Abstracts – Conference Global Perspectives on Responsible AI 2020

For further information also see our website https://responsible-ai.org/

Foundations for Fair Algorithmic Decision Making

Krishna Gummadi, Max Planck Institute for Software Systems

Algorithmic (data-driven learning-based) decision making is increasingly being used to assist or replace human decision making in a variety of domains ranging from banking (rating user credit) and recruiting (ranking applicants) to judiciary (profiling criminals) and journalism (recommending news-stories). Recently concerns have been raised about the potential for discrimination and unfairness in such algorithmic decisions. Against this background, in this talk, I will discuss the following foundational questions about algorithmic unfairness:

(a) How do algorithms learn to make unfair decisions?
(b) How can we quantify (measure) unfairness in algorithmic decision making?
(c) How can we control (mitigate) algorithmic unfairness? i.e., how can we re-design learning mechanisms to avoid unfair decision making?

Moral Principles and Algorithms

Bernhard Nebel, University of Freiburg

Algorithms seem to lack any moral dimension because they are simply a rule for calculating a value of some mathematical function. However, one can design algorithms that compute answers to moral questions, once one has formalized moral principles. We did that in the context of AI planning systems and came up with generalizations of action-centric moral principles such as deontology, utilitarianism, do-no-harm principle, and the double effect doctrine. As it turns out, it is computationally demanding to validate plans according to most of those principles.

AI for Medicine and Healthcare

Tonio Ball, University of Freiburg

Artificial intelligence (AI) systems based on applied machine learning (ML) are increasingly used in medicine and healthcare. Such AI systems may improve and/or automatize

diagnostics, provide treatment recommendations, enable new kinds of online quality management such as by ML-based anomaly detection, steer therapy and interventions on multiple levels from scheduling to surgery robotics, and may also improve healthcare operations management, for example by prediction of bed occupancy or patient waiting times. In many of these emerging applications, interpretability of the ML models is a key requirement, but often still also a research challenge. In this presentation I will give an overview of medical AI systems currently developed at the University Medical Center, and how we approach their interpretability.

How to make sure Machines behave themselves?

Iyad Rahwan, Max Planck Institute for Human Development

Machine Intelligence plays a growing role in our lives. Today, machines recommend things to us, such as news, music and household products. They trade in our stock markets, and optimize our transportation and logistics. They are also beginning to drive us around, play with our children, diagnose our health, and run our government. How do we ensure that these machines will be trustworthy? This talk explores various psychological, social, cultural, and political factors that shape our trust in machines. It will also propose an interdisciplinary agenda for understanding and improving our human-machine ecology.

Liability for AI and other Algorithmic Systems

Christiane Wendehorst, University of Vienna

The challenges posed by AI and modern digital ecosystems in general – such as opacity ('black box-effect'), complexity, and partially 'autonomous' and unpredictable behaviour – are similar, irrespective of where and how AI is deployed. However, at a somewhat lower level of abstraction, and closer to what regulators might actually wish to address, the potential risks associated with AI appear as normally falling into either of two dimensions: (a) 'physical' risks, i.e. death, personal injury, damage to property etc. caused by unsafe products and activities involving AI; and (b) 'social' risks, i.e. discrimination, total surveillance, manipulation, exploitation etc. and general loss of control caused by inappropriate decisions made with the help of AI or otherwise inappropriate deployment of AI.

As far as 'physical' risks are concerned, I will argue in the paper that they should be subject to more traditional frameworks, including with regard to liability (e.g. product liability, vicarious liability, existing and extended regimes of strict liability). These frameworks need to be fully adapted to the challenges posed by digital ecosystems, including AI.

The 'social' dimension of risks is much more AI-specific, and much more difficult to address. This is where all the AI-specific regulatory components currently discussed as part of a new regulatory framework, such as ensuring inclusiveness of training data, ensuring that decisions are explainable, information duties, impact assessment, human oversight etc. are fully justified. The dilemma faced by a regulator is that between ensuring a sufficient level of protection across the board, without any significant gaps or loopholes, and avoiding too much

uncertainty and/or red tape. This also holds true for liability. In my paper I will describe different legislative techniques and evaluate their respective benefits and drawbacks, coming up with a suggestion for how to structure a new European legislative framework for AI.

