RegHorizon

PEOPLE, PROJECTS, PARTNERSHIPS

ETH zürich

logitech AUS AMERICAN UNIVERSITY IN SWITZERLAND DiPLO

AI POLICY CONFERENCE

SUMMARY OF EVENTS

NOV 16-17, 2020

Contents

RegHorizon & ETH Zürich: AI Policy Conference Summary of Events

November 16-17, 2020

Abstract

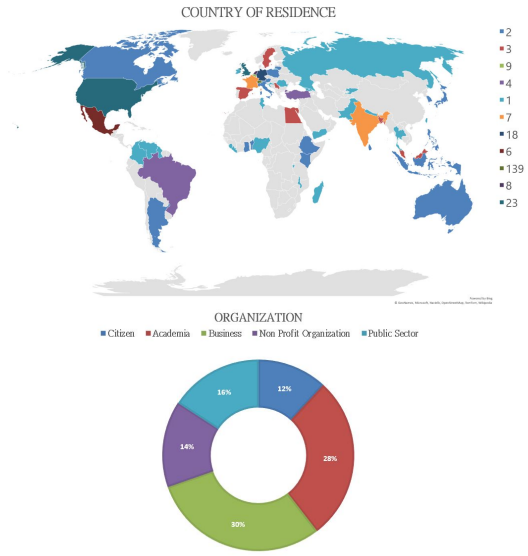
Can policy be used as a tool to build trust in Artificial Intelligence (AI) technologies? How can we align AI Regulation across different geographies? How can society get AI-ready and shape a better future for all stakeholders?
To address these issues, *RegHorizon*, in collaboration with ETH Zürich, created an unbiased platform for a timely multi-stakeholder discussion among policymakers, academia, business, and society. This paper provides an informed summary of the main conference takeaways, panel discussions, and academic review sessions on research in AI and AI governance.

Keywords— artificial intelligence, policy, governance, technology, regulation, data, ethics, decision-making, algorithmic bias, trust, trustworthiness, transparency, accountability

1 Introduction

Technologies based on Artificial Intelligence are a force for good and have the potential to address many societal challenges. Yet, they pose many threats and concerns, including to our privacy and social justice. It is thus crucial to think about how AI Governance can be used to uphold trust and ensure the trustworthiness of those who develop and deploy AI solutions to build towards a better digital future for all.

To kick-start this needed discussion, RegHorizon, a Swiss-based strategy consultancy firm, together with ETH's Center for Law and Economics, hosted an interactive online conference on AI Policy that took place on November 16 & 17, 2020. Participants had the opportunity to engage in an open exchange of best practices in regulatory management and stress-test new AI governance ideas alongside international business leaders, government representatives, and academics. There were around 400 registrations, with participants coming from varied geographies and occupational backgrounds (see figures below).



Across five different panel discussions and two academic paper review sessions, the conference explored three main questions:

1. What role can policy frameworks play to instill trust in AI and protect society and businesses against harmful misuse?

2. How can we address geoharmonization of AI policy to manage cross-border business risks, enhance competitiveness and foster innovation?

3. How to get AI ready as an individual and as an organization to maximize the benefits from this powerful tool for future generations?

Section 2 of this publication provides detailed individual summaries of the five panel discussions, and an outline of the key-note speech delivered by Dr. David Weinberger. Each panel summary focuses on the core challenges in the current AI regulatory framework identified by panelists and is complemented by the panelists' suggestions for ways forward in AI regulation.

Section 3 of this publication summarizes the key challenges and next steps in AI regulation that were discussed across different panels. The challenges and suggestions are grouped under sections pertaining to regulatory approaches, capacity building, business concerns, and ethical considerations around development and deployment of AI. This section also provides RegHorizon's objectives on setting up multidisciplinary projects based on selected findings of the conference.

Section 4 comes as a separate segment as it summarizes the academic papers presented as part of the AI Governance Research Workshops hosted by ETH Zürich. The paper presentations are organised in four subsections depending on the topic. Section 4.1 covers papers on Human-AI interaction, and section 4.2 covers papers on Natural Language Processing and Blockchain. The academic papers in section 4.3 are concerned with regulatory issues in AI, and section 4.4 summarises papers discussing the technical aspects of AI.

2 Panel Discussions

The AI Policy Conference was launched with a Panel on *Fostering Innovation and Growth*. The core challenges in balancing innovation and growth and the elements of effective AI policy were addressed, including those that promote competition and mitigate business risk.

Panel 2 dealt with *Data Privacy and Consumer Protection* including the differences in dealing with regulating data versus regulating AI. This panel further underlined the need for regulations to adapt quickly to the developments of AI and other new technologies.

The key-note speech by Dr. Weinberger highlighted the main challenge for regulators of having to regulate the unknown. According to Dr. Weinberger, regulators should increase their understanding of technology and be inclusive in the policymaking process in order to regulate effectively.

Panel 3 on *AI Policy Geo-Harmonization* summarized the current status of policy developments in the EU, which was followed by a call for more inclusive discussions. The needs of both large and small businesses were addressed and panelists discussed a few steps that can be taken to become more efficient on global cooperation in AI policy.

Panel 4 *AI Policy in Healthcare* highlighted the ways to increase adoption of AI in healthcare, including the role of policy. Conference participants had the chance to hear from clinicians on how education, culture, and cross-disciplinary focus when developing AI can support adoption. Further challenges addressed were data-management in healthcare and how to make data high-quality, more accessible and more meaningful for developers, reviewers, or regulators.

Panel 5 on *Getting Ready to Manage Business Risk and Complexity* explored the specific roles of governments and educators, innovators, businesses and citizens in ensuring better outcomes and faster deployment of AI solutions.

2.1 Summary of Panel 1: Fostering Innovation and Growth

List of panelists:
• Katarzyna Gorgol, Digital Affairs at EU Delegation to UN,
• Aldo Podestà, CEO at L2F,
• Dr. Jochen Friedrich, Technical Relations Executive, IBM,
• Dr. Ron Chrisley, Centre for Cognitive Science, The University of Sussex, UK,
• Dr. Christian Busch, Swiss State Secretariat for Education, Research and Innovation,
• Dr. Jovan Kurbalija, Executive Director of DiploFoundation and Head of the Geneva Internet Platform.

Panel Moderator:
• Dr. Katharina E. Höne, Research Associate in Diplomacy and Global Governance, DiploFoundation

Panel Topic: What role can policy play to help manage business risks, enhance competition and foster innovation and growth?

The first panel of the conference kicked off with a discussion on the most relevant elements of an effective AI policy from a European perspective. According to Ms. Gorgol, that is the policy that successfully promotes innovation. The EU Commission's goal to establish an AI ecosystem of excellence and trust in Europe is well documented by the EU Commission's 'White Paper on Artificial Intelligence'

published in February 2020 [1]. The key objectives revolve around embracing the strengths of academic institutions and established organisations, and encouraging the private sector to invest in AI made in Europe. Other elements include a focus on increased investment, availability of data, and help to the SMEs.

Mr. Podestà elaborated on the impact regulation has on innovation and growth from an entrepreneurial perspective as it can be instrumental to ensure product and industry quality. In his words, when regulation caters to ethical concerns, it increases the quality of an industrial ecosystem, which in turn increases trust in AI by users and ultimately creates business opportunities.

Regarding managing business risks in AI, Dr. Friedrich stressed the need to identify highly critical areas where the impact of AI solutions is high, such as in democratic decision-making, infrastructure, and ownership of data. He mentioned that although algorithms themselves may be difficult to regulate, their inputs and outputs can be regulated more easily. Performance standards, and standards in the area of trustworthiness, explainability and prevention of bias are needed. It is also important to ask whether AI technology follows existing regulation or whether new regulation is needed.

