from a development perspective.

Part of the problem for regulators around the world, including India, has been to establish a threshold for regulatory intervention. Risk based approaches are tempting in this regard, but to work, there needs to be open, inclusive, and transparent dialogue around risk identification and assessment. For this process to be meaningful, it is essential that civil society has the knowledge and capacity to evaluate the impact of AI; transparency and expertise are two sides of the same coin. These capacities are currently limited in India, and greater investments are needed in interdisciplinary research and public communication. Trust in judicial systems and institutions is also paramount - the absence of adequate grievance redressal mechanisms for many digitally enabled governance interventions in India do not bode well for building such trust.

Rather than thinking of AI governance in terms of specific high or low risk products and services, it is more fruitful to think of AI as a field of research - how we enable more responsible AI research and innovation? It is also helpful to adopt an infrastructural lens when thinking about AI governance. This focuses our attention on a wider range of issues that need to be governed - from the political economy of AI innovation trajectories, to the invisible labor enabling AI growth, to the societal impacts of AI. It also helps establish a certain set of values for anchoring or steering AI governance.

Finally, ethical frameworks may be inadequate for industry self-regulation, but at a societal level, we need to have far greater conversations about the ethics of automated and algorithmic decision making - we need to make important societal choices about where and how we want to introduce AI systems. At a time of growing surveillance and authoritarianism around the world, drawing a clear red line on the use of automated facial recognition and emotional recognition systems, by public and private actors, should be a priority.

ABOUT THE AUTHOR

Urvashi Aneja

Urvashi Aneja is Founding Director of Tandem Research, an independent research collective based in India, that generates policy insights at the interface of technology, society, and sustainability. Her work focuses on the ethics, political economy, and governance of emerging digital technologies in the Global South. She is also Associate Fellow at Chatham House.

Part IV Global Efforts from the International Community

AI Governance in 2020: Toolkit for the Responsible Use of AI by Law Enforcement

By Irakli Beridze

With the worst pandemic the world has seen in more than a century, 2020 has undoubtedly been a life-changing year, with political, social, and cultural upheaval across the world. At the same time, years of technology advancements seemed to happen in mere months and the first COVID-19 vaccines appeared in a record time. In fact, Artificial Intelligence (AI) was one of the technologies responsible for that achievement, contributing to speed up the development of a messenger Ribonucleic Acid (mRNA) based vaccine.

AI adoption was indeed accelerated during the COVID-19 outbreak, not only in medical research but also to restrict the movement of populations, causing some controversies. From contact tracing apps to facial recognition cameras that monitor travellers' temperature, the use of AI for tracking and surveillance raised concerns of fundamental freedoms, such as the right to privacy. While the application of technology can play an important role in containing the spread of the virus, the use of AI must stay proportionate, necessary and legitimate. To avoid potential pitfalls that could undermine fundamental rights, as well as infringe public trust in national institutions, governments should strive to advance AI governance, and guarantee this technology is developed for the good of societies.

Attending to the specificities of every sector and the respective different technical solutions AI can provide, a global AI governance can not curb possible harmful effects in all disciplines. Although there are general principles of Ethics, Fairness, Accountability and Transparency that must be cross-cutting, sectoral initiatives might be more useful and effective.

The Centre for AI and Robotics of the United Nations Interregional Crime and Justice Research Institute (UNICRI) seeks to support law enforcement agencies and other key stakeholders in the criminal justice system, to understand the risks and benefits of AI, and exploring their use for contributing to a future free of violence and crime. Recognizing that the use of AI in law enforcement is often a highly sensitive and controversial subject, and there is no specific guidance on this matter, the value in advancing AI governance for law enforcement has become increasingly prominent. In this sense, the UNICRI Centre for AI and Robotics, together with INTERPOL's Innovation Centre, have undertaken to develop an operationally oriented Toolkit for the Responsible use of AI by Law Enforcement, that

can support and guide in the design, development and deployment of AI in a responsible manner. This ambitious goal was set up and fuelled by several discussions at the INTERPOL-UNICRI Global Meetings on AI for Law Enforcement that brought together experts from all over the world. The last edition, in November 2020, convened more than 600 participants from law enforcement, academia and industry to share experiences, learn from one another and contribute to shaping this Toolkit. 2020 was not all gloom and doom, progress was made in many fields, by many actors, including in this growing and important AI governance field. With a long road and much work ahead, we are nevertheless moving in the direction of responsible AI for law enforcement.

ABOUT THE AUTHOR

Irakli Beridze

More than 20 years of experience in leading multilateral negotiations, developing stakeholder engagement programmes with governments, UN agencies, international organisations, private industry and corporations, think tanks, civil society, foundations, academia, and other partners on an international level.

Mr Beridze is advising governments and international organizations on numerous issues related to international security, scientific and technological developments, emerging technologies, innovation and disruptive potential of new technologies, particularly on the issue of crime prevention, criminal justice and security. He is supporting governments worldwide on the strategies, action plans, roadmaps and policy papers on AI.

Since 2014, he has initiated and managed one of the first United Nations Programmes on AI, initiating and organizing a number of high-level events at the United Nations General Assembly, and other international organizations and finding synergies with traditional threats and risks, as well as identifying solutions that AI can contribute to the achievement of the United Nations Sustainable Development Goals.

He is a member of various international task forces, including the World Economic Forum's Global Artificial Intelligence Council, the UN High-level panel for digital cooperation, the High-Level Expert Group on Artificial Intelligence of the European Commission. He is frequently lecturing and speaking on the subjects related to technological development, exponential technologies, artificial intelligence and robotics and international security. He has numerous publications in international journals and magazines and frequently quoted in media on the issues related to AI.

Irakli Beridze is an International Gender Champion supporting the IGC Panel Parity Pledge. He is also recipient of recognition on the awarding of the Nobel Peace Prize to the OPCW in 2013.

Global Cooperation on AI Governance: Let's Do Better in 2021

By Danit Gal

As countries continue to isolate in attempts to curb the COVID-19 pandemic and many industries grind to a halt, AI research, development, and applications continue to accelerate, largely unabated. This further exacerbates two pressing AI governance challenges: 1) growing global fragmentation in the use and regulation of AI, and 2) the widening gap between rapid AI deployment and lagging regulation. A good starting point to address these challenges is inclusive global cooperation on AI governance. But that is far easier said than done.

In many respects, 2019 and 2020 were good years for government-level global cooperation on AI governance, as international institutions stepped up to respond to the abovementioned challenges. The 2019 OECD AI Principles, subsequently informing the G20's AI Principles, were the first successful instance of such cooperation on AI governance. Established in June 2020, the Global Partnership on AI serves as an AI cooperation platform for like-minded countries with 19 state-level members to date. On the same month, the United Nations' Secretary General announced he will create a multi-stakeholder advisory body on global AI cooperation, including member-states.

While it remains to be seen how these initiatives continue to develop, existing structural barriers prevent them from becoming truly global. In all instances, equitable Global South representation, and that of many other marginalized groups, is sorely missing. Existing global cooperation initiatives on AI governance are built by developed and often Western countries that then seek to engage others. By design, they cater to the values and interests of the selected few who are already well-represented in datasets used to train selectively beneficial algorithms.

Failing to engage the underrepresented global majority bodes ill for any ambition of curbing global fragmentation. It leaves the majority of regulators and users from developing countries lagging behind the accelerating applications of AI, those already regulating AI, and those benefiting from AI regulation. As is often the case with new technologies, AI colonialism is increasingly accompanied by AI regulation colonialism, further eroding the agency and sovereignty of those who need it most. More so than imported technologies like AI, imported policies and regulations rarely acculturate.

We must do better in 2021. Internationally, global AI governance cooperation initiatives must equitably engage Global South entities and underrepresented communities, avoiding tokenization. Regionally, AI governance initiatives must invest in shared techno-regulatory capacity building and coordination. Domestically, regulators must consult the public to ensure diversity, inclusion, and feasibility before aligning regionally and internationally. Instead of a centralized approach to global cooperation that amplifies existing power structures, a decentralized, multi-leveled cooperation approach will empower and benefit a far larger proportion of those in need of AI governance.

We will do better in 2021 if we make room for many more people at the AI governance table, internationally, regionally, and domestically. Our collective ability to ensure AI goes well for the whole of humanity depends,

to a large extent, on our ability to govern its development, deployment, and use. The more global cooperation on AI governance becomes, the better.

ABOUT THE AUTHOR

Danit Gal

Danit Gal is Associate Fellow at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. She is interested in technology ethics, geopolitics, governance, safety, and security. Previously, she was Technology Advisor at the United Nations, leading work on AI in the implementation of the United Nations Secretary-General's Roadmap for Digital Cooperation. Danit serves as the vice chair of the P7009 IEEE standard on the Fail-Safe Design of Autonomous and Semi-Autonomous Systems. She also serves on the Executive Committee of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, and as Founding Editor and Editorial Board member of Springer's AI and Ethics journal, Advisory Board Member of the AI & Equality initiative at the Carnegie Council for Ethics in International Affairs, Executive Committee Member of the AI4SDGs Cooperation Network at the Beijing Academy of AI, and Advisory Board Member of EPSRC and UKRI's Trustworthy Autonomous Systems Verifiability Node.

AI in Pandemic Response: Realising the Promise

By Seán Ó hÉigeartaigh

Responding to COVID-19 has galvanised the AI research and governance communities in 2020, as it has the world at large. AI and digital technologies were touted early on in many quarters as being key to the response; and indeed, the early signs were promising. Blue Dot, an AI-based outbreak identification and tracking system, provided one of the first warnings of the outbreak in January. Thousands of papers described AI-based applications across the prediction and response process, from outbreak modelling, to drug discovery, to hospital logistics planning, to symptom analysis, to surveillance in support of quarantining.

A year in, however, it is clear that the role for AI has been somewhat limited. Why? I have been part of the Global Partnership on AI's AI and Pandemics working group, which has spent 6 months analysing applications, limitations, bottlenecks and solutions. I offer some individual reflections:

1. Hype vs boring reality. Technology is wonderful, but the core of pandemic response remains tried and trusted techniques – investments in public health, including sufficient PPE and hospital health capacity; (manual) contact tracing; quarantining, and so forth. AI can play a role in supporting this, but for now still a limited one. Indeed, over-focusing on technological solutions can divert attention and resources away from the basics.

2. A new challenge. AI, and particularly ML systems, work well when we can learn from the past to predict or act in the future. COVID-19 was a new disease, with new symptoms affecting populations in a novel way, and creating a relatively novel set of challenges and pressures on society. This makes it more difficult to train and deploy AI systems with confidence in many contexts. In particular:

3. Data. Many potential AI applications, from deep learning-based analysis of lung CT scans, to predicting population health outcomes, need large, diverse, representative datasets to be trained to sufficiently robust performance. In practice, data has been sparser, and scattered across research groups and countries.