Liability for Artificial Intelligence in Private International Law

Prof. Dr. Jan von Hein, University of Freiburg

Overview:

1. Introduction

2. The current European Framework

2.1. The subject of liability

2.2. Non-contractual obligations: the Rome II Regulation

2.2.1. Scope

2.2.2. The general rule (Article 4 Rome II)

2.2.3. The rule on product liability (Art. 5 Rome II)

2.2.4. Special rules in EU law (Article 27 Rome II)

2.3. Contractual obligations: the Rome I Regulation

2.3.1. Scope

2.3.2. Choice of law (Article 3 Rome I)

2.3.3. Objective rules (Articles 4 to 8 Rome I)

2.3.4. Special rules in EU law (Article 23 Rome I)

3. The draft regulation of the European Parliament's JURI Committee

3.1. Scope

3.2. The law applicable to high risk systems

3.3. The law applicable to other systems

4. Evaluation

5. Summary and outlook

In April 2020, the JURI Committee of the European Parliament presented a draft report with recommendations to the Commission on a civil liability regime for artificial intelligence. The draft regulation (DR) proposed therein is noteworthy from a private international law perspective as well because it introduces new conflicts rules for artificial intelligence (AI). In this regard, the proposed regulation distinguishes between a rule delineating the spatial scope of its autonomous rules on strict liability for high risk AI systems (Article 2 DR), on the one

hand, and a rule on the law applicable to fault based liability for low risk systems (Article 9DR), on the other hand. The latter rule refers to the domestic laws of the Member State in which the harm or damage occurred. In my presentation, I analyse and evaluate the proposal against the background of the already existing European regulatory framework on private international law, in particular the Rome I and II Regulations.

Technological Autonomization and its Effects on Antitrust

Stefan Thomas, University of Tübingen

Increased efficiency in decisionmaking is a main driving force behind the development and the use of artificial intelligence in businesses. The processing of great amounts of data, the ability to recognize patterns in these data in short time, and the capability of a system to adopt to new information and to pursue strategies autonomously, allow for quicker decisions and a reduction in errors. These implications can be described as phenomena of "technological autonomization". While this can increase economic efficiency and thereby consumer welfare, it also comes with a risk to competition. Autonomously acting computer-agents may achieve collusive equilibria without the affected firms actively inducing such conduct. Traditional enforcement paradigms governing the cartel prohibition can fall short of tackling these cases, if no active involvement of market participants in the control of the relevant artificially intelligent systems can be established. Artificial intelligence can also precipitate unilateral conduct harmful to competition. If artificial intelligence is relied upon by dominant gatekeepers, especially platform-operators, it can precipitate decisions that foreclose markets or reduce innovation incentives through imitation. To the extent that those gatekeepers benefit from these decisions without being actively involved in the decisionmaking process, this can pose challenges to the antitrust laws. If the decision-criteria are unknown, it is difficult to undertake a counterfactual-analysis in order to explain the causality of the decision for potential harm to consumers. Coping with these challenges calls for a reevaluation of established antitrust principles and epistemological doctrines. The traditional antitrust law's reliance on individual human conduct and normative notions might need to be gradually substituted for by effects-related criteria.

From Corporate Governance to Algorithm Governance: AI as a Challenge for Corporations and Their Executives

Jan Lieder, University of Freiburg

Every generation has its topic: The topic of our generation is digitization, especially Artificial Intelligence (AI). From a conceptual perspective, AI applications will have a major impact on corporate law in general and corporate governance in particular. In practice, AI applications pose a tremendous challenge for corporations and their executives. As algorithms have already entered the board room, law makers must consider legally recognizing e-persons as directors and managers. The applicable law must deal with effects of AI on corporate duties of boards and there liabilities. The interdependencies of AI, delegation of leadership tasks and

the business judgment rule as a safe harbor for executives are of particular importance. A further issue to be addressed is how AI will change the decision-making process in corporations as a whole. This topic is closely connected with the board's duties in Big Data and Data Governance as well as the qualifications and responsibilities of directors and managers.

A European Approach to regulating AI - the EU Commission White Paper of February 2020

Jens-Peter Schneider, University of Freiburg

Europe is a continent like others with various perspectives on AI and its regulation. These perspectives vary between EU Member States as well as between socio-economic groups within national or European arenas of debate about AI. Thus, it would be superficial to pretend to present "the" European approach to regulating AI. Instead the program wisely announces a much more modest presentation about "a" European Approach - a formula also the EU Commission used for its White Paper on AI of February 2020. The White Paper is one step in an ongoing process of shaping Europe´s digital future. As a scholar of European Administrative Law I contextualize the White Paper by presenting the evolving dynamics in the digitalized public governance of the EU Single Market. The White Paper is structured in accordance with the "twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology". The Commissions aims at the best of two worlds or in the phrasing of the White Paper at establishing an "ecosystem of AI excellence" as well as an "ecosystem of trust" in AI applications. As a lawyer I will focus on the second dimension which guides the proposals of the Commission for a regulatory framework for AI. More particularly, the Commission intends to adjust the existing EU legislative framework relating to AI on one hand side and to propose new legislation specifically on AI on the other side. In the final part of my talk I will present some first ideas about the way forward with regard to a framework for AI applications used by public bodies and administrative authorities. A major component could be risk-based impact assessments equally fostering learning about AI risks and exploitation of AI advantages. A guiding principle should be "accountability by design" complementing individual rights to privacy with structural safeguards for accountability of responsible public users of AI applications.