A recurring problem when talking about AI is the term's inclusive nature as it refers to technologies involving big data, automation, statistics, and algorithms. In the words of Dr. Chrisley, such a terminological confusion limits the precision and variety of instruments used in future policies.

According to Dr. Busch, the main objective of Swiss AI Policy is to provide the best framework conditions for AI research, innovation and the use of AI applications, while at the same time address the critical challenges of transparency, bias, and liability. Switzerland applies the tech neutrality principle, and leaves large freedom to individual actors while binding them to respect all existing laws.

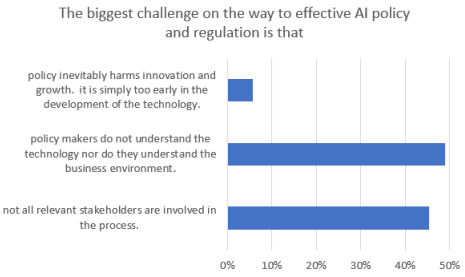
Dr. Kurbalija's perspective was that AI is an amplifier of the digital divide, in the same way as other digital technologies. While the situation in developing countries is improving, the gap between the developed and developing world is widening. This divide is growing not only between countries but also within societies, across generations and professions. Moreover, developing countries are being asked either to apply AI technologies without questioning them; or being left out of the AI policy debate altogether.

One point of discussion encouraged by the panelists was the importance of distinguishing between policy and standards. Standards are developed by groups of specialists, while policy has to go through a legislative process, which means that working solely on standards could go against the principle of inclusion. It was agreed that standards are generally a way to implement policy objectives, and often policy and standards complement each other.

On a final note, over the past years, we have witnessed the use of AI technologies becoming more widespread in designing policies and engaging with regulators. According to Dr. Chrisley, AI systems assisting with policy design about AI technologies themselves are likely to become a reality soon.

The poll results from Panel 1 are provided below.

[1] White Paper on Artificial Intelligence: A European Approach to Excellence and Trust, *European Commission*, 2020



Suggestions from Panel 1 on ways forward on AI regulation:

• The process of developing policies needs to be inclusive, and regulators need to speak to the broadest possible range of stakeholders. The communication channels with the businesses need to be improved. In the EU, this has already been done through many platforms such as the EU Alliance in AI or by engaging in cross-public consultations.
• Regulators should prepare in advance for the next generation of AI systems to not lag behind when new technological developments occur.
• Regulators should look at high-risk areas of impacts of AI, and how existing legislation can be adapted as new tech is developed.
• Countries need to develop educational competencies and ensure that people can understand and deal with new tech.
• It is crucial to identify the new component(s) of AI, and separate the new technologies' concerns from statistical problems such as bias and robustness.
• To reduce the digital divide gap, policymakers could use SDGs as benchmark objectives for AI developers against which to measure the performance of the AI algorithms regarding important societal issues.
• The diplomatic world has a role to play to ensure a smooth interplay between standardization and regulation.

2.2 Summary of Panel 2: Data Privacy and Consumer Protection

List of panelists:
• Paul-Olivier Dehaye, Founder of PersonalData.IO,
• Dr. Jan Kleijssen, Director for Information and Action Against Crime, Council of Europe,
• Prof. Marcel Salathé, Academic Director of EPFL Extension School,
• Miguel Amaral, Directorate of Public Governance at OECD,
• Leila Delarive, CEO at Empowerment Foundation and Amplify.

Panel Moderator:
• Elliott Ash, Assistant Professor of Law, Economics, and Data Science at the Center for Law and Economics, ETH Zürich

Panel Topic: What role can policy play to promote trust in AI and protect society against harmful uses?

According to Mr. Dehaye, there is a lot of uncertainty about the risks associated with AI technology. Although some regulations exist, such as data protection and privacy, the lack of enforcement is a core problem. This issue once again stresses the issue of trust in governments and whether they are effective in that role.

For Ms. Delarive, the main difficulty in creating AI policy is the lack of a universally accepted and legally binding definition of what constitutes AI. Different AI systems will have different risk profiles and should be treated differently. Ultimately, it should be the human that is responsible for the decisions made by AI systems.

Prof. Salathé pointed out the need to increase the speed and technical competency of regulators. According to him, while education has a

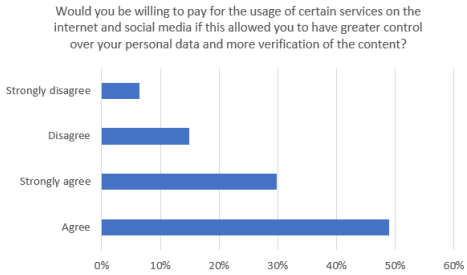
principal role in increasing the awareness and technical ability of society as a whole, the core challenge is the investment in time and money that this requires.

According to Dr. Kleijssen, ethical frameworks give guidance but are not binding. When an AI application directly affects human rights and democracy (such as the use of facial recognition by the police, or AI in the judiciary), it creates risks that should be regulated. The Council of Europe is working on making it obligatory for governments and private companies to conduct a human rights impact assessment before applying AI technology. The COE is also negotiating a treaty on AI and working on certification of AI applications in partnership with the IEEE.

There was a general discussion and different points of view on the speed with which new rules should be established for new technologies and for AI in particular. Some argued that governments already acted too fast, with the risk of compromising on key issues such as data privacy (e.g. in the case of covid-19 tracking apps), while others believed that Governments were too slow. While some groups are questioning whether to regulate or not, according to Dr. Kleijssen, "the train has already left the station." Forty-seven Governments, including EU member states, Switzerland, Russia, and Turkey have already decided that they want to regulate AI.

Mr. Amaral provided a good checklist for regulators to mitigate the risk of choosing the wrong regulatory options. Before moving ahead with regulation, governments should look at which existing regulations apply to their specific problem, the right timing and approach to regulate, and how to keep regulation purpose-suitable through constant real-time monitoring of the effects of specific regulations.

The final point of discussion revolved around the panel poll results (figure below), with most participants willing to pay additionally for their privacy to be protected. This indicates that the current privacy requirements are not being adhered to, or that the citizens do not have confidence in the regulators to enforce privacy law. The panelists agreed that regardless of whether citizens are willing to pay, customers should not be asked to pay for companies to respect the law.



Suggestions from Panel 2 on ways forward on AI regulation:

• Regulators need to increase their capacity to assess users' risks and require that information about when and how they are using AI is provided. Inspiration can be taken from California's Automated Decision Systems Accountability Act [2], which makes it mandatory for companies to indicate when their customers are interacting with AI operating assistants.
• With the aim of increasing technical know-how, governments need to allocate budgets and create educational environments to overcome time and money constraints for individuals. It is imperative that the youth, and more importantly, the middle and older generations have access to technological knowledge.
• To move forward with AI policy, regulators need to start from the ethical frameworks as a base and look at existing laws on human rights. The next step would be to develop general le-

[2]Personal Rights: Automated Decision Systems, *California State Assembly Member Ed Chau*, 2020

gal principles that apply to critical uses of AI, and move to create sector-specific regulations.
• Those who design and deploy AI technologies could be required to have a license by law, based on professional qualifications.
• Governments need to work in a coordinated manner within a country and also across national boundaries.
• Governments need to consider combining hard and soft-law approaches, e.g. making use of self-regulation, ethical business practices, and industry standards.
• Governments need to adopt a more iterative and flexible approach to AI policy. There is a need to move away from static to dynamic governance and incorporate constant review and revision for pre- and post-impact of technologies.
• Governments should leave more space for experimentation, such as via regulatory sandboxes. This is important not only for businesses to try out new methods but also for governments to learn about the new technology and its implications, both in terms of risk and opportunity.
• Governments need to be at the forefront of what is going on, e.g., through scenario planning and early engagement with the business community.