4. Ethics and governance. Researchers and governments pursuing AI techniques, and the digital systems that would have provided the data necessary for AI, have struggled with legal, regulatory and ethical challenges. Health data especially is subject to specific protections and ethical considerations (for good reason), and navigating data access and responsible use across jurisdictions can be an opaque process, especially for smaller groups. In the UK, a centralised approach to digital contact tracing, which would have been a rich resource for ML analysis, was initially favoured by government; however well-justified concerns about privacy and data governance from civil society groups contributed to the adoption of a decentralised approach.

The COVID-19 crisis still has years to run, during which AI can be far more useful. There are pressing steps in AI governance that can be taken to support this. I highlight in particular the recommendations on global research- and data-sharing, and data governance described in the GPAI report [1]. There is also a role for privacy-preserving ML and other approaches to secure data use. To avoid mistakes and ensure public trust, we

must be able to address ethics and governance with urgency, through better use of foresight and ethics by design; having a strong focus on and support for robustness and reliability; and through fast-moving oversight bodies and cross-society consultative methods [2, 3]. Above all, it is essential that we build on this crisis to be better-prepared for future pandemics. In this crucial challenge, the potential for AI to aid us remains immense.

References

[1]http://thefuturesociety.org/wp-content/uploads/2020/12/Responsible-AI-in-Pandemic-Response.pdf

[2]Tzachor, A., Whittlestone, J., Sundaram, L., & hÉigeartaigh, S. Ó. (2020). Artificial intelligence in a crisis needs ethics with urgency. Nature Machine Intelligence, 2(7), 365–366. https://-doi.org/10.1038/s42256-020-0195-0

[3]Cave, S., Whittlestone, J., Nyrup, R., O hEigeartaigh, S., & Calvo, R. A. (2021). Using AI ethically to tackle covid-19. BMJ, n364. https://-doi.org/10.1136/bmj.n364

ABOUT THE AUTHOR

Seán Ó hÉigeartaigh

Seán Ó hÉigeartaigh is the Director of the AI: Futures and Responsibility programme (AI: FAR) at the Leverhulme Centre for the Future of Intelligence (CFI), an interdisciplinary centre that explores the opportunities and challenges of artificial intelligence. The AI:FAR programme focuses on foresight, ethics and governance related to artificial intelligence.

He is also the Co-Director of Cambridge's Centre for the Study of Existential Risk (CSER), a research centre focused on emerging global risks and long-term challenges.

Seán's research spans the impacts of artificial intelligence and other emerging technologies, horizon-scanning and foresight, and global risk. He led research programmes on these topics at the Future of Humanity Institute (Oxford) from 2011 to 2015, was founding Executive Director of the Centre for the Study of Existential Risk from 2014 to 2019, and co-developed both the Strategic AI Research Centre, and the Leverhulme Centre for the Future of Intelligence. His paper *An AI Race: Rhetoric and Risks* (with Stephen Cave) won joint best paper at the inaugural AI Ethics and Society Conference. He has a PhD in genome evolution from Trinity College Dublin.

From Principles to Actions: Governing and Using AI for Humanity

By Cyrus Hodes

In 2020 the Global Governance of AI, or at least discussions in various fora on the ethics of AI, have gathered momentum and everyone now is (or should be) focusing on translating these principles into action.

As usual, one of the most active and globally influential platforms is the OECD, since the adoption of their principles by the G20 last year in Japan, reaffirmed this year at the Saudi Arabia S20. The launch of the AI observatory (OECD.AI) was successful into bringing together truly cross-disciplinary views, conducting deep dives into AI policy, exploring countries' AI initiatives, and very importantly, better assessing the impact of AI systems through a framework that understands Contexts, Data & Input, AI Models, and Tasks & Output.

In Europe, the EU Parliament's Panel for the Future of Science and Technology (STOA) has launched a partnership with the OECD Global Parliamentary Network, with a focus on the promotion of trustworthy and human-centered AI, as well as shared reflections on the future development of AI. This initiative brought together Members of Parliament from 42 countries and represents an important step toward multilateral coordination on the governance of AI.

An important milestone in AI Governance in 2020 was the launch of the UN Secretary General's Roadmap for Digital Collaboration in May, based on the recommendations of the High-Level Panel on Digital Cooperation. Besides Recommendation 3C on Artificial Intelligence, Recommendation 1B on Digital Public Goods has validated and highlighted the work of the Global Data Access Framework (GDAF). In particular, it calls for the utilization of big data and artificial intelligence to create "digital public goods in the form of actionable real-time and predictive insights, critical for all stakeholders, including the United Nations, as they can serve to identify new disease outbreaks, counter xenophobia and disinformation and measure impacts on vulnerable populations, among other relevant challenges". Whereas the Secretary General goes on to point out "efforts such as the Global Data Access Framework, which is aimed at developing technical infrastructure to enable and scale up the sharing of data in all modalities to speed up the processes for creating quality digital public goods".

The GDAF exercise is co-led by the UN Global Pulse initiative, the AI Initiative of The Future Society (TFS) and the Noble Intelligence initiative of McKinsey, and has over 120 stakeholders, including major technology firms, academic institutions, non-governmental organizations and UN agencies. We are aiming at publishing a blueprint in the first quarter of 2021, followed by a Minimum Viable Product (MVP) of how various types of data could be shared or given access to, in order to run AI systems to help us get to the Sustainable Development Goals.

In 2020, the World Bank has played an important role to help emerging countries adopt and implement national AI strategies (and the AI Initiative was honored to take part in this exercise, the same way we are working on the Rwanda National AI Strategy, with GIZ and the World Economic Forum).

Another relevant platform advancing the international governance of AI is the Global Partnership on AI (GPAI), launched by France and Canada. 2020 saw this partnership consolidate its global role to discuss Responsible AI, Data Governance, Innovation and Commercialization as well as the Future of Work. A very timely subgroup has been created under Responsible AI for focus on AI-based solutions for the Pandemic Response. GPAI's 15 founding members are Australia, Canada, France, 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. There is a strong drive to open GPAI to the Global South and finding a mechanism to have China join as one of the world leading AI powers only makes sense to me.

The Future Society has worked closely with GPAI and their members and published two seminal reports on both the Responsible AI grou,p as well as on the pandem-

ic response, both can be accessed here: https://g-pai.ai/projects/responsible-ai/ and here: https://g-pai.ai/projects/ai-and-pandemic-response/.

Part of the work on the pandemic, nicely springing the OECD AI Principles into action, is the CAIAC (pronounced "kayak") project that I have the pleasure to co-lead, together with Stanford Human-Centered AI (HAI), Stability.ai, various UN partners, starting with UNESCO, and supported by the Patrick J. McGovern Foundation. CAIAC is a platform designed to increase our common knowledge about the COVID-19 virus and to give unique decision-making support. It dynamically maps knowledge about the virus and its impact by connecting localized initiatives and interventions. Such a platform is important in policy making, giving our leaders direct access to relevant knowledge, based on human intelligence and augmented by AI, and available data sets to tackle health aspects of the crisis, as well as the social, economic and financial responses needed to get out of this crisis in a globally coordinated fashion. We know other pandemics will emerge, and we are also facing dormant crisis highlighted by the UN Sustainable Development Goals, starting with Climate Change as a strong case in point calling for globally coordinated actions, augmented by AI systems following up the CAIAC knowledge gathering and sharing model.

If anything, 2020 has highlighted the dire need for global coordination in times of crisis. One can only regret that partisan politics, instead of common sense, has limited global coordination through essential multilateral actors (the WHO being a case in point), but we remain hopeful that in this decade of action for the UN SDGs, more AI for SDG initiatives will blossom, such as the AI4SDGs think tank from the Beijing Academy of Artificial Intelligence (BAAI), with the support of Baidu, Megvi, Yitu and Didi, or the Oxford Initiative on AI×SDGs led by Saïd Business School with the support of Facebook, Google and Amazon. This also clearly points out for an increased role of the UN in bringing together not only governments, but also leading AI platforms in the East and the West, to foster AI based projects to tackle the complex systems that are the SDGs. Hopefully 2021 will allow us to bring these projects together.

Finally, we have been working on a couple of impactful projects with world governments and UN partners (e.g.: UNICRI) to deploy AI systems against human trafficking and to educate law enforcement agencies on existing AI tools to identify and get rid of Child Sexual Abuse Material (CSAM) online. These are very concrete, practical cases where we can leverage the power of AI while helping government and policy makers tackle humanity's challenges. We hope that, as world leaders better embrace the potential of AI systems and implement various AI Principles, large scale, ambitious AI for good programs will be deployed in 2021 with a strong global coordination, emulating collaborations of the CERN or ISSS programs.

ABOUT THE AUTHOR — Cyrus Hodes

Cyrus Hodes recently served as the Advisor to the Minister of Artificial Intelligence at the UAE Prime Minister's Office and has been leading for the past 3 years the Global Governance of AI Roundtable (GGAR) at the World Government Summit in Dubai. He is a Partner at FoundersX Ventures, a cross-stage venture capital firm based in Silicon Valley. Cyrus is the co-founder and Chair of the AI Initiative at The Future Society - a 501(c)3 incubated at Harvard Kennedy School - where he engages a wide range of global stakeholders to study, discuss and shape the governance of AI. At the AI Initiative, Cyrus co-leads the Collective and the Augmented Intelligence Against COVID-19 (CAIAC) project, together with Stanford HAI, advised by leading UN agencies, and co-leads the Global Data Access Framework (GDAF) project with the Executive Office of the UN SG (Global Pulse) and McKinsey (Noble Purpose AI). He is part of the Global Partnership on AI (GPAI), a member of the OECD Expert Group on AI (ONE AI), member of the EU Parliament Panel for the Future of Science and Technology (STOA), a Constituent of Recommendations 1B (Digital Public Goods) and 3C (AI) of the UN Secretary-General's High-level Panel on Digital Cooperation, a member of the Council on Extended Intelligence (MIT-IEEE), and is a co-author of IEEE Ethically Aligned Design. Cyrus is part of the steering committee of AI Commons and is its Data Initiative lead, is a member of the AI Ethics Board of Smart Dubai, and is an AI Governance Advisor at the Shanghai Institute for Science of Science (SISS). Cyrus was educated at Sciences Po Paris, where he later was a Lecturer in International Security, holds a M.A. (Hons) from Paris II University in Defense, Geostrategy and Industrial Dynamics and a M.P.A. from Harvard Kennedy School of Government.

Part V Regional Developments from Policy Practitioners

AI Is Too Important to Be Left to Technologists Alone

By Eugenio Vargas Garcia

In a year tragically disturbed by COVID-19, we have learned a key lesson. Much like a virus, the overarching impact of AI will not be confined to national borders. The pandemic demonstrated that we need international cooperation to successfully tackle cross-border issues in everyone's interest.

As a general-purpose technology with multiple capabilities and long-term implications, AI can pose challenges that must be prevented or mitigated, by pooling resources and expertise in defining agreed parameters to safely develop its full potential.

Despite considerable progress on many fronts, the international governance of AI still lacks satisfactory norms, policies, safety measures, and technical standards. There are no regimes or normative instruments at the global level that go beyond high-level principles.