2.3 Summary of Key-Note Speech

Dr. David Weinberger, senior researcher at Harvard's Berkman Klein Center for Internet & Society

The talk by Dr. Weinberger, a renowned author, academic, and scientist working on the effect of technologies on ideas and society, addressed the metaphysical panic caused by the Blackbox nature of Neural Nets. Dr. Weinberger mentioned that as these complex models are increasingly applied in critical scenarios from science, politics, and society, there is a desperate call for making the Blackbox transparent and explainable to humans.

According to Dr. Weinberger, it is not the Learning System that is a Blackbox, but rather the world itself. We are now seeing a Copernican moment where human rationality as the ideal model for intelligence and task-solving is challenged. Demanding transparency is a forensic tool for humans, and stochastic systems that do not share this feature of human thought outperform in many areas and will continue to grow more unexplainable to humans.

The speaker addressed two constraints to ensure the applicability of Learning Systems in society, with the first one being utilitarian/outcome-based. We can define what we care about from a human perspective, giving rise to many conflicting goals. For example, would we trade the low environmental impact of a self-driving car for affordability. The take-away in this respect is that we need to define a unified approach because individual product optimization stands in the way of global outcome optimization.

The second constraint is deontological/principle-based. A commonly used principle is fairness, which is, at first glance, simple. But as can be seen in Dr. Weinberger's example of gender discrimination, there are many ways to define fairness. The different definitions, such as equal opportunity, gender-blindness, or demographic fairness, give rise to very different recommended outcomes. The conclusion is that fairness is not simple, and we need to include everyone in the discussion on what it means.

In conclusion, the world is overall messy, and forcing Learning Systems to adhere to human

standards of explainability might take away much of their power. The challenge for regulators is to regulate what we don't understand. In order to do so, the best way forward is to get close to technology, think creatively, and be inclusive in the policymaking process.

For the interested readers who want to delve deeper into how AI technology is enabling us to take advantage of all the chaos it's revealing, we refer to Dr. Weinberger's most recent book. [3] The book explores the ways AI is changing how we think about ourselves and our most basic strategies for addressing the future: from how we approach our everyday lives to how we make moral decisions and how we run our businesses.

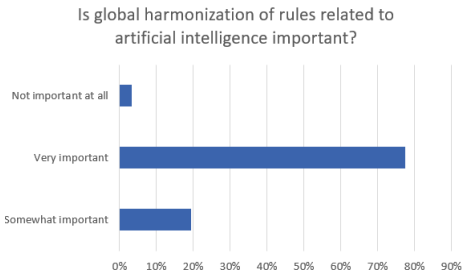
2.4 Summary of Panel 3: AI Policy Geo-Harmonization

List of panelists:
• Eva Kaili, Chair, Future of Science and Technology Panel, European Parliament,
• Ivana Bartoletti, Technical Director - Privacy, Deloitte,
• Robert Mandelin, FIPRA, former Director - General of DG CONNECT - EU Commission,
• David Campos Pavon, Group Data Privacy Officer and Vice President, Nestle.

Panel Moderator:
• Ayisha Piotti, Co-founder and Managing Partner, RegHorizon

Panel Topic: What are the main and crucial next steps in creating and implementing a cross-border coordinated global approach to AI policy?

The first panel of the second conference day opened up with a poll directed at the conference participants. The results are provided below.



This panel's discussion kicked off with why achieving a geo-harmonized AI policy solution has become increasingly essential for companies. According to Mr. Campos, clear and harmonized rules for AI help build trust in AI and help businesses decide where to invest in AI deployment.

Ms. Bartoletti noted that, although there is already an abundance of legislation on discrimination, international tools for human rights and equal opportunities across countries, regulators need to perform a fitness test of current legislation to see if we can prevent the harms of automated decision making. According to her, there is a growing alignment on the universal values that should form the basis of global policy. Privacy, fairness, and trustworthiness are increasingly being discussed and gaining importance beyond the EU, in countries such as Brazil, USA, and China.

According to Ms. Kaili, the contrasting data standards across different jurisdictions in the EU pose significant data quality challenges. Data governance is therefore a crucial priority for the EU, which has the aim of finalizing legislation on its governance by 2021. The European Parliament's objective is to improve data access for both low-risk and high-risk AI applications, improve inter-operability of data

[3] Everyday Chaos: Technology, Complexity, and How We're Thriving in a New World of Possibility, D, Weinberger, 2019, *Harvard Business Press*

across countries, and improve talent management in this field by building an ecosystem of excellence.

Mr. Campos and Ms. Bartoletti highlighted the business priority areas that regulators should concentrate on from a global perspective. AI Governance should help companies build trust in technology as technology adoption by society would otherwise be impossible. Regulators should also avoid conflicting regulation, understand the consequences of specific regulations in other areas, and facilitate effective data flows by relaxing data localization requirements. From the auditing side, guidance for companies on the due diligence process is needed, with clear instructions on the entire AI life cycle from development to deployment. These regulations should be easy to comply with to ease the burden of SMEs' compliance. Finally, citizens should be protected through an established process, including a mechanism for redress.

It was also suggested that not only should SMEs have a stronger say in the policy making process, but they should also be protected from takeovers by more prominent companies through strengthened competition laws.

Suggestions from Panel 3 on ways forward for AI regulation:

• Three ways to ensure speed for regulators are creating an ethic of mutual recognition and acceptance of each other's standards, creating networks of excellence both for AI innovators and regulators, and alignment of standards.
• Regulators need to favour an adaptive regulative approach of co-creation of regulatory benchmarks. This involves establishing some basic standards which companies can apply, and facilitating reporting on progress in an auditable way.
• Public authorities need to devote resources to monitor the co-creation of the regulatory process.
• There was a consensus that the EU is playing a leadership role in the domain of AI policy but more needs to be done in terms of moving from a regional to a global approach. Here the multilateral institutions have a role to play, and it's the quality of their output that will make a difference.

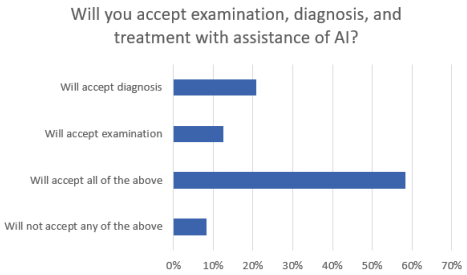
2.5 Summary of Panel 4: AI Policy in Healthcare

List of panelists:
• Dr. Ceri Thompson, Deputy Head of Unit eHealth, Wellbeing and Ageing, European Commission,
• Prof. Dr. Philippe Ryvlin, Head of Department of Clinical Neurosciences, Vaud University Hospital,
• Dr. Sean Khozin, Global Head of Data Science Innovation, Jannssen R&D, Johnson Johnson
• Gilles Lunzenfichter, CEO and Co-founder, Medisanté,
• Dr. Gabriel Krummenacher, Lead of Data Science, Zuehlke Group.

Panel Moderator:
• Sanja Fabrio, Co-founder and Managing Director, RegHorizon

Panel topic: How can policy mitigate risk and ensure safety and performance of AI solutions, to unleash the huge potential of AI in healthcare for patients worldwide?

The opinions of the attendees regarding the opening panel poll are summarized below.



This panel kicked off with a discussion on the drivers behind the adoption of AI-based solutions in healthcare and the role of policy. According to Dr. Khozin, trust is the underlying theme and one way of building the trust of patients in AI innovations is through proactive regulatory mechanisms. This way, we can ensure that AI solutions are scalable and the right incentives are accessible. Another key area where AI policy can help is ensuring that AI solutions built on healthcare data are both performative and respectful of different settings. From the perspective of an AI developer, Dr. Krummenacher mentioned that having access to vast and diverse data and a good understanding of the clinical environment in which it will be used is fundamental for designing good algorithms for healthcare.

According to Mr. Lunzenfichter, it is highly relevant for policy improvement to develop appropriate reimbursement policies for digital or AI-supported healthcare technologies, including those related to remote patient monitoring. In the European Union, large policy disparities currently exist between countries. Without adequate incentives, physicians and care providers have neither the motive nor the capacity to use new technological solutions or contribute to their development. With incentives set through proper reimbursement schemes, we might see a shift from the current consumer-centric approach to medical AI solutions, controlled by global tech companies and device manufacturers, towards an approach led by medical professionals.

Physician participation supports the clinical integrity of AI technologies, which could significantly increase trust in them. For example, clinical studies could provide regulatory authorities with an adequate basis for evaluation regarding the approval of new AI solutions. Transparent and efficient regulatory approval procedures would facilitate reimbursement later when a new solution is deployed at scale. According to Dr. Ryvlin, it is important to address these problems pragmatically while focusing on the performance of any proposed solutions. Another advantage of a physician-centered approach is that medical professionals already have considerable experience with many critical issues, including bias, data privacy, fairness, and security of sensitive data.

On that note, more representative and diverse training data would help with building more performant systems and reduce bias in algorithms. Remote patient monitoring could contribute high quality, granular data at the scale needed to train and advance AI solutions. However, leveraging these data's full potential requires that they be shared efficiently and securely with AI developers.

According to Dr. Thompson, the priorities of the EU's current regulatory framework guiding big data in healthcare are securing better access for citizens to health data, encouraging innovation around digital tools and services, and connecting health data, as per the 2018 Communication on Digital Transformation of Health and Care in Digital Single Market [4]. Organizational, technical, and semantic interoperability is crucial in enabling the secure sharing of data such as e-health records, genomic data, and medical imaging, especially across borders. To address these issues, the

[4]Communication on Enabling the Digital Transformation of Health and Care in the Digital Single Market: Empowering Citizens and Building a Healthier Society, *European Commission*, 2018

European Commission is developing principles and technical specifications to create Europeanized data spaces, including on healthcare, as laid out in the EU Data Strategy [5] adopted in 2020. A single digital market for healthcare could securely grant AI developers and innovators access to high-quality, unbiased, and representative healthcare data to improve the performance and assessment of AI algorithms.

Suggestions from Panel 4 on ways forward for AI regulation:

• To address the issue of the medical AI solution's performance, high-level clinical studies need to be performed, and reimbursement schemes need to be put in place. These schemes are a catalyst for ensuring innovation and a level playing field for smaller developers. Reimbursement is also a tool to reduce the digital divide between the 'haves' and 'have-nots' and should be considered a priority.
• We need pragmatic and proactive policies when it comes to making the required investments for reimbursement and AI research. While facilitating the adoption of those solutions we also need to make sure that they are inclusive, reduce some inherent biases and don't create a digital divide.
• To ensure the most scalable and privacy-preserving approach, the data ownership should be transferred to the individual.
• Regulators should define norms for developing, validating, and testing AI models that are risk-specific. This would give clarity to AI developers and help speed up AI development.

2.6 Summary of Panel 5: Getting Ready Managing Business Risk and Complexity

List of panelists:
• Dr. Ekkehard Ernst, President at Geneva Macro Labs Future of Work, AI Specialist at ILO,
• Dr. Alberto-Giovanni Busetto, Group Head of Data & AI, Adecco Group,
• Dr. Joanna Bryson, Professor of Ethics and Technology, Hertle School of Governance,
• Luca Brunner, Managing Director, Cognitive Valley Foundation,
• Dimitrios Psarrakis, Technology & Innovation Policy Specialist, EU Parliament.

Panel Moderator:
• Dr. Aileen Nielsen, Chair, NYC Bar Science and Law Committee, Law and Tech Fellow, ETH Zürich

Panel Topic: How can we get ready to optimize the economic & societal opportunities offered by AI while managing the risks?

According to Mr. Psarrakis, in order to optimize the benefits and limit the risks of AI, the first step for governments needs to be to develop infrastructure for data, and the second step would be to introduce standards. By introducing standards in the various technical stages of AI development, governments can reduce ambiguity, solidify supply, and help AI adoption by reducing AI demand volatility. Standards can also help make the moral implications less ambiguous by addressing privacy and ethical aspects.

According to Prof. Bryson, we need accountability for AI systems, which is achieved

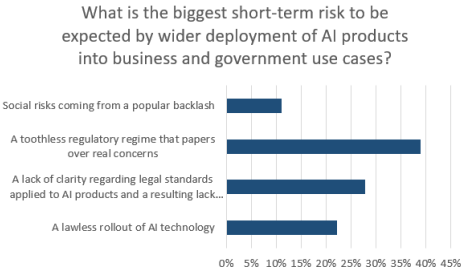
[5] A European Strategy for Data, *European Commission*, 2020

through transparency and enforcement at all AI development stages. If we manage to provide accountability mechanisms in various aspects of system development, we can also ensure the transparency of the whole process.

Dr. Ernst elaborated on the impact of AI on jobs and pointed out that the fears related to job losses in the developed world are overblown. Unlike developing countries whose economies are more labour-intensive, developed countries will be impacted much less by AI adoption. Next, the concept of 'known unknowns' arises as AI helps to increase the productivity of existing jobs by a currently unknown amount. According to Dr. Ernst, the extent to which this will happen will depend on the standards we set and the regulatory framework we provide. The networking potential of AI is the 'unknown unknown', as its potential to increase collective intelligence is currently untapped. Overall, we cannot efficiently implement the technology if we cannot bring the regulators and the service providers together.

On the topic of ensuring that AI is not only developed by a few, Dr. Busetto argued that inclusiveness in AI development is vital since it will foster diversity and mitigate biases in measurement and development. More incentives to innovate will be created by an environment that is open and diverse. Having an open environment allows for transparency and helps mitigate biases and statistical fallacies, which would otherwise go unnoticed. Finally, Mr. Brunner touched on the importance of inclusiveness for better technological governance, as inclusiveness enables interdisciplinary platforms to determine the flow of AI governance.

The results of the final poll of the conference are provided below.



Suggestions from Panel 5 on ways forward for AI regulation:

- Governments need to partner with market players and efficiently use regulatory sandboxes. This way, there can be a trustworthy market and a better general knowledge of critical topics such as development bias, robustness, and biometric systems.
- Regulators need to enforce processes that show due diligence: how the system was designed, what the intentions were, and how they will be checked and implemented.
- To have AI deployed at scale, we need to bring together the regulator and the service provider. With such a collaboration, we can improve AI technology to better serve humans in various sectors such as the labour market, business operations, education, and transactions.

3 Main Conference Takeaways and Next Steps

3.1 Core Challenges in AI Policy and Potential Solutions

The two-day online multi-stakeholder discussion among policymakers, business, academia, and society raised crucial concerns around the current AI regulatory framework. More importantly, there were some excellent suggestions from panelists on how we can make the future with AI better, safer and more equitable for all relevant stakeholders. This section serves as a recap on the core challenges and potential remedies in AI policy, which can be grouped based on the following concepts:

1. Regulatory Approaches

Challenges
- There is currently no universally-accepted definition on what constitutes AI, which limits the precision of regulatory tools and policies.
- Despite existing regulation, there are regulatory gaps which are often sector-specific.
- Contrasting data standards across different jurisdictions pose significant challenges for data quality.
- Countries are moving with different speeds when it comes to regulating AI.