Political tensions, growing competition, polarization, and tribalism do not seem conducive to major agreements in a very short time. As a result, the absence of collaborative forms of governance, combined with mistrust towards the multilateral system, can complicate efforts to cope with AI risks in the long run.

In a fragmented landscape, if states fail to coordinate properly, future regulation of AI can become Balkanized. A "splinternet" scenario, with opposing blocs holding mutually incompatible rules, should be avoided.

A do-nothing approach is hardly an option. In a normative vacuum, driven by a logic of race to the bottom, governments and private companies may push even harder for rapid AI development, regardless of considerations based upon law, ethics, safety, or security.

The need for AI governance is clear. As the technology becomes more ubiquitous, demands will grow stronger to set in motion cooperation to prevent harm. Predictability by means of norm-setting can pave the way to responsible strategies to minimize disturbing scenarios.

Realistically, a regulatory AI agency now seems too far away. At this juncture, multi-stakeholder forums, on a voluntary basis, could fill the gap if focused upon recommendations geared at prevention rather than regulation per se.

Multilateralism can play a role moving forward. The United Nations, for instance, with its unmatched range and universality, could offer a neutral, nonpartisan, legitimate platform for facilitating negotiations.

UNESCO has made great strides in advancing a much-needed discussion on a draft for the first global standard-setting international instrument on the ethics of AI.

The UN Secretary-General, in his Roadmap for Digital Cooperation, identified three crucial tasks: increase representation from the Global South in AI deliberations; improve overall coordination of existing initiatives; and capacity-building, particularly in the public sector.

UN-sponsored consultations were held in 2020 on creating an AI Advisory Body to provide expert advice and set the stage to effective consensus-building. The ultimate goal should be reaching solutions that can accommodate all concerns as much as possible.

“War is too serious a matter to leave to soldiers”, Clemenceau once said. Similarly, AI is too important to be left to technologists alone.

AI policymaking will require bridging the gap between the technical community and political leaders, governmental officials, diplomats, and parliamentarians. Interconnecting these two worlds is not only advisable, it is critical for success.

ABOUT THE AUTHOR

Eugenio Vargas Garcia

Eugenio V. Garcia is diplomat, PhD in International Relations from the University of Brasilia, and researcher on artificial intelligence and global governance. Currently Minister-Counsellor and Chargé d’affaires ad interim in the Brazilian Embassy in Conakry, Republic of Guinea. Former senior adviser on peace and security, Office of the President of the United Nations General Assembly, New York (2018 - 2020). Worked previously in the Brazilian Embassies in London, Mexico City, Asuncion, and in the Permanent Mission of Brazil to the United Nations in New York. Held different positions in the Ministry of Foreign Affairs, including on Asia-Pacific affairs, diplomatic planning, and as adviser to the Foreign Minister (2005 - 2009) and the Deputy Foreign Minister (2014 - 2015), as well as Head of the United Nations Division (2015 - 2018). Visiting research associate at the University of Oxford (1999 - 2000) and professor at the College of Latin American Studies, National Autonomous University of Mexico (2004 - 2005). Published seven books on foreign policy and international affairs. Former Brazilian junior chess champion (1985). His main areas of academic research include AI, the impact of new technologies in peace and security, and the role of multilateral organizations.

The Governance Approach of Artificial Intelligence in the European Union

By Eva Kaili

Artificial Intelligence grows to be a critical technology which is expected not only to change business models and our life as consumers, it most importantly challenges traditional models and notions of citizenship. European Union is in the core of this transformation, aspiring to introduce to the World a third path; neither the model of surveillance capitalism advanced by the United States nor the model of digital imperialism advanced by China. Europe brings to the front a third approach that accelerates innovation and the uses of Artificial Intelligence without compromising the privacy rights of the citizens, without discounting the ownership and value of the data that the people produce, and without violating the safety and quality of life of the people in the labyrinth of applications and means of data collection.

Europe introduces a model of governance based on the fundamental notion of technological neutrality, where the regulator sets the principles and the market comes and applies the principle by defining the standards of the product or service. The game changer in this approach is that the standards comply to strict and fundamental principles for customer protection, social inclusion, human rights, privacy and non-discrimination, that EU is a global model of safeguarding.

The European Institutions worked systematically in the last year to establish this competitive framework that we aspire to become the global standard in AI. This framework aims to elevate people's trust in AI and ensure that in the digital age, people co-exist with intelligent systems without fearing exclusion, manipulation, oppression or discrimination. Retaining freedom of choice in a human-centric AI that would prevent brain computer interfaces challenging the nature and future of humanity. In contrast to the trends of the Fourth Industrial Revolution towards inequalities and dehumanization, technology and innovation best practices need now to be bent back towards the service of humanity, and Europe could lead as a global rules and standards setter for the Fifth Industrial Revolution. The European principle-based framework for AI systems must translate and establish by law, with respect to our rights in the digital age. At the foundational level, this framework must guarantee higher transparency and accountability, define the liability of AI systems, establishing standards to trace, audit, explain, appeal and reverse decisions made by AI during its entire lifecycle.

The rapid development of automation in Europe must not reflect mistakes of the past; AI algorithms and systems must be trained on diversified sets, and their objectives have to be clearly defined and controlled to avoid risks of bias and discrimination or data poisoning. Decisions made by AI must be aligned with the collective ethical fabric that defines Europe throughout the lifecycle of intelligent systems and follow clear red lines, such as the risk-based approach that would, for example, completely ban research in autonomous lethal weapons or conscious AI. The high-risk applications should alert us of the fragility of our common values, as well as our rights as citizens, our responsibilities as policy-makers and our obligations to future generations.

ABOUT THE AUTHOR

Eva Kaili

Eva Kaili is a member of the European Parliament, part of the Hellenic S&D Delegation since 2014. She is the Chair of the Future of Science and Technology Panel in the European Parliament (STOA) and the Centre for Artificial Intelligence (C4AI), Member of the Committees on Industry, Research and Energy (ITRE), Economic and Monetary Affairs (ECON), Budgets (BUDG), and the Special Committee on Artificial Intelligence in a Digital Age (AIDA).

Eva is a member of the delegation to the ACP-EU Joint Parliamentary Assembly (DACP), the Delegation for relations with the Arab Peninsula (DARP), and the Delegation for relations with the NATO Parliamentary Assembly (DNAT).

In her capacity, she has been working intensively on promoting innovation as a driving force of the establishment of the European Digital Single Market. She has been the draftsperson of multiple pieces of legislation in the fields of blockchain technology, online platforms, big data, fintech, AI and cybersecurity, as well as the ITRE draftsperson on Juncker plan EFSI2 and more recently the InvestEU program.

She has also been the Chair of the Delegation to the NATO PA in the European Parliament, focusing on Defence and Security of Europe.

Prior to that, she has been elected as a Member of the Hellenic Parliament 2007—2012, with the PanHellenic Socialist Movement (PASOK).

She also worked as a journalist and newscaster prior to her political career.

She holds a Bachelor degree in Architecture and Civil Engineering, and Postgraduate degree in European Politics.

The Third Way: the EU's Approach to AI Governance

By Charlotte Stix

Good intentions should be followed by concrete actions. It is clear that this principle also holds for the development and deployment of AI that performs in a reliable and secure manner. Recent years were filled with scoping the ethical and policy landscape with regards to AI, spearheaded in the European Union (EU) by the High Level Expert Group on AI, an independent advisory group to the European Commission and fortified through the Coordinated Plan on AI, where European Union Member States agree to cooperate and coordinate their actions on a number of related policy areas. We can now see this work growing roots. Most notably, the EU has taken concrete steps towards regulating AI. While the process will be long, the first outlines are already evident and have been presented in the White Paper on AI: A European Approach to Excellence and Trust, where proposals for ethical considerations, legal obligations and technical infrastructure intertwine to create what the European Commission sees as two sides of a coin: an "ecosystem of trust" and an "ecosystem of excellence". This brief piece will elaborate further on the ecosystem of trust.

The framework for the "ecosystem of trust" aims to outline the regulatory framework that the European Commission will solidify in early 2021, with the goal of regulating all high risk AI systems within the EU. As a global first, the risk based approach to regulation is ambitious and advocates an approach that adequately addresses risks of AI systems whilst promoting their development and uptake.

Trust and trustworthiness continue to form a red thread for the EU's activities with regards to AI. Indeed, the ability to encourage a flourishing ecosystem with trustworthy AI for users, citizens and the environment is seen as a major advantage for the EU to harness and a risk-based approach to regulation as one that may help ensure proportionality.

Building on the Ethics Guidelines on Trustworthy AI's seven key requirements for trustworthy AI, the regulatory proposal uses this non-binding framework to develop requirements for high-risk AI systems, that is, those that fall under the regulatory scope. High-risk AI systems, or cases, are defined by two criteria: if the sector itself is high risk (e.g. healthcare, transport etc.) and if the intended use involves high risk (e.g. injury, death, significant material/immaterial damage). Mandatory requirements for AI systems in the proposed legal framework are applicable if both of the aforementioned criteria hold (i.e. cumulative criteria). The White Paper also suggests that cases such as biometric identification and the use of AI for recruitment processes should be seen as high risk. Moreover, non-high risk applications may have the option to partake in a voluntary labelling scheme also, drawing on the seven key requirements for trustworthy AI.

The road towards a proportionate regulation for AI and implementing it within the ecosystem is still long and rocky, but the EU is taking steps towards achieving this goal.

ABOUT THE AUTHOR

Charlotte Stix

Charlotte Stix is an experienced technology policy expert with a specialization in AI governance. Her PhD research at the Eindhoven University of Technology critically examines ethical, governance and regulatory considerations around artificial intelligence. In that context, she serves as Fellow to the Leverhulme Centre for the Future of Intelligence, University of Cambridge and as Expert to the World Economic Forum's Global Future Council on Neurotechnologies.

Most recently, Charlotte was the Coordinator of the European Commission's High-Level Expert Group on Artificial Intelligence. Formerly, she was a Researcher at the Leverhulme Centre for the Future of Intelligence, University of Cambridge, a Fellow to the World Economic Forum's AI Council, consulted Element AI as an Advisor on their European AI engagement and served as a Policy Officer at the European Commission's Robotics and AI Unit where she managed over 18m in projects.

Charlotte is awarded as a 2020 Forbes' 30 under 30 (Europe) and was named a Young Global Shaper by the World Economic Forum.

A Year of Policy Progress to Enable Public Trust

By Caroline Jeanmaire

In the last year, AI regulatory frameworks have finally begun to emerge, advancing beyond mere "AI Principles". One cornerstone of this change is the realization that, rather than impeding innovation, regulation can fuel development by inspiring public trust. To hold the public trust and avoid another AI Winter, AI governance must factor in a "blind-spots radar" to reduce the risk of unforeseen calamities.