Suggestions
- It is important to work on how existing regulatory frameworks can be adapted to apply to AI systems. The next step would be to develop general legal principles around critical uses of AI, and then move to creating sector-specific regulations.
- Governments should adopt a more iterative and flexible approach to AI policy. They should leave more space for experimentation via regulatory sandboxes, and incorporate real-time monitoring of the impact of AI-based solutions. This approach ensures that regulators are up to speed with the development of AI technologies.
- The regulatory process needs to be inclusive of all stakeholders.
- Governments need to work in a coordinated manner both within and across countries.

2. Capacity Building

Challenges
- There is an urgent need for society to become aware of and understand the broader implications (risks and benefits) of technical systems such as AI.
- There is a perceived lack of technical competence among the regulators of AI.

Suggestions
- Education has a key role to play in increasing awareness and technical ability of society as a whole, which is why governments need to allocate budgets and create educational environments.
- Another way forward is to establish an ecosystem of excellence and trust that includes regulators, vibrant start-ups, and academic institutions working together on joint projects in the area of AI.
- Governments need to be at the forefront of what is going on, e.g., through scenario planning and early engagement with the business community.
- The diplomatic community and international organizations have a role to play in building capacity, sharing best practices and helping to reduce the digital divide.

3. Business Concerns

Challenges
- Technology adoption by society would be impossible without end-users' trust in technology, with transparency about performance and outcomes being of high relevance for end-users.

Companies using AI technologies thus need to demonstrate trustworthiness in their applications to be trusted by the public.
- Certain regulations limit the free flow of information within companies that operate across regions.

Suggestions
- Regulation ensures product and global industry quality by increasing trust in AI technologies by users, and creates further business opportunities for AI entrepreneurs.
- Regulators should define norms for developing, validating, and testing AI models that are risk-specific. This would give clarity to AI developers and thus speed up AI development.
- Harmonization of (clear) rules for AI will help businesses decide where to invest in the deployment of AI technologies.
- Guidance for companies on the due diligence process is needed, with clear instructions on the entire AI life cycle from development to deployment.
- SMEs should have a stronger say in the policy making process, and they should also be protected from takeovers by bigger tech companies through effective updates on competition laws.
- Regulations should be easy to comply with to ease the burden for SMEs.
- Regulators should avoid conflictive regulation, understand the consequences of specific regulations in other areas, and facilitate effective data flows by decreased data localization requirements.

4. Ethical Considerations for AI Policy

Challenges
- There are conflicts regarding the universal values that should form the basis of global policy, including cultural differences. This gives rise to different scenarios for AI-based decision-making.
- There is no mechanism of redress for citizens and consumers, which in turn impacts the adoption of AI systems.

Suggestions
- We need to identify highly critical areas where the impact of AI solutions is high, such as the use of facial recognition by the police, AI in the judiciary, and automated job hiring. Another option would be to require a licence by law for those that design and deploy AI technologies.
- It is important to ensure that the key challenges of transparency, bias, and liability regarding AI solutions are addressed.
- The performance of AI algorithms should be tested against certain benchmarks (eg. SDGs) to reduce the digital divide between and within countries.

3.2 Next Steps for RegHorizon and ETH Zürich

RegHorizon will be working in the months ahead to build on selected findings of this conference.

Special focus will be given to working with small and mid-size businesses that are the foundation of economic output in all societies, to:
• increase their awareness of upcoming regulatory requirements and identify their impact on company's business,
• build their competencies in responsibly translating regulatory requirements into a business strategy,
• bring their voice on policy needs and suggestions to relevant regulators in Switzerland, Europe, and beyond,
• put together a sector-specific or issue-specific project including relevant regulatory feedback.

Our joint platform will continue to bring all relevant voices to the table including those of policy makers, businesses, academics, ethicists, social change makers, and representatives of civil society. It will ensure that a

dialogue on AI Policy developments and societal and business needs that was started at the November 2020 conference continues, including through different formats such as debates, online events, multi-stakeholder projects, and further conferences.

These efforts will help further build on the outcomes of the first AI Policy Conference and prepare the ground for the Second AI Policy conference that will take place later in 2021.

4 AI Policy Research Workshops

4.1 Human-AI Interaction

4.1.1 *"Black Box" Medicine: Does AI raise a distinctive ethical challenge?*

Alexander Stremitzer, Kevin Tobia

From a legal perspective, all medical treatments are considered a bodily harm procedure and can only be performed after the patient has consented. Since the advent of AI-supported diagnostics has been so majorly disruptive to the field, Profs. Stremitzer and Tobia ask whether our understanding of informed consent is still applicable when AI is involved.

In a pilot experiment, the authors presented participants with one of two scenarios of a patient getting a biopsy to detect cancer: In one, the biopsy is performed directly by the consulting doctor, while in the other, the doctor uses an AI diagnostic algorithm without disclosure of this use to the patient. Participants are then asked whether they think that the patient consented to the procedure. While results showed clear differences between the two scenarios (i.e., a much lower mean consent score in the AI scenario), Stremitzer and Tobia are aware that several confounding factors might have been at work. Namely, they have identified *the analysis*, *regulatory approval, diagnostic accuracy*, and *decision opacity* as possible confounders.

For this reason, they have designed a second vignette experiment, where one scenario includes delegating the analysis to another doctor while in the other scenario the analysis is delegated to an AI. The design also varies based on whether the patient is informed of the delegation and on the prediction accuracy of the analyst (doctor or AI). Regulatory approval and opacity are kept constant across conditions.

While results clearly demonstrate that the biopsy analysis delegation should be declared to the patient to allow for informed consent, there are surprisingly no significant differences between the human doctor versus the AI conditions. Stremitzer and Tobia argue that AI diagnostics might not warrant a distinctive concept of informed consent after all. In their future work, they want to investigate how decision opacity might influence the understanding of consent.

4.1.2 *We and It: An interdisciplinary review of the experimental evidence on human-machine interaction*

Daniela Sele, Marina Chugunova

This paper is an extensive literature overview covering 118 experiments, 8 observational studies and 12 literature reviews from a wide range of disciplines such as psychology, marketing, decision-making, and medicine. The aim is to create a coherent picture of human-machine interaction through three big subsegments: general human-machine interaction, workplace interactions and incorporating automated agents in a decision-making process.

1. General H-M Interaction

One of the more interesting findings is that Human-Machine interactions are social interactions. People apply gender and ethnic stereotypes to machines, reciprocate kind behaviour and exhibit human countenance like smiling and silence fillers when interacting with chatbots. Yet, there have been studies that show that not only are these interactions different from human-to-human interactions but also depend on the level of anthropomorphism of the machine.

The fact that we respond emotionally to machines brings both positive and negative implications. It has been shown that the use of automated agents increased the likelihood of intimate partner violence disclosure. On the other side, humans display reduced emotional response which may lead to more unethical behaviour when interacting with machines.

2. Human-Machine interaction in the workplace

There have been a few conflicting results regarding productivity and perceived fairness in human-robot teams. While some studies show that reduced social pressure decreases human production, other studies show the opposite effect with an increased human-to-self task allocation. There have also been contradicting findings on humans' willingness to accept automated managers. Some papers present an increase in perceived fairness of employees when given robotic managers, while there is also counter-evidence on how humans see the machine's competence as insufficient to justify its superior position.

3. Decision-making

Author Sele identified three main phenomena when machines are in charge of decision making: *algorithm aversion, automation bias*, and *algorithm appreciation*. As a result of humans' adverse reaction to machines, humans are less-forgiving to algorithmic mistakes than human-made mistakes. In contrast to this, some studies show how people exhibit over-reliance on automated decision support and preference for algorithmic judgements over human-made decisions.

Finally, Sele mentioned that there is some inconsistency in the literature concerning triggering aversion/appreciation and attributed it to one or more of the following effects:

- *Distribution of agency*, which means that people are averse to fully delegate tasks to automatic algorithms since it reduces their input in the decision-making. However, when humans remain in charge, they are appreciative of the additional decision support;
- *Context / Type of Task*, which means that people might be more algorithm-adverse regarding a moral task such as the life of a human. On the other side, when the task is analytical, algorithms are more likely to be looked upon;
- *Performance Expectations*, which means that many people believe that a machine's performance cannot make up for humans' uniqueness.

4.1.3 *Human Bias in Algorithmic Choice*

Tobias Gesche

In the first of three experiments, the author explored whether the relative costs of false positive and false negative errors are reflected in the human designers' choice of threshold values, which determine whether an algorithm's probabilistic estimate is high enough to take action or not.

The second experiment explored whether the threshold varies with intrinsic preferences for specific decisions, and presence of conflict of interest. The main finding was that the algorithm design is highly dependent on the human designer, as preference for outcomes and conflict of interest have a strong effect on the choice of threshold values. When humans are deciding for themselves only, they tend to set much higher/lower thresholds as a means of taking back control of decision-making over the algorithm. On the other side, when the decision affects others, they seem to be confident in letting the algorithm make the choice.

The third experiment investigated whether threshold choice is adjusted to changing circumstances such as changing incentives, varying scale of consequences (represented by the number of people affected by the choice made by the algorithm), and changing statistical expertise. The results showed that conflict of interest and the initial bias have a lasting effect, even when the incentives change. However, if the initial choice was unbiased it does not debias following choices but lead designers to see it as a moral license to make a biased decision next. Importantly, the effect of such bias does not get smaller when the consequences affect more people, which implies that human biases scale. Finally, changes in statistical expertise amplify the effect that conflict of interest has.

4.2 Natural Language Processing and Blockchain

4.2.1 *Letting Text Speak to Economic Data*

Nandan Rao, Elliott Ash, Thiemo Fetzer, Carlo Schwarz

The paper presented is still a work in progress and is motivated by a more general objective to connect text data to classical economic data. More precisely, it aims at relating various text corpora to occupation titles. The occupation titles and occupational codes originate from the US Dictionary of Occupational Titles (DOT) and O*Net. The text corpora of interest are large, unlabeled data sources consisting of particular specialized language, namely online job listings, and patents. Job listings correspond to an application close to the original domain, while patent texts to a domain that is further away from occupation titles. The setting considered is thus clearly a transfer learning problem.

The work presented is methodological and follows an approach based on word embeddings. The language model used is called Starspace, which leverages embeddings at the sentence level. First, a word embedding is learned for each word in the vocabulary. Then, the subsequent summation over all word embeddings in a given sentence results in the sentence embedding. This procedure's rationale is that sentences (and thus their embeddings) should be semantically similar across the same document, and dissimilar across different documents. The word embeddings learned during training are different from classical views on word embeddings since they are meant to be composed. Then, a multinomial classifier is trained on embedded DOT descriptions (the observations) and O*Net (the labels).

The preliminary results reported are mixed in terms of linear separability of occupation titles, even in the DOT training corpus. However, the results are encouraging when applying the model on job ads and comparing the predicted occupation title with the actual job title. When using the model to patents, the classifier is very unsure about the occupation title predictions. The most critical challenge relates to these particularly low-class probabilities returned by the model. They may be due to the patents not clearly reflecting an occupation, differently from what is observed for job listings. This is to be expected given that the model is transferred to a dataset "further away" from the corpus it was trained on.

4.2.2 *Infochain: A Decentralized, Trustless & Transparent Oracle on Blockchain*

Naman Goel, Cyril van Schreven, Aris Filos-Ratsikas and Boi Faltings

Latest advances in AI and blockchain technologies allow traditional decision making systems of public interest to be implemented in an au-

tomated, decentralized and transparent manner. One of the most crucial components of these systems is data. Any data-driven system is only as reliable, decentralized and transparent as its data acquisition pipeline. For example, smart contracts on a blockchain allow transactions and agreements between real-world agents to be executed in a decentralized manner without requiring a third party. But the smart contracts rely on an external entity to get data about real-world events that trigger the execution of such transactions or agreements. This external entity is called an oracle. Existing oracle solutions rely on trusted third party data sources to acquire data about the real-world. This idea is problematic for obvious reasons. First, it requires users to trust a third party, which is against the fundamental principles of blockchain. Secondly, it doesn't provide any guarantee on the correctness of data. *Thus, to reliably exploit the benefits of decentralized applications running on blockchain, a decentralized and trustworthy solution for acquiring data must be provided.*

This paper examined the challenges of acquiring correct data without relying on trusted data providers and for the first time, delivers a fully working solution to the decentralized oracle problem. The authors showed how a peer-consistency mechanism, based on crowdsourcing and game theoretic incentives, can be used to acquire high quality data from self-interested data providers. The mechanism is robust even when the data providers have external incentives to provide incorrect data. The paper further proposed ***Infochain***, a completely decentralized peer-consistency based truthful data collection system in Ethereum. Infochain addresses various practical challenges that arise in Ethereum implementation of peer-consistency mechanisms. The paper also provided compelling empirical evidence about the economic feasibility of such a system in practice.

4.2.3 *Gender attitudes in the judiciary: Evidence from U.S. Circuit Courts*

Arianna Ornaghi, Elliott Ash, Daniel L. Chen

Although the population of law students in the US has been balanced across gender for the last two decades, women are still significantly underrepresented at the top of the legal profession, with only 26 % of all sitting judges in U.S. Circuit Courts being female in 2018. One of the questions that arise naturally is whether this gap can be attributed to the discriminatory treatment of female judges by their peers due to gender attitudes.

To study this question, Ornaghi, Ash, and Chen proposed a novel measure of gender attitudes based on use of gender stereotypes in text. More precisely, the authors represented the language used in the opinions authored by a judge using word embeddings trained with Glove. Then, they defined *gender slant* to be the cosine similarity between a gender and stereotypical dimensions, with gender dimension being the average vector difference for male and female word sets, and the stereotypical dimension being the difference between word sets for career and family. That is, they develop a measure of gender slant based on how strongly judges associate men with careers and women with families in their writing.

The authors were cautious about the data collection and preprocessing performed in this study to ensure that causal interpretation is feasible. Given that judges are quasi-randomly assigned to cases at the circuit-year level, cases assigned to judges with higher/lower slant are comparable and the effect of being assigned a slanted judge is well identified. They also conditioned on judges' biographical attributes to ensure that other characteristics do not confound the slant effect.

A few different regression models were fit to study whether gender slant affects interactions with female judges. One of the more interesting models was a differences-in-differences designed regression, which compared appealed cases decided by female or male district judges that were later assigned to judges with a different slant for a potential reversal. It was discovered that slanted judges are more likely to reverse opinions authored by female rather than male district judges. Other models' findings included that slanted judges are less likely to assign opinions to female judges and also less likely to cite female judges. Lastly, the authors validated the newly proposed gender slant measure by showing that slanted judges vote more conservatively in gender-related cases..