Regulators in the United States have maintained a light touch while prioritizing public trust. Building on AI principles developed by the Defense Department and the intelligence community, in December 2020, President Trump issued an Executive Order outlining nine principles agencies must meet when designing, developing, or acquiring AI applications. It also calls for comprehensive inventories of AI deployments within agencies and a roadmap for improved policy guidance on AI use. These regulatory principles encourage innovation by reducing uncertainty. "The ongoing adoption and acceptance of AI will depend significantly on public trust", the executive order states. "Agencies must therefore design, develop, acquire, and use AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values". To oversee and implement the U.S. national AI strategy, the White House established the National Artificial Intelligence Initiative Office in January 2021. The next few months under a Biden administration will provide more clarity around the duties of care that developers and manufacturers must have towards the public.

In parallel, the European Union followed principles developed by the High-Level Expert Group on AI in proposing new regulatory frameworks. In 2020, the European Commission published a White Paper that identified seven key requirements AI technologies should respect as well as a regulatory framework adopting a risk management approach. In addition, the Commission recommended including stand-alone software within the scope of product regulation and ensuring the safety of AI systems across economic actors in supply chains.

Regulators in the US and the EU understand the risks of technology backlash, knowing that accidents or failures could provoke public fear, as has happened before. The Three Mile Island Nuclear Accident in 1979 inhibited the development of the US nuclear power industry for 30 years, while in Europe the public has shied away from genetically modified foods due to various food safety issues including the Mad Cow disease. Conversely, the aviation industry enjoys public confidence, thanks to robust safety certification processes and visible processes for investigating accidents.

However, regulators on both sides of the Atlantic have yet to verify risk assessment criteria and risk management strategies. Processes to monitor and reassess technology for the appearance of "unknown unknowns" still need to be developed. Over 150 experts have signed the "Foresight in AI Regulation Open Letter" to the European Commission urging them to keep and strengthen the proposed regulation, even though some groups may downplay potential risks related to AI. AI ethics and safety enable innovation, and regulation cannot be simply a façade if systems are to inspire public faith and confidence in the long-haul.

ABOUT THE AUTHOR

Caroline Jeanmaire

Caroline Jeanmaire is the Director of Strategic Research and Partnerships at the Center for Human-Compatible AI at UC Berkeley. She focuses on building a research community around AI safety and relationships with key stakeholders. Her research focuses on policies that ensure the safety and reliability of AI systems as well as models for international cooperation around AI. Before working at CHAI, she was an AI Policy Researcher and Project Manager at The Future Society. She notably supported the organization of the first and second Global Governance of AI Forums at the World Government Summit in Dubai, with over 200 attendees. Caroline was a Youth Delegate to the United Nations for two years with the French delegation. She has a dual master's degree in International Relations from Peking University and Sciences Po Paris and a bachelor's degree in political sciences from Sciences Po Paris. Caroline was named one of the 100 Brilliant Women in AI Ethics in 2021.

From Human-Centric to Planetary-Scale Problem Solving: Challenges and Prospects for AI Utilization in Japan

By Arisa Ema

The Japanese government has been promoting its information policies under the concept of Society 5.0, a human-centric society that simultaneously achieves economic development and resolution of social issues through a system that integrates cyberspace and physical space. The Cabinet Office released the "Social Principles of Human-Centric AI" in March 2019. Since then, relevant ministries and agencies have formulated AI-related guidelines, such as guidelines for the development of medical diagnostic imaging support systems (Ministry of Health, Labor and Welfare), guidelines for contracts in the agricultural sector (Ministry of Agriculture, Forestry and Fisheries), a handbook for AI utilization (Consumer Affairs Agency), and certification systems for AI education (Ministry of Education, Culture, Sports, Science and Technology and Ministry of Economy, Trade and Industry and CSTI).

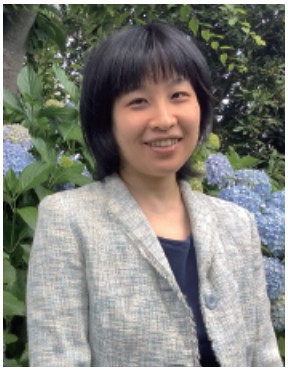
On the other hand, the response to COVID-19 exposed the fact that we were completely unprepared for the transition to a Society 5.0 society. The fact that government and private companies signed an agreement on COVID-19 to promote the use of data can be regarded as a step forward. However, data sharing between the central and local governments and the medical institutions was found to be ineffective. In addition, we found out that the Tokyo Metropolitan Government was using fax machines to report the number of infected people.

The problem of data sharing is not limited to COVID-19, in fact, has been pointed out previously. Particularly, in Japan, there are more business-to-business companies than business-to-consumer companies. In other words, there are many cases where data acquisition, AI model development, and service providers are different. Therefore, issues such as AI safety and fairness must be addressed not only by one company, but by all parties involved in a multi-layered dialog. Moreover, start-up companies lack resources, making it difficult for them to deal with risks. From this perspective, the Japan Deep Learning Association, whose members are mainly start-ups, is considering how to assess and respond to AI systems in the external environment, including insurance, auditing, accident investigation, consumer protection, whistle-blowing systems, and standardization, in addition to AI governance within companies since the summer of 2020.

Last but not the least, AI governance must be discussed internationally. The 2nd French-German-Japanese AI Symposium was organized in November 2020. The joint statement of the first conference held two years ago emphasized a human-centric approach. Therefore, in line with this direction, the theme of the second conference was human-centric AI. However, we are now facing planetary-scale challenges, such as the COVID-19 pandemic, climate change, and community fragmentation. Therefore, the joint statement in 2020 proposed that planetary-scale problems, from anthropocentric to environmental, should be addressed. It is important to work on AI governance for planetary-scale issues in collaboration with various stakeholders and organizations.

ABOUT THE AUTHOR

Arisa Ema



Arisa Ema is Assistant Professor at the University of Tokyo and Visiting Researcher at RIKEN Center for Advanced Intelligence Project in Japan. 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) (http://sig-air.org/) 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) and chairing AI governance study group. She was also a member of Council for Social Principles of Human-centric AI, The Cabinet Office, which released “Social Principles of Human-Centric AI” in 2019. She obtained Ph.D. from the University of Tokyo and previously held a position as Assistant Professor at the Hakubi Center for Advanced Research, Kyoto University.

India's Strategies to Put Its AI Economy on the Fast-Track

By Raj Shekhar

2020 presented unprecedented challenges to governments globally, but especially in developing countries, where nationwide measures to contain the spread of the COVID-19 risked putting other public policy priorities in the pipeline on indefinite hold. Yet, in India, the Narendra Modi government commendably managed to drive critical consultation and dialogue on several cardinal policies, targeting the improvement of India's AI readiness to support the domestic growth of both public and private AI enterprises, in line with the Prime Minister's vision for AtmaNirbharBharat (or self-reliant India).

Considerable progress was made by the Indian government in devising frameworks for data governance in India. The Joint Select Committee of the Indian Parliament met several times during the year for a clause-by-clause examination of The Personal Data Protection (PDP) Bill, 2019, given its undisputed significance in deciding the fate of the Indian digital citizenry and that of the Indian digital economy. The Ministry of Electronics and Information Technology (MeitY) proposed (a) the Data Centre Policy, detailing inter alia a set of enabling regulations and programs for building domestic workforce and infrastructural capabilities to realize the data localization mandate enshrined in the PDP Bill, 2019; and (b) the Non-Personal Data (NPD) Governance Framework, recommending inter alia the enactment of a central NPD statute and the establishment of a central NPD authority under the statute to unlock the economic, social, and public value of data, whilst securing an equitable, innovation-friendly, and privacy-respecting regime for non-personal data sharing in the country. The National Institution for Transforming India (NITI) Aayog, the Indian government's premier policy think tank, proposed (a) setting up a National Data and Analytics Platform (expected to launch this year), to publish government data in an open, user-friendly format in order to enable research and innovation, evidence-based public policy-making, and participatory governance in the country; and (b) the adoption of Data Empowerment and Protection Architecture, enabling citizens to securely and seamlessly share their personal data with third party institutions to benefit from easy, hassle-free access to a host of banking and insurance products online.

Additionally, recognizing the whole gamut of ethical concerns emerging around the development and deployment of various AI use-cases globally, NITI Aayog released working documents on Responsible #AIforAll and Enforcement Mechanisms for Responsible #AIforAll. The documents emphasized inter alia the need to uphold constitutional fundamental rights in defining principles for responsible AI governance in India, and recommended a national-level, multidisciplinary Council for Ethics and Technology, tasked with helping sectoral regulators to develop proportionate, risk-based AI regulations and other measures to foster the creation of a responsible AI ecosystem in India.

To equip India's next generation of professionals with AI-ready skills, the Ministry of Human Resource Development (MHRD) in the third National Education Policy proposed to bring about a pedagogical transformation in the Indian educational system, by means of introducing inter alia courses on AI in the school and university curricula.

2020 concluded with the Ministry of Science and Technology (MST) releasing the 5th National Science, Technology, and Innovation (STI) Policy with an overarching bearing on the improvement of India's AI readiness. The Policy detailed the Indian government's vision for achieving technological self-reliance, by means of well-grounded, dynamic, and inclusive institutional governance mechanisms, aimed at enabling open and wide access to all publicly-funded STI research, increasing public and private funding of STI research in priority sectors, and developing essential human capital for the STI ecosystem.

Even as these proposals from MeitY, NITI Aayog, MHRD, and MST for national-level policy shifts may have had their weaknesses, they have appreciably paced up stakeholder engagement on a range of pressing AI governance issues in India. This demonstrates a marked improvement from 2019 in the government of India's positioning, vis-à-vis AI governance that promises to lend the Indian AI economy the critical foundations it needs to pick up steam and compete in the global AI marketplace with resilience.

ABOUT THE AUTHOR

Raj Shekhar

Raj Shekhar is the Founder and Executive Director at AI Policy Exchange, an international cooperative association of individuals and institutions working at the intersection of AI and public policy, with the mission to produce deliverables that can create an AI-literate society and inform better AI policies. As an Affiliated Scholar at CITRIS Policy Lab, University of California, Berkeley, Raj is collaborating with Dr. Brandie Nonnecke (Founding Director) on research related to the governance on AI. As Consultant (Data, AI) at International Innovation Corps (IIC) of The University of Chicago, Raj is supporting (a) operations of the Open Data Working Group, an initiative by IDFC Institute and IIC to advance India's open data aspirations, and (b) IIC's engagement with the Ministry of Electronics and Information Technology, Government of India aimed at building capacity for data and AI innovation through policy and program implementation. Raj also is an Affiliate at The Future Society and sits on the Founding Editorial Board of Springer Nature's AI and Ethics Journal.

Earlier last year, Raj was featured as a leader in the Responsible Tech ecosystem in the Responsible Tech Guide brought out by All Tech Is Human in partnership with NYU Alliance for Public Interest Technology. He was recently declared a DIET Champion for 2021 by DataEthics4All for his demonstrated commitment to Data and Diversity, Inclusion and Impact, Ethics and Equity in Teams and Technology (DIET).

Raj holds an integrated Bachelor's in Law and Humanities from Dr. Ram Manohar Lohiya National Law University, Lucknow, and a Master's in Public Policy from the Institute of Public Policy, National Law School of India University, Bengaluru.

"Cross-Sector GPS": Building an Industry-Agnostic and Human-Centered Future of Work

By Poon King Wang

In 2020, under the guidance of Singapore's Advisory Council of the Ethical Use of AI and Data, the Infocomm Media Development Authority and the Personal Data Protection Commission collaborated with the Lee Kuan Yew Centre for Innovative Cities (LKYCIC), to launch A Guide to Job Redesign in the Age of AI.

It is Singapore's first industry-agnostic guide to help companies see how they can take full advantage of AI, and at the same time manage its impact on employees with practical human-centered strategies.

The need for human-centricity is well understood. The value of an industry-agnostic approach to achieve human-centricity was made clear by the pandemic in 2020.

COVID-19 showed us what upskilling initiatives lack when entire sectors are disrupted. The workers in those sectors suffer because many current worker upskilling efforts are sector-specific. Such efforts are, however, inadequate when there are hardly any options left for workers within their professions and sectors.

This predicament will become more prevalent in the future because the pandemic has also accelerated digitalization. The acceleration will cause more sectors to see widespread disruption from automation and remote work (that could lead to outsourcing). Sector-specific efforts will hence be similarly inadequate to help workers in their disrupted professions and sectors.

This is where an industry-agnostic strategy is valuable. An industry-agnostic strategy means we will be able to help workers find opportunities outside of their professions and sectors.

The LKYCIC's future of Work research has developed such an industry-agnostic approach that is effectively a "cross-sector GPS". By using AI and a tasks-skills stack we have built, we can chart clear, concrete, step-by-step pathways from disrupted professions/sectors to growing professions/sectors.

Our industry-agnostic strategy is grounded in the converging consensus across research and industry, that tasks are the right resolution to study the economic and technological impacts on jobs. By identifying the tasks shared by different jobs in different sectors, we can chart cross-sector pathways that become the basis for designing initiatives to help affected workers. Our "cross-sector GPS" thus expands options across sectors for workers, helping them navigate the current and future crises better.

Our "cross-sector GPS" combines the power of AI algorithms with our tasks-skills stack. The tasks-skills stack links national and international databases in industry and occupational data. The research underpinning it all draws on and integrates insights across the disciplines of data science, engineering, AI, labor economics, occupational psychology, and organizational studies.

By contributing our research to create Singapore's first industry-agnostic A Guide to Job Redesign in the Age of AI, we will strengthen the broader efforts to build a trusted and progressive AI environment that benefits citizens, companies, and governments.

ABOUT THE AUTHOR

Poon King Wang

Poon King Wang is the Director of the Lee Kuan Yew Centre for Innovative Cities at the Singapore University of Technology and Design (SUTD), where he also heads the Smart Cities Lab and the Future Digital Economies and Digital Societies initiative. He is concurrently Senior Director of Strategic Planning at SUTD.

King Wang is on the World Economic Forum's Expert Network on Cities and Urbanisation, Konrad Adenauer Stiftung's Strong Cities 2030 Network, and the Future of Work working group of the Global Partnership on Artificial Intelligence. He was also on two national taskforces on the future of adult learning, and on future services and digital economy. He has co-authored and published *Living Digital 2040: Future of Work, Education, and Healthcare*, which was recently translated into Korean.

His teams' research is recognised in Singapore's National AI Strategy for helping to build a trusted and progressive environment for AI. His team also collaborated with the Infocomm Media Development Authority's Personal Data Protection Commission (under the guidance of the Advisory Council of the Ethical Use of AI and Data) to launch A Guide to Job Redesign in the Age of AI, Singapore's first industry-agnostic guide to help companies manage AI's impact on employees.

AI Governance Readiness: Rethinking Public Sector Innovation

By Victor Famubode

2020 has been an awkward year for human civilization. This is evident in the way humans have responded to the pandemic. On one hand, the pandemic exposed vulnerabilities in our global innovation system and conversely accelerating new opportunities through broader digital adoption. In various cases, at the heart of this digital adoption is the use of artificial intelligence across diverse sectors and domains.

For government, a significant milestone in AI's application during this pandemic has been the collaborative use of such general-purpose technology with other technologies and data to combat the spread of COVID-19. For instance, Egypt's government through the "United Nations Development Programme (UNDP) launched an automated testing service of COVID-19 symptoms with sign language chatbot". Notably, the application of AI use cases in high-stake domains, such as public healthcare, comes with its own risk and it ultimately needs to be acknowledged by governments across African countries.

The rapid growth of AI's application across the continent shows the need to rethink public sector innovation. This would entail moving away from tokenization of democratizing responsible ways of deploying and implementing AI systems to actual operationalization of these systems. While there has been slow progress in adopting guardrails for data protection as "24 African countries out of 53 countries have adopted data protection regulation and laws" (Privacy International 2020), implementation and compliance of these regulations would be a key determinant of addressing AI and data governance. However, governments in many parts of the continent are still confronted with weak institutions. This ultimately means adopting and scaling AI Governance would require addressing the challenges confronting these institutions.

Importantly, 2020 showed progress in awakening African governments to the ethical implications of AI. A major step has been to partner with institutions to assist with designing national strategies that incorporates ethical guidelines. A key example is "Rwanda's Ministry of ICT and Innovation (MINICT) and Rwanda Utilities Regulatory Authority (RURA) with implementation by GIZ FAIR Forward engaging Future Society to support the development of Rwanda's national artificial intelligence strategy".

Going into 2021, major ethical issues such as algorithms bias, surveillance, digital divide and privacy need to be thoroughly addressed. In addressing them, key considerations for African governments towards readiness in AI Governance include:

1. Leverage a community pipeline for designing, deploying and implementing AI systems. This ultimately helps to incorporate diversity and variations into data and models being built for public consumption.
2. Re-evaluate public procurement processes to significantly capture wider consultation and transparency in procuring AI systems from private sector technology companies. It assists with spotting key safety and security vulnerabilities early on before deployment.
3. Build capacity in policymakers, developers and

society planners in responsible use of AI systems.
4. Create more public awareness on both the benefits and the limitations of AI systems.
5. Incorporate risk and impact assessments from design to implementation phase across each public sector domain that these AI systems are expected to be used.

ABOUT THE AUTHOR

Victor Famubode

Victor is currently a committee member for IEEE Global Initiative for Autonomous and Intelligent Systems - P7003 Algorithms bias and considerations. He is an AI governance and policy specialist with focus on helping government institutions with designing data governance frameworks. His experience cuts across media, government and consulting with primary focus on ensuring better public policy choices are made within the technology sphere.

AI Governance in Latin America and Its Impact in Development

By Olga Cavalli

Latin America is a region that offers a vast and diverse geography and a fantastic biodiversity. It is the home of the biggest rivers, mountains and many natural resources and beautiful places, enhanced with rich culture and well-trained human resources.

The recent events related with COVID-19 pandemic are having a profound impact in the regional economies. In this complex scenario, the use of artificial intelligence can introduce changes and innovation in the national and regional industries, making them more productive. According to a report by the Inter-American Development Bank (IADB), the use of artificial intelligence can increase the GDP of the biggest economies of the region in the next decades.

Although national priorities have changed in 2020, putting more focus on areas like health and economy, several governments in the Latin American region are working on public policies and national strategies to promote the use and development of AI at the national level.

Mexico was one of the first countries in the world to create a national AI strategy, Brazil has launched the National Internet of Things plan, which includes the creation of AI laboratories with focus on strategic areas like cybersecurity and defense, Chile is working with the community and national experts to develop its own plan for AI, in Argentina the national AI strategy is being developed and Colombia has published its Ethical Framework for AI.

The challenge for the Latin American region is to work towards a multistakeholder plan that must include government, private sector, civil society, technical community, and academy, where efforts and resources must be focused on the use of state-of-the-art technology and the enhancement of education in strategic areas and industries. The region has the highest inequity of the world, and the use of AI should improve this undesired situation.

One of the problems of the region is its limited voice and low participation in those spaces, where the global AI governance and ethics frameworks are defined. Those spaces are usually dominated by developed economies, and if the voice of Latin America is not raised, the results could be unbeneficial for the needs of the region.

There are concrete efforts to overcome this problem and enhance the relevant participation of the regional experts in international negotiations, like the South School on Internet Governance, SSIG, which trains students and young professionals to become the leaders of the region in these negotiations. The training is free for participants. Since its creation in 2009 and along the twelve editions, it has trained thousands of fellows from the region and other countries, giving them tools on how to navigate the international technology ecosystem and creating among them and international experts a very strong and valuable network.

There is an important competence from a world dominated by constant innovation from highly developed countries, but all the countries of the region have innovative companies and human resources that have relevant conditions to position themselves as leaders based on AI applications.

ABOUT THE AUTHOR — Olga Cavalli

Olga Cavalli is Professor of Internet infrastructure at the Economy School of University of Buenos Aires in Argentina.

She is director of the South School on Internet Governance SSIG. She co-edited the book *Internet Governance and Regulations in Latin America*, and is also a frequent contributor writing papers and chapter books in her areas of expertise.

She was chair of the Study Group on Internet of Things and Smart Cities (2015 - 2017). The ITU has recognized her as an innovator in this field.

During 2007 - 2014 she was distinguished by the United Nations Secretary General, being chosen to be a part of the Internet Governance Forum's Multistakeholder Advisory Group, MAG. Olga is very active in ICANN where she has recently been appointed as GNSO Council Member.

Olga participated the National AI Strategy development group. Her current research includes the impact of AI and 5G in the development of countries in Latin America.

Her education includes a PhD in Business Direction, a master's degree in Business Administration, a master's degree in Telecommunications Regulation, and a degree in Electronic and Electric Engineering. She is fluent in Spanish, English, Portuguese and German, and can understand French and Italian.

Olga lives in Buenos Aires, Argentina.

Artificial Intelligence in Latin America

By Edson Prestes

The discussion about the Governance of Artificial Intelligence in the LATAM region still is incipient. Although there are some national plans on AI in the region, debates about Governance per se are far from being broad and inclusive. They are limited to specific segments of the society and do not consider a multi-level and intergovernmental participation in the process. In fact, there is a huge informational gap between the main actors that prevent to understand each other and also the complexity of the domain. There is a clear need of human and institutional capacity building to empower citizens and states in the region to create an effective and useful governance mechanism. No governance mechanism can be developed adequately if the main stakeholders do not understand the limits of the technology; the implication of its wide use; the opportunities that can be created by AI-based technology; the multidisciplinary nature of AI domain; the need for agile regulatory mechanisms; the need of collaboration across nations; the need for accountability and transparency practices; and the importance of public engagement.

The national plans for AI in the region are strongly influenced by international ethical and regulatory principles, like those elaborated by OECD, WEF, etc. However, these biased plans seem to be partially adequate to LATAM reality. We can not simply import or copy mature models developed in other regions without a deep reflection on their local impact. We have particularities. Some are intrinsic to the region, while others were already addressed by High Income Countries. To cite some of them: some communities do not have yet electricity, then they will neither take full advantage of the AI ecosystem nor have their values represented in the AI models. Some countries in the region do not consider education as priority - the most important component for AI, and therefore, we have witnessed year by year a decreasing in the investment on Education. In addition, LATAM region lacks strategies and a rich ecosystem to prevent “brain drain”.