4.3 Regulatory Issues in AI

4.3.1 *Accuracy bounding: A regulatory path forward for the algorithmic society*

Aileen Nielsen

This paper proposed a way to tackle the issues arising from models performing too well, with the primary motivation in the growing concern about machine learning and AI. People are typically afraid that current biases will not only be perpetuated by using trainable models, but also exacerbated. Other concerns are targeted at surveillance systems that become increasingly better and cost-efficient and the potential impacts on the labour market through the use of AI.

The solution proposed by Nielsen was to "bound" the final accuracy output of a machine learning model, which can be implemented both directly or indirectly. Direct accuracy bounding refers to the limitation of a model output where extreme accuracy is undesirable, with proposed applications in tackling concerns in privacy, surveillance, and civil rights. Alternatively, accuracy bounding can be deployed to limit model accuracy indirectly by mimicking human behaviour via adding noise to the output. In the concrete example of facial recognition, the user interface could return multiple individuals, with some being the most likely predictions produced by the model and others being merely random.

The author underlined that the accuracy bounding should always be context-specific. The use of accuracy bounding could have other beneficial impacts, such as drawing attention to the inherent trade-offs when using metrics in fairness applications. In terms of social acceptance, accuracy bounding is already in line with regulations such as the EU's General Data Protection Regulation (GDPR).

Concluding remarks highlighted certain limitations of accuracy bounding. The main criticisms discussed were that accuracy is not necessarily desirable for all applications. It does not necessarily work in every conceivable implementation and constitutes only one among many possible policy instruments in solving AI-related issues.

4.3.2 *What's in the box? The legal requirement of explainability in computationally aided decision-making in public administration*

Jacob Slosser, Henrik Olsen and Thomas Hidlebrandt

Algorithmic decision making (ADM) systems are being applied in various public administrative decision-making processes. ADM has shown its advantages such as faster response time, better cost-effectiveness, and better quality. However, it raises several legal concerns, such as explainability, responsibility, and accountability.

There is a difference between legal and causal explainability. A neuron-level explanation of a decision can be compared to a human judge explaining his metabolism in detail. What is mandatory in many legal systems is a truthful explanation, addressing all essential aspects of a decision and how they were weighted. ADM systems can theoretically provide this. In Slosser et al's view, machines should not be held to higher standards than their human counterparts. The authors called this an Automated-Human Spectrum with Legal-Explainability Thresholds as a guideline for ADM explainability. To ensure that machines have the same quality of explanation as humans, an administrative Turing Test was proposed, with 80% of explanations to be provided by the algorithm, and 20% by human case-workers. Sometimes, a case is handled by both a human and an algorithm. When a human supervisor can't distinguish between explanations, the algorithm can be safely used as an ADM tool.

New rulings shall be continually fed into the system to make it more precise in its wordings and more adjusted to the corresponding legal system. This *Administrative Turing Test* uses the behavioral definition of intelligence: *If we can't distinguish by its decisions whether a judge is human or not, then its exact inner workings should not be the standard used for deciding whether AI may be used as decision support in public administration.* The main point of contention is whether ADM systems can achieve the level of explanation that human judges can provide. Legal decisions often require a high amount of socio-cultural and empathetic understanding of the case, including knowledge of the decision's legal nuances and implications. Elsewise, the systems could only be used in situations with low stakes.

Slosser et al argued that the current push for mathematical/causal explainability of Neural Nets might hurt some of their most robust features. This notion was also touched upon by Dr. Weinberger's key-note on the black-box nature of algorithms (see Section 2.3).

4.3.3 *AI Initiatives in Swiss Enterprises: Governance Mechanisms to Increase Transparency and Fairness*

Michael Weiser, Mael Schnegg, Patrick Lanter

As AI becomes widely used in corporations, it is imperative to explore the ethical risks of employing strategies from algorithms that perform tasks without human involvement and control.

Although various players (e.g., Google, Singaporean Government, European Union) have proposed mechanisms for AI governance and regulation, it is still unclear how corporations implement these proposals worldwide. Due to Switzerland's tradition of self-regulation, studying Swiss companies' approaches to AI governance is an interesting setting to understand AI governance's natural development.

Weiser et al. examined four large Swiss corporations to determine the current state of AI governance initiatives for increasing fairness and transparency in their algorithms. The authors conducted semi-structured interviews with several representatives of each company, and plan to further enrich their dataset using documentation, codes of conduct, and ethics committees' decision history.

Based on the data collected so far, the authors made several observations. Most notably, there are hardly any implemented AI systems for decision-making support in Swiss enterprises, as most corporations are still in an exploratory phase. Only one company has an

AI-specific Audit Team. Weiser et al. posit that companies do not see the need for AI governance systems yet since AI is not yet a core part of business operations and companies still expect some form of human involvement in the decision-making processes. This is reflected in the companies' treatment of AI ethics, as 1) none of the ethics training include ethics behind AI systems, 2) some companies have not yet established ethics committees, and 3) existing ethics committees do not provide support to developers on various AI issues during the development process.

Weiser et al. concluded that currently enterprises lack specific governance processes and tools to address AI's ethical challenges appropriately. However, as AI begins playing a larger role in corporations' core functions, we need to understand better if and under what circumstances existing governance mechanisms fail for AI systems.

4.4 Technical Aspects of AI

4.4.1 *A Parallel Evolutionary Multiple-try Metropolis Markov Chain Monte Carlo Algorithm for Sampling Spatial Partitions*

Wendy K. Tam Cho, Yan Liu

Current methods for determining districts for first-past-the-post voting systems present many challenges and questions regarding how partisanship can affect the fairness of the electoral maps drawn. Leveraging computers and algorithmic methods to generate maps is one approach towards creating more representative and fair boundaries. These methods, however, require a novel approach to approximate what would otherwise be an NP-Hard problem.

Cho and Liu presented an evolutionary Markov Chain Monte Carlo (EMCMC) algorithm that can leverage powerful supercomputers to parallelize and rapidly sample spatial partitions for creating distributions of electoral maps. Human collaborators can then deliberate on these results to determine a fair outcome.

One of the main motivations behind this work was to generate a representative distribution of accurate maps while adhering to the many constraints involved in drawing electoral boundaries. While the use of simple Monte Carlo (MC) or Markov Chain Monte Carlo (MCMC) simulations have been previously used, they typically failed to create a distribution that matches the true distribution of possible valid maps. EMCMC was capable of producing a result much closer to the true distribution. One of the key elements of (E)MCMC is using a Metropolis-Hastings algorithm to create and evaluate random samples, which create a representative distribution when aggregated.

Another major element in the research was related to leveraging parallel processing of the Blue Waters supercomputer. Alongside extensive map restrictions, the generation of all possible maps is an intractable problem, necessitating an approximation method such as EMCMC. To improve performance, the algorithm utilizes sending and receiving buffers to reduce down-time needed for data transfers, a methodology that was shown to scale well with increased parallelization.

This algorithm's development has applications in determining whether partisan gerrymandering has potentially occurred. It provides human map-makers with a tool to help generate and decide on an appropriate electoral map. Ultimately, it works collaboratively with humans, makes information more publicly available, thus dismantles data/expertise monopoly, increases engagement on democratic theory with greater access to information, and challenges how we conceptualize fair-

ness.

4.4.2 *POTs: Protective Optimization Technologies*

Carmela Troncoso, Bogdan Kulynych, Rebeka Overdorf, and Seda Gürses

Although the rise of AI has undoubtedly improved our everyday lives in the past decades in many ways, we have also become increasingly aware of the severe risks it poses to our privacy and social justice. Currently, the predominant approach to mitigate these risks is to implement AI governance policies, through which Machine Learning (ML) owners should be guided to make their algorithms "fair."