Of course I have a long list of concerns, but all of them touch on some very basic points: education and information. Considering Brazil, it is common seeing people giving their data to receive purchase discounts or being forced to give their data to get access to benefits. Some Brazilians accept the current Brazilian Government’s view that fake news is an example of freedom of expression; and not a human rights violation. Thus, the question that naturally appears is how a governance mechanism for AI can protect people from abuses or human rights violations caused directly or indirectly by AI if the citizens do not know their rights and obligations; or worse, do not have adequate formal education to have this understanding? No trustful or reliable governance mechanism can exist or survive if the people are not placed in the centre of the debate, i.e., when the view of the main stakeholders (i.e.: government or companies) is profit-oriented or distorted.

ABOUT THE AUTHOR

Edson Prestes

Edson Prestes is Professor at Institute of Informatics of the Federal University of Rio Grande do Sul (UFRGS), Brazil. He received his BSc in Computer Science from the Federal University of Pará (1996), Amazon, Brazil, and MSc (1999) and PhD (2003) in Computer Science from UFRGS. Edson is Senior Member of the IEEE Robotics and Automation Society (IEEE RAS) and IEEE Standards Association (IEEE SA). Over the past years, he has been working in different initiatives related to Standardisation, Artificial Intelligence, Robotics and Ethics. For instance, Edson is Member of the United Nations Secretary-General's High-level Panel on Digital Cooperation; Member of the UNESCO Ad Hoc Expert Group for the Recommendation on the Ethics of Artificial Intelligence; South America Ambassador at IEEE TechEthics; Chair of the IEEE RAS/SA 7007 - Ontological Standard for Ethically Driven Robotics and Automation Systems Working Group; Member of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Advisor at The Future Society; Advisor at the Carnegie AI & Equality Initiative at Carnegie Council for Ethics in International Affairs; Member of UNESCO IFAP Information Accessibility Working Group and Past Associate Vice President on the IEEE RAS Industrial Activities Board.

2020: A Key Year for Latin America's Quest for an Ethical Governance of AI

By Constanza Gomez Mont

The year 2020 was a year that deepened the ethical questions on the impact digital technologies have in the lives of people, especially AI technologies. What did giving up certain levels of privacy to respond to a health crisis mean in the short and long-run? What does it mean that only families with access to digital tools could continue with their studies and work? How does the increased use of digital platforms impact the strengthening of already existing data monopolies? What does a lack of digital literacy mean in a context where most basic services migrated online?

For Latin America, a region with one of the greatest social inequalities, these questions are not taken lightly. The quest for answers on how to ensure a more just and rights-based approach to data and AI-driven technologies has taken various institutions in the region to develop in 2020 key initiatives on ethical AI governance.

For example, the Inter-American Development Bank rolled out the initiative fAIr LAC, that helps governments and entrepreneurs adopt responsible AI practices through the development of guides and the creation of hubs in Mexico, Uruguay, Costa Rica and Colombia in collaboration with public and private partners. Moreover, the IEEE Global Partnership co-founded with C Minds the Latam Circle with the vision for prioritizing human well-being with autonomous and intelligent systems in the region and fostering a meaningful participation of Latin American experts in the development of global AI ethical standards; UNESCO started an online community as a follow-up of the regional consultation of the ethical AI global instrument. Other institutions such as UNICRI and the Eon Resilience Lab kickstarted an effort to include the voice of regional experts in the development of a global toolkit of rights-based AI for crime prevention and justice; and networks such as IA2030Mx, in Mexico, and AI Latam, with a regional umbrella, are strengthening the AI ecosystem. This list of actions are some of various other regional efforts to advance ethical AI governance.

Moreover, some governments did not stay behind in 2020. The government of Uruguay published a tool to evaluate the impact of AI systems; the government of Colombia published a draft of an AI ethical framework; the Federal Institute of Access to Public Information in Mexico helped kickstart the first policy prototype on transparency and explicability of AI systems of the region; the government of Chile started the development of the AI Policy that includes a pillar of ethics which will be published in 2021.

These are only a few of existing initiatives that add to past years efforts in countries such as Brazil and Argentina. There is no doubt that Latin America has a very long road ahead to consolidate and scale these initiatives as well as deploying more coordinated actions. And most importantly, making sure that the access and benefits of digital technologies and AI are distributed to all, specially as it still faces its fight against the pandemic and enters a phase of economic recovery. In these times of great need of rethinking ethical questions and accelerating inclusive answers, the design of AI's ethos for the region and how it is governed will have a profound impact on Latin America's prosperity. It will be key that it lives up to expectations, contributes to leading practices of the Global South and have a meaningful participation in the development of global governance AI processes.

ABOUT THE AUTHOR

Constanza Gómez Mont

Constanza is a social impact strategist and practitioner that brings governments, companies, multinational organizations, and local communities together to co-design initiatives that accelerate the positive impact of emerging technologies.

She is the founder and president of C Minds, a women-led action tank that works in the intersection of new technologies, society, and the environment. Through C Minds leadership, she has worked for over 12 years in the development of policies and initiatives in the field of digital economy, AI, and social innovation in emerging economies, with a special focus in Latin America. Moreover, she is co-founder of the global initiative AI for Climate.

Her passion for building communities and creating regional platforms has led her to lead the AI Governance working group of WEF's Global Future Council on AI for Humanity; co-found and chair the IEEE Global Partnership Latam Circle; be invited as part of the High-Level Group for the drafting of UNESCO global instrument on AI ethics; and assist the Inter-American Development Bank (IDB) in the founding of the AI for social impact regional initiative, fAIr LAC; among other key initiatives.

Her work has been featured in international media and has been recognized by the Paris Peace Forum, the WEF and the Government of the United Kingdom, among others.

Towards a Regional AI Strategy in Latin America

By Jean García Periche

The 2020 global crisis paralyzed the world and made us all rethink the future of our species. The pandemic forced us to digitize and accelerate the adoption of Artificial Intelligence (AI). As AI disruption becomes more present, it is crucial for the international community to coordinate comprehensive strategies that respond to the global challenges posed by cognitive technologies. In Latin America, there is still a lacking state of consciousness when it comes to AI governance, implementation, and deployment. As the rest of the world rapidly acknowledges the centrality of AI in governing the future of humanity, Latin America struggles in coming to terms with a compelling narrative that can leverage the power of computer intelligence.

Although there is a growing number of tech unicorns in the region, AI implementation projects are often small scaled and rarely transcend the pilot stage. The massive AI-driven disruptions in society will create deep structural changes in the economy. If Latin America does not move fast, it may risk falling into irrelevance. While AI advances at exponential rates, LATAM's power structures remain vertically driven and highly inefficient.

However, there is hope in the midst of chaos. With a population of more than 600 million, Latin America is an ideal place to develop machine learning systems that can be deployed at scale and harness the value of abundant data. In this way, Latin America can be key to defining the future of this technology and become a new international agent that has real power in the global governance of Artificial Intelligence. Among the 100 countries best prepared to use AI, 15 are in Latin America.

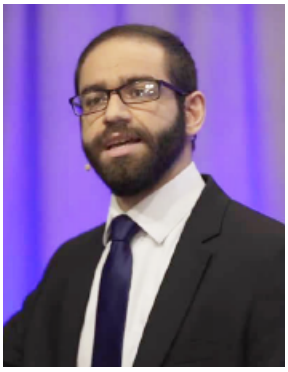
As some States start developing and implementing national AI strategies and digital transformation frameworks, this past year has seen remarkable progress in some Latin American countries. In terms of digital government, Colombia is the third most advanced country within the OECD. Similarly, Mexico, Uruguay, Chile, and Brazil are taking important steps in AI-readiness and crucial governance issues such as data privacy. Nonetheless, the limited voice and participation in global AI governance forums from Latin America is astounding. If LATAM criteria are not included within emerging policy standards, future establishment of AI governance could hinder progress in Latin America.

A fundamental element to note is that national initiatives are necessary, but not sufficient. Without unifying criteria and standardizing frameworks, no single country in Latin America can become an AI leader by itself. Latin America needs regional cohesion and a sound level of political unity. In this decade, the region needs to build a strong international coalition around a Regional AI Strategy to integrate this technology as an essential tool for leapfrogging into a new stage of development.

By making AI a priority, Latin America will be able to upgrade its socioeconomic and political systems to the XXI Century. To do that, it needs unity and international cooperation. In the Age of AI, unity in Latin America is synonymous to survival.

ABOUT THE AUTHOR

Jean García Periche



Jean García Periche is the co-founder and Chief Government Officer of GENIA Latinoamérica, a research and development (R&D) regional platform with the mission of including Latin America into the global development of Artificial Intelligence. Jean is also a Fellow from Singularity University at NASA Ames Research Center, and works as an Advisor and strategic foresight researcher to senior-level officials in the public sector. Previously, he founded Global Neo, an organization to develop new global governance models with the implementation of blockchain-based systems and decentralized token economies. He is also the President of the Centre for Political Innovation in the Dominican Republic, working on smart cities and the sustainable development of communities through circular economy models.

AI Policy Making as a Co-Construction and Learning Space

By José Guridi Bustos

All over the world countries have developed (or are developing) AI policies and strategies. It looks like there is some kind of imperative to plan how to take advantage of this third AI summer, however, it is not clear why or how. The latter is evident if we analyze strategies over the world, finding many focuses (e.g., research, development, ethics), governances (e.g., public, private, hybrid), processes (e.g., top-down, bottom-up, policymakers-centered, academia-centered) and so on.

All the fuss is understandable when we realize that AI is a general purpose technology that is heavily shaping (and being shaped by) society. In the next five to ten years, AI will drastically change the way we perceive and process information, work, interact among each other, and many other fundamental components of our lives. Thus, policymakers feel the urge to foster socioeconomic development and to regulate AI to prevent people from harm. However, even though this anxiety exists, there is little clue on how to do it right.

When developing AI strategies, the epistemic hierarchy, in which scientists and policymakers try to safeguard their credentials and authorities by demarcating their expert knowledge versus "lay" knowledge emerges. The latter has proven to be problematic for democracy [1] and frames policy construction within a deficit model of innovation in which lack of R&D is the main obstacle to overcome [2]. This framing can fail to recognize AI as a sociotechnical construct, in which the main question should be, how we will be as human beings along with it. Additionally, this framework might lead to policy instruments disconnected from a thorough policy discussion about means and ends of the AI policy.

In Chile, social riots in 2019 and the COVID-19 pandemic opened a window of opportunity to design a participatory methodology for building the AI National Policy, since they lowered resistance to public participation, especially in policymakers and authorities. We developed an open call for self-convocated roundtable discussions, in which anyone anywhere could participate and contribute, being the only requisite to use the proposed axis as guide (i.e.: First, Enabling Factors; Second, Development and Adoption; Third, Ethics, regulatory aspects and socioeconomic impacts). This process was both a co-construction exercise towards the first draft, and learning space where academia, industry, government and society interacted, taught from their expertise and learnt from each other. More than 7,000 people participated in discussions and webinars and had the opportunity to contrast their discussion with the draft in a recently closed public consultation.