Troncoso et al argued that this approach, with its focus on algorithmic fairness, is insufficient to address the issue of harmful AI adequately. Instead, they outlined a new framework based on so-called Protective Optimization Technologies (POT)s, gave examples of how some existing technologies can already be seen as POTs, and elaborated their approach through two case studies.

One of the most pervasive AI systems in use today are optimization-based, developed to capture and manipulate behavior and environments to extract value. Such systems introduce broader risks and harms for users and environments (including non-users) than the outcome of a single algorithm within that system. Besides, the harmful impacts of optimization-based systems may go beyond the bias and discrimination measured in algorithmic outputs. Hence, commonly used frameworks of fairness are too narrow to capture the full range of harmful AI risks. Moreover, the authors maintained that the current notion of AI governance implicitly makes the following assumptions: 1) the system's goal is just and moral, and 2) the ML owners have both the means and the incentives to mitigate risks of harm in their systems. Both assumptions seem questionable in the real world.

The authors proposed the development of technologies aimed at mitigating the negative externalities of optimisation systems. For example, a credit-scoring algorithm that unfairly prohibits a specific group of users from receiving loans (e.g., based on their ZIP code). Here, a POT could be the systematic poisoning of the ML algorithm through manipulation of the training data by a critical mass of "Robin Hood"-type users, (i.e., users which already have a high credit score can leverage this to take out a loan under the discriminated ZIP code, then immediately pay it back, thus incrementally shifting the decision boundary in favor of the discriminated group).

Troncoso et al. acknowledged that designing and deploying POTs is not trivial and comes with its own set of challenges. For example, POTs may elicit transitions in the system state that result in other externalities or lead to an arms race between different negatively impacted populations. However, POTs are more suited to address the issue of harmful AI than current notions of algorithmic fairness and AI governance. Only POTs can provide means of intervention for affected parties from outside the system, and can serve to correct, shift, or expose any harms that systems impose on populations and their environments beyond algorithmic discrimination.

POTs could broaden our understanding of AI governance as to what it can and should be governed and who can enforce governance.

4.4.3 *Really Useful Data: A Framework to Evaluate the Quality of Differentially Private Synthetic Data*

Christian Arnold, Marcel Neunhoeffer

Synthetic data serves many purposes across research, government bodies, and businesses with data-driven decision-making processes. However, an evaluation of the effectiveness of synthetic data in representing the original data set is currently lacking. Arnold and Neunhoeffer proposed a new benchmark to evaluate differentially private synthetic data.

Synthetic data allows for data analysis while still preserving privacy when there are certain restrictions on the ethical handling of the data. Differentially private synthetic data aims to protect privacy with principled guarantees while maintaining data utility. The creation of synthetic data also allows reproducibility of sensitive research, potential increase in access to government-collected information, and businesses' outsourcing of data analytics.

Determining that synthetic data is representative or suitable is not a trivial task, with the diversity of data types being one potential issue. Ensuring data types (continuous vs. integer values) and logical consistency (no impossible data combinations) are two examples of where synthetic data could create inconsistent results. Additionally, suitability presents many challenges as many analyses are context-dependent. General utility metrics may not capture or ensure that the context-specific signals captured in the original data are preserved when synthesized.

The aim of the framework of Arnold and Neunhoeffer was to advance the quality-privacy frontier by providing a way to compare synthetic datasets systematically. One step towards better data synthesis is designing a benchmark that allows researchers to ensure their synthetic data maintains utility. Specifically, these benchmarks should provide details on training data and generalization similarity, and the specific and general utility of the synthesized data. Specifying the goals of the analyses, such as inference vs. prediction, allows for a more selective accommodation of data reproducibility.

List of Speakers & Moderators

Panel 1 — Fostering Innovation & Growth

Katarzyna Gorgol
EU Commission

Dr. Jochen Friedrich
IBM

Dr. Jovan Kurbalija
DiploFoundation

Aldo Podestà
L2F

Dr. Christian Busch
SERI, Swiss Government

Dr. Ron Chrisley
The University of Sussex, UK

Panel Moderator
Dr. Katharina E. Höne
DiploFoundation

Panel 2 — Data Privacy & Consumer Protection

Paul-Olivier Dehaye
PersonalData.IO

Dr. Jan Kleijssen
Council of Europe

Prof. Marcel Salathé
EPFL Extension School

Miguel Amaral
OECD

Leila Delarive
Empowerment Foundation

Panel Moderator
Dr. Elliott Ash
ETH Zurich

Keynote Speaker

Dr. David Weinberger
Harvard's Berkman Klein Center for Internet & Society

Panel 3 — AI Policy Geo-Harmonization

Eva A. Kaili
European Parliament

Ivana Bartoletti
Deloitte

List of Speakers & Moderators

Robert Madelin
FIPRA

David Campos Pavon
Nestlé

Panel Moderator
Ayisha Piotti
RegHorizon

Panel 4 — AI Policy in Healthcare

Dr. Ceri Thompson
DG CNECT, European Commission

Prof. Dr. Philippe Ryvlin
CHUV (Vaud University Hospital)

Gilles Lunzenfichter
Medisanté

Dr. Sean Khozin
Janssen R&D, Johnson & Johnson

Dr. Gabriel Krummenacher
Zühlke Group

Panel Moderator
Sanja Fabrio
RegHorizon

Panel 5 — Getting Ready: Managing Business Risk & Complexity

Dr. Ekkehard Ernst
ILO

Dr. Alberto-Giovanni Busetto
The Adecco Group

Dr. Joanna Bryson
Hertie School of Governance, Berlin

Luca Brunner
Cognitive Valley

Dimitrios Psarrakis
EU Parliament

Panel Moderator
Dr. Aileen Nielsen
ETH Zurich

Acknowledgement

We would like to offer our heartfelt thanks to our speakers, moderators, event participants, as well as readers. We would like to extend our gratitude also to our event partners, Logitech, American University in Switzerland, and DiploFoundation, for your support and financial contributions. We thank everybody who helped make AI Policy conference a success with a special thanks to Sabah Bassam from RegHorizon, and Teodora Bujaroska from ETH Zurich for their valuable contributions to the conference and this publication, respectively.
We thank you all for your trust and involvement as our efforts towards a sustained dialogue on the future of AI policy continue.

Ayisha Piotti
RegHorizon

ayisha.piotti@reghorizon.com

Sanja Fabrio
RegHorizon

sanja.fabrio@reghorizon.com

Dr. Elliott Ash
ETH Zurich

ashe@ethz.ch

About the Organizers

RegHorizon is a Swiss based **strategy consultancy** that helps you manage your risks and **position your organization for the future**. We help you to **build trust** with customers, employees, governments, and civil society **by developing and promoting policy solutions in emerging fields**. We help you to **optimize your stakeholder outreach** by providing actionable advice, cutting edge tools, and **executive training programs**. We **foster collaboration** among our partners (businesses and innovators, top academics, global and regional decision-makers and civil society influencers) through regular **workshops and events**.

The Center for Law & Economics is part of the Department of Humanities, Social and Political Sciences at **ETH Zurich**. We are **an interdisciplinary research center** with three professorships and numerous postdocs, as well as Ph.D. students, and scientific assistants. Our **research interests** are broad, including **intellectual property law, law and tech, law and AI, behavioral experiments, and governance**. We study these issues using tools from statistics, machine learning, and natural language processing. In these areas we are a **leading research center**, regularly hosting international conferences like the AI Policy Conference, as well as seminars and collaborations with peer institutions across the globe.

Contact Details

RegHorizon
Saint Sulpice
Vaud, Switzerland

www.reghorizon.com

ETH Zurich
Zurich
Zurich, Switzerland

www.ethz.ch