In thinking about AI governance for the following decades, experiences like the Chilean process should be analyzed and improved in order to design institutions that acknowledge the sociotechnical nature of technology. Countries, in particular emergent ones, should design open processes and build their own development models from their strengths and weaknesses, refraining from just importing international experiences.

References

[1]Jasanoff, S. (2005). Judgment Under Siege: The Three-Body Problem of Expert Legitimacy. S. Maasen & P. Weingart, Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making (pp. 209-224). Springer Netherlands. https://-doi.org/10.1007/1-4020-3754-6_12

[2]Pfotenhauer, S. M., Juhl, J., & Aarden, E. (2019). Challenging the “deficit model” of innovation: Framing policy issues under the innovation imperative. Research Policy, 48(4), 895-904. https://doi.org/10.1016/j.respol.2018.10.015

ABOUT THE AUTHOR

José Guridi Bustos

José Guridi is the Chief of the Future and Social Adoption of Technology Unit at the Chilean Ministry of Economy, Development and Tourism. Previously he served as an advisor at the Future Team of the Ministry of Science, Technology, Knowledge and Innovation, where he led the creation of Chile’s Artificial Intelligence National Policy. José is also an adjunct professor of the School of Engineering at Pontificia Universidad Católica de Chile.

José is part of the OECD network of experts on AI, the fAIr LAC initiative of the Interamerican Development Bank and affiliate of The Future Society. He holds an engineering degree with a Master of Sciences in Industrial and System Engineering from Pontificia Universidad Católica de Chile.

Part VI Emerging Initiatives from China

Artificial Intelligence and International Security: Challenges and Governance

By FU Ying

In recent years, the rapid advance of artificial intelligence (AI) technology has brought about enormous opportunities. As technological revolution often comes with unforeseeable security challenges, special attention must be given to the moral and technological hazards of AI weaponization. It has now become urgent for mankind to consider how to effectively balance the benefits of the technology and the security risks of its weaponization, and how to find appropriate pathways for AI governance.

As it stands, while AI-enabled weapon systems can have powerful military effects, they are not fully reliable and potential challenges associated with their applications abound. AI has inherent technical defects which may make it hard for attackers to restrict the range of their strikes, thus exposing the attacked to excessive collateral damage and causing unintended casualties of civilians. Since big data-based algorithms and training data sets may inevitably introduce biases into real application AI systems, and training data sets may be contaminated by other countries, AI may provide wrong recommendations to decision-makers and mislead military commanders into making wrong deployments. Moreover, AI's deficiencies in interpretability, learning, and common sense will magnify the risks of battlefield conflicts during human-machine collaboration and even stimulate spiral escalation of international crises.

International cooperation is essential if humanity is to tackle the common challenge of the global governance of AI, which cannot be addressed by any country alone. Countries need to exercise restraint in the military field and work together to build international governance mechanisms in this regard. Assisted decision-making systems that are not cognizant of responsibility or risk should be prohibited. When AI-enabled weapons are used, the scope of damage by their strikes must be limited so that collateral damage and escalation of conflict can be prevented. The development and use of AI-enabled weapons must conform with existing norms of international laws. The data security of AI should be given high priority, and the whole process - from data mining and collection, to data labeling and classification, and data use and monitoring - should be regulated and restricted so as to prevent the forming of wrong models which may causing decision-makers to make wrong judgments.

The current stage represents a critical window of opportunity for establishing international norms on AI security. It is important for China and the US to have dialogue and

cooperation in this regard as they may be able to contribute wisdom for collaboration in AI governance at the global level. The two countries should start official discussions on how to establish international norms and regimes, explore areas of cooperation on the basis of their respective interests and concerns, exchange and translate relevant documents, and carry out policy dialogue and academic exchanges. Such efforts will also help to reduce potential risks to the bilateral relations and the global security.

ABOUT THE AUTHOR

FU Ying

Chairperson, Center for International Security and Strategy, Tsinghua University (CISS). She is Vice-Chairperson of the Foreign Affairs Committee of China's 13th National People's Congress (NPC).

FU Ying started her career with China's Ministry of Foreign Affairs (MFA) in 1978 and had long engaged in Asian affairs. She served successively as Director of Division and Counselor in Asian Affairs Department of MFA. In 1992 She joined UN peacekeeping mission in Cambodia. She was appointed Minister Counselor at Chinese Embassy in Indonesia in 1997, Chinese Ambassador to the Philippines in 1998, and Director General of Asian Department of MFA in 2000. She then was appointed Ambassador to Australia (2004-2007), and Ambassador to the United Kingdom (2007-2009). She served as Vice Minister of Foreign Affairs for European Affairs and then for Asian Affairs (2009-2013).

FU Ying was elected deputy to China's 12th (2013) and then 13th (2018) National People's Congress (NPC). She was the Chairperson of the Foreign Affairs Committee and spokesperson of the 12th NPC (2013-2018). She is now a Vice-Chairperson of the Foreign Affairs Committee of China's 13th NPC (2018-2023).

FU Ying also serves as the Chair of Center for International Security and Strategy (CISS) since 2018, and appointed an adjunct professor at Tsinghua University in 2018. She was appointed the Honorary Dean of the Institute for AI International Governance of Tsinghua University (I-AIIG) in 2020.

China Continues to Promote Global Cooperation in AI Governance

By ZHAO Zhiyun

In 2020, the sudden outbreak of COVID-19 pandemic is not only threatening people around the world, but also generating great challenges to global governance. Various AI-based solutions, are not only playing positive roles in tracing the origin of the virus, preventing and controlling its outbreak, and researching and developing vaccines etc., but also speeding up "contactless economy" such as online shopping, online education and tele-medicine.

Global collaboration is needed to combat the COVID-19 pandemic. So does the AI governance, especially when AI technology application has been accelerated across the world. Since 2020, to implement Development Plan for A New Generation of Artificial Intelligence and the Governance Principles of the New Generation of Artificial Intelligence: Developing Responsible Artificial Intelligence, China is actively advocating for international cooperation in AI governance while supporting more research and enriching practices on AI governance. The highlight of this year's G20 summit, President Xi Jinping proposed to hold a workshop on AI when appropriate to implement G20 AI principles. Chinese government has also launched the Global Initiative on Data Security, which, being the critical area of AI governance, clearly calls to address issues of data governance by means of co-discussion, co-establishment and co-sharing.

These measures and initiatives have fully demonstrated the consistent efforts that China has made in global cooperation of AI governance. In the future, by focusing on issues of AI governance, China will continue to establish platforms, expand channels and widen international cooperation and exchange to get broader consensus to promote AI development in a healthier manner, together with players all around the world.

ABOUT THE AUTHOR

ZHAO Zhiyun

ZHAO Zhiyun, PhD in Economics, Professor, Doctoral Supervisor, the Party Committee Secretary of Institute of Science and Technology Information of China (ISTIC), Director of New-Generation Artificial Intelligence Development Research Center of Ministry of Science and Technology of the People's Republic of China (MOST). ZHAO Zhiyun is granted with the Special Government Allowance provided by the State Council, and selected for "New Century Million Talents Project", National Cultural Expert and Theorist of "Four Groups" and Leading Talent of the "Ten Thousands Talent Plan". She is well-known as a leading talent in economic theories and policies, and S&T management and policies. She especially has unique insights in emerging technology and industrial development. She pays great attention to the issue of AI governance, and focuses on promoting related research and cooperation between China and other countries. She has won outstanding achievements in the construction of theoretical system, in the promotion of technological progress, and in the related disciplinary construction. She has published more than 30 academic monographs, 4 Chinese translations, and more than 130 academic papers. As the Principal Investigator, she takes charge of nearly 30 national, provincial and ministerial research projects, including National Key Research and Development Project, National Sci-Tech Support Plan and National Soft Science Major Project.

Steadily Taking Off: China's AI Social Experiment Is in Full Swing

By SU Jun

At present, new technologies, new applications, and new business formats such as AI, big data, and the Internet of Things are in the ascendant. Particularly, at this critical moment when humanity unites to collaborating to fight against COVID-19 and to struggle for high-quality economy recovery in the midst of the global pandemic, AI has shown the tremendous power to promote economic development and social reconstruction, which has become one of the most powerful driving forces for the world to march forward. However, from the perspective of socio-technical and public policy theory, and public perception, AI has not only empowered development, but also arisen many challenges in the governance field, such as law, privacy, ethics, and security. The evasion and resolution of these problems requires research through evidence-based methods to explore and grasp the comprehensive social impact of AI on human society.

Since experts and scholars in China launched the initiative in 2019, all parties have responded actively. We have used social experiments to study the comprehensive impact of AI and carried out a large-scale investigation based on evidence. In 2020, overcoming the influence of the pandemic, the AI social experiment in China has successfully gone through overall planning, top-level design, prioritized development, demonstration and deployment. The academic research, organization and team building, talent training, base construction of AI social experiments have been carried out in an orderly manner throughout the country. The full cycle of public policy research, i.e.: "theoretical research - policy implication - decision making - policy implementation - organization engagement" has been fulfilled.

From "proposing an initiative" to "taking off steadily", the beginning of the AI social experiment is inseparable from the broad consensus of building an intelligent society with humanism. An intelligent society with humanism is a people-oriented society, with highly developed science and technology, wide application of intelligent technology, comprehensive balance of instrumental rationality with value rationality, harmonious coexistence of people - environment - technology, open and inclusive society atmosphere, and humanistic spirit. This consensus is rooted in China's humanistic tenet of developing emerging technologies, which originates from the vivid practice of AI social experiments, and leads to the vision of a common intelligent society for mankind in the future.

Driven by this social consensus, the central government coordinated the layout, commissioned the AI social experiment expert group, and compiled the AI social experiment planning. Tsinghua University, Zhejiang University, Peking University, Beijing Normal University, Renmin University of China and other universities have taken actions to establish ethical norms and operating procedures for AI social experiments. Pilot experiments in urban governance and rural e-commerce have been carried out. Natural experiments, quasi experiments, questionnaire surveys and other experimental methods have been used to study the impact of AI applications on individuals and social organizations. Under the premise of scientific sampling and ethical review, the Chinese academia, industry and government cooperated closely to participate in areas including public health, education, pension system, environmental protection, urban governance, agricul-

ture and rural development in provinces and cities including Beijing, Shanghai, Zhejiang, Guangdong and Hubei. More than 10 provinces and cities have built hundreds of AI social experiment scenes, and formed a batch of innovative and outstanding cases. Therefore the foundation for the development of long-term, wide-field, multi-disciplinary AI social experiments has been built up.

The intelligent society is the future we will live in, and also the blue ocean we have never reached. Now, AI social experiments have taken off steadily, and related work has also been carried out in an orderly manner. Under the guidance of human-oriented values and scientific evidence-based research paradigm, AI social experiments will help human society transform from the age of industry to intelligence. The AI social experiment will provide objective and accurate factual basis and humanistic solutions for the successful transformation of the times.

ABOUT THE AUTHOR

SU Jun

SU Jun is the Cheung Kong Scholar Chair Professor in School of Public Policy and Management at Tsinghua University. He serves as the dean of Institute of Intelligence Society Governance（ISG）, Tsinghua University, the director of the Center for Science, Technology and Education Policy (CSTEP) at Tsinghua University and the director of Think Tank Center of Tsinghua University. He is the deputy director of the Advisory Committee of the Public Administration under the Ministry of Education.

Artificial Intelligence Governance Requires "Technical Innovation + Institutional Innovation"

By LI Xiuquan

The accelerated development and application of artificial intelligence are putting urgent demands on governance in ethics, security, privacy, fairness, and other aspects. New laws, institutional design and governance rules are needed to guide and standardize technological development. The issue of artificial intelligence governance is not only a matter of system design, but also a matter of technology research and development. The increasingly urgent requirement for technological innovation in artificial intelligence governance needs to be addressed.

Each type of governance problem contains many technology innovation needs. The current big data-driven machine learning methods represented by deep learning uses a black box model. The internal structure and mechanism of the model are not transparent to users, and lack an understanding of the increasingly large and complex machine intelligence. When there is an error, it is very hard to trace the origin and explain, which brings great difficulties to the system designer in legal and normative design. The research of more transparent model algorithms and the development of interpretable, understandable, and predictable intelligent algorithms will provide technical approaches for solving the dilemma of responsibility identification and overcoming ethical constraints.

Similarly, sample attacks, sensor interference, and deep learning framework vulnerabilities may also bring challenges to the security of intelligent systems. A safe and reliable artificial intelligence system should have strong security performance, can effectively respond to various deliberate attacks, and avoid security accidents caused by abnormal operations and malicious attacks; governance such as privacy violations and discrimination or prejudice also requires technology innovations, such as user privacy data desensitization, federated learning, and small data incremental learning to cope with these risks and challenges, and enhance public's trust in AI technology.

In order to meet the challenges of artificial intelligence governance, the integration of technological innovation and institutional innovation has been deepening in 2020. Artificial intelligence technology developers are trying to devote more and more energy to the development of technical solutions to security, privacy, and fairness issues. In the future, the development of responsible artificial intelligence technology should become the next important direction for theoretical innovation and technology development. It is necessary to solve governance problems through technological approaches, and share responsibilities with legal and administrative parties to jointly guarantee the healthy development of the AI industry.

ABOUT THE AUTHOR

LI Xiuquan

Dr. LI Xiuquan is now Research Fellow of Chinese Academy of Science and Technology for Development (CASTED), and Deputy Director of New Generation Artificial Intelligence Development Research Center of Ministry of Science and Technology. He received his Ph.D. degree, in field of Computer Science, from Tsinghua University. He is also joint PhD in Information Science, University of Hamburg, Germany. He has many years of research experience in AI fields, such as multidimensional time series data modeling and prediction, and brain-controlled robot system based on EEG. His current research area is big data and AI technology foresight and evaluation, industrial technology roadmap and AI innovation policy research. He has strong interest in the study of the frontier trend of intelligent transformation, and the demands for innovative policies in various aspects of AI development such as research, industry and governance. He has presided over 10 research projects such as “Research on the Major Strategic Issues of Chinese Intelligence Economy and Intelligence Society development”, “Research on the Leading Trends and Policies of Artificial Intelligence at Home and Abroad”.

Developing Responsible AI: From Principles to Practices

By WANG Guoyu

2020 is a year in which artificial intelligence continues to develop rapidly in the context of the global epidemic, and also a year in which the ethical principles of artificial intelligence has become ever more abundant. By April 2020, more than 160 guidelines have been included in the AI Ethics Guidelines Global Inventory developed by the nonprofit organization Algorithm Watch. Publishing bodies involve various kinds of institutions and organizations, including governments, academia, enterprises, and so on. Many of these ethical principles are directly titled "Responsible AI". In view of this alone, the development of "Responsible AI" can be said to have become an international consensus.

However, the development of "Responsible AI" requires not only principles and appeals, but also actions and practices. To this end, it is necessary to clarify the subject and object of responsibility, as well as the authority of responsibility, and to address the issues of who is responsible, for which is responsible, and to whom is responsible. For example, for issues brought about by the applications of AI such as privacy breaches, discrimination, and liability in autonomous driving, who is the responsible subject, the designer, the engineer, the corporate decision-maker, or the regulator? Is it the individual, the institution, or the state? When an action involves multiple subjects from different institutions, how should responsibility be allocated? When corporate interests and social ethics conflict, how should they be prioritized? In particular, when the development of AI involves national core competitiveness on the one hand and conflicts with ethical principles on the other, should it be developed (e.g.: military applications of AI) and how should it be developed (e.g.: face recognition technology)? Who has the authority on accountability and how is it done? These questions cannot be avoided if the ethical principles of AI are to be brought to ethical practice.

To answer these questions, we need to develop further theoretical discussion of these issues. However, it is more important to clarify the issue of responsibility in the practice of developing AI and to promote the construction of responsible mechanisms and the culture. To this end, the China Computer Federation (CCF) established an interdisciplinary "Committee on Ethics and Professional Conduct" (CEPC) in 2020, with computer experts and philosophers co-chairing and being committee members. At the China National Computer Congress (CNCC) 2020, the committee held a special forum on "More Harmony for the World-Professional Ethics in Information Technology". On the one hand, not only the history and dimensions of computer - and AI professional ethics in the era of information technology, but also the theoretical path to develop responsible AI, were analyzed from the theoretical level. On the other hand, the path to the development of responsible AI, the AI ethics education in universities and the framework of artificial intelligence agile governance were also explored and discussed in depth. The Young Computer Scientists and Engineers Forum (CCF YOCSEF) of the CCF also discussed how to incorporate fairness, transparency, acceptability, and sustainable development into the design of the third generation of artificial intelligence. The CCF is trying to promote the conversion of responsible principles of AI into responsible practices for engineers at four levels: ethical, structural, educational and technical.

To follow the development of “Responsible AI” into actions, it is also important to consider that AI cannot be viewed as just an isolated, closed system of algorithms and technologies. The ethical issue of artificial intelligence is an ethical issue in the social-technical system where humans and AI live in symbiosis and technology and society interact to shape each other. The development of “Responsible AI” needs to face not only the technical ethics of AI and the professional ethics of engineers, but also the uncertainty and social-ethical challenges in the social context that includes multiple dimensions such as economics, politics, and culture. In particular, it is important to make AI benefiting the society in the framework of global ethics. Therefore, it is necessary not only to awake the moral sensibility of AI practitioners and strengthen the research on the responsibility and norms of AI in application scenarios, but also to continue to promote the international dialogue on AI ethics and the construction of transnational accountability mechanisms.

ABOUT THE AUTHOR

WANG Guoyu

Dr. WANG Guoyu is a Professor of Philosophy at Fudan University. She is also a Distinguished Professor of “Chang Jiang Scholars” of MOE. She serves as the Director of the Center for Biomedical Ethics and Applied Ethics at Fudan University. Prof. WANG is also Vice President of Committee on Ethics of Science Technology and Engineering of the Chinese Society for Dialectics of Nature (CSDN), the co-chair of the CCF Committee on Ethics and Professional Conduct (CEPC), and the director of the ethics committee of the International Human Phenotype Group Project. WANG Guoyu has been engaged in high-tech ethics and governance research for a long time. She conducts as PI several National Key Projects.

Promote the Formation of "Technology + Regulations" Comprehensive Governance Solutions

By WANG Yingchun

In recent years, many AI governance principles and initiatives have been issued by relevant organizations and institutions around the world. It has become an international consensus to promote the "Human-Centered " and "Responsible" development of AI. The transformation of these consensus principles into action plans has become a general trend.

On July 10, 2020, the Governance Forum of the World Artificial Intelligence Conference was held in Shanghai. With the theme of "Developing Responsible Artificial Intelligence", the forum discussed a comprehensive range of governance solutions, centered around the theme of "Technologies + Regulations" as a combined driving force to practice the ethical principles of artificial intelligence. At the forum, the "Working Plan for Collaborative Implementation of AI Governance Principles" (9 items of Shanghai AI governance collaborative action) was released, which consists of a framework - "one platform, four tasks and four systems".

The "one platform" means to build a platform for global network of cooperation and exchange. We should form a global community for AI governance research and cooperation. Drawing on multi-cultural wisdom, exploring accessible and inclusive safety assurance and solutions for a win-win outcome, we need to build a multilateral, multidisciplinary and multi-agent consultative governance mechanism, in order to promote the formation of a global joint effort to construct s a common AI governance system.

The "four tasks" are goals to carry out standard specifications, industrial self-monitoring, optimal practice and trustworthy technologies. First, all stakeholders should cooperate to research and formulate technical standards and application specifications, clarify the threshold for industrial access, and ensure the legitimate rights and interests of relevant parties. Second, to establish ethical commitment and review system of scientific research and enterprise products, and form the compliance and self-discipline process and operation guide within the industry. Third, constantly summarize cases and relevant experiences ofhigh-quality practices, extract specifications from practice, and then guide the industry through best practices. Finally, we should develop trustworthy technologies on the basis of transparent algorithms and privacy protection, develop solutions for ethical evaluation and other relevant regulations studies, and promote trustworthy solutions through the cooperation between public and private sectors.

The "four systems" include evaluation system, regulatory system, talent system and security system. We should build a system of compliance indicator, methods and platform of evaluation, to carry out governance qualification assessment and classification certification; to improve an open and transparent cross-departmental collaborative regulation system to realize the supervision of related products and applications; promote interdisciplinary research, set up compulsory courses for governance and cultivate talents with multi-displinary skills; to build a social security system to cope with the impact of artificial intelligence on employment structure, to develop relevant insurance products and to enhance public digital literacy.

To realize the "human-centered" vision, we need compre-

hensive governance solutions that integrate technologies and regulations. We call for global peers to work together to develop products and solutions that are more adapted to social values around typical scenarios, and actively form consensuses and solve specific problems in positive and steady practice.

ABOUT THE AUTHOR

WANG Yingchun

WANG Yingchun, PhD, Head of Research Department of Science, Technology and Society at Shanghai Institute for Science of Science, areas of expertise include innovation transformation and innovation governance, and science, technology and society. He initiated and organized a multidisciplinary AI research group to conduct systematic research on AI. He has undertaken a number of consulting projects entrusted by the Ministry of Science and Technology and the government of Shanghai municipality, and has continuously participated in the research and policy drafting of the government's AI policy. He led the organizing work of the Governance Forum under World AI Conference in Shanghai. At the moment, he is also responsible for the running of the Secretariat of the Expert Advisory Committee of the National New-generation AI Innovation and Development Pilot Zone in Shanghai.