The Dual System of Responsible AI in Law and Ethics

Weixing Shen, University of Tsinghua

There are different understanding of the concept of AI in different countries, different industries and different stages. With various risks arising in the expansion of the application field of computer technology and the improvement of intelligence, China and other countries treat responsible AI as an essential part of AI technology development plan.

There are two goals of Responsible AI, one is to ensure the security and controllability of technology, and the other is to promote AI to provide universal benefits for human beings. For achieving these goals, it is necessary to answer three questions:

(1) Which AI development may threaten human welfare;

(2) What measures are currently taking to develop Responsible AI;

(3) How to achieve comprehensive AI risk governance through scenario theory.

Although these questions are fundamental, they are no specific answers about them. This article will answer these questions from the perspective of governance rules toolkit, which are ethical system and legal system. Specifically, building the common values of Responsible AI in the ethical system, and making the security bottom line of AI development in the legal system. Using this dual system is beneficial to provide a pleasant innovation environment for the prosperity of AI.

The goal of AI ethics system is AI for Good, and make AI safe and controllable to serve humanity. The specific approach is through ethical norms and Guidelines for Good Practice, and the best practice is to formulate international conventions and set up an international Artificial Intelligence Ethics Committee. In June 2019, China's Ministry of Science and Technology issued *New Generation AI Governance Principles - Developing Responsible AI* in June 2019. Also, Chinese enterprises put forward or participated in some initiatives in the field of AI governance. The ethical system has its more flexible mechanism and larger goals, which has an subtle influence on shaping the values of AI technology. However, this system is highly dependent on autonomy to achieve its goals. Therefore, it has the function of setting a benchmark, but there is no function of protecting the bottom line.

Although it is conservative and hysteretic in the legal governance system, its governance effect is significant. The goal of Responsible AI in the legal system is to provide different legal rules for different governance objects. Besides, the technical objects of its governance include data and algorithms, and the behaviour objects of its governance include producers, controllers and users.

The goals of data governance are privacy protection and data security. All countries have a set of systematic countermeasures, and China is not an exception. The data governance in China is through legislation and technical standards. In addition to the *Cybersecurity Law* and the current fragmented legislation, there are formulating a uniform *Personal Information Protection Law* and *Data Security Law* to deal with it.

Meanwhile, it formulated a series of national standards of *Information security technology—Personal information security specification* and *Baseline for classified protection of cybersecurity.* It is worth mentioning that the *Civil Code* of *POR*, which has just been promulgated, stipulated the right of data, privacy and personal information separately in terms of civil rights. Among them, data belongs to property rights, privacy belongs to primary civil rights, and personal information also belongs to personality interests worth protecting.

The COVID-19 crisis has also brought great challenges to the protection of personal information. On the one hand, the public welfare has been enhanced by the application of Big Data. On the other hand, we are also we have noticed the problems in protecting people's right of personal information under the COVID crisis. Currently China is trying to achieve the AI governance goal by utilizing soft guidelines designed for the protection of personal information and establishing certain important typical judicial cases. In this process, the current personal information rights system and data security management system are the core mechanisms in the promotion of law enforcement and technical standards. Meanwhile,

the traditional concept of *Informed Consent*, *Principle of Proportionality* and *Responsibility Distribution* are also very important in the framework of legal governance of responsible AI.

The governance of algorithm is another key issue in the development of responsible AI. Transparency, controllability and interpretability of algorithm are very critical not only for technologist but also for jurists. At present, the governance of algorithms in China is focused on the fields of e-commerce and Smart Products. Some legislation and judicial cases have already been formed in the fields of personalized recommendation, tort liability, etc. Actually, the development of Responsible AI should be constructed on the basis of *Explainable Algorithm*. In the future it will be necessary to establish a dual mechanism of interpretability for personal users and professional institutions, in order to meet the needs of the *right to know* and to assess risks of AI in a professional way. It will also be necessary to make a proper balance among the responsibilities of producer, controller and user of AI.

AI technology has brought unprecedented potential risks, but we need to realize that a large number of AI technologies are still in the initial stage of development, it is very important to build dynamic regulatory objectives and flexible regulatory measures. Dual system of ethics and law could accelerate the realization of Responsible AI. Today countries and organizations around the world have issued more than 40 AI-related ethical principles and standards, but there still is an urgent need to establish an international convention for Responsible AI in order to strengthen the common sense in worldwide. It will be necessary to summarize the risks of the application of AI in different public and private sectors, and accordingly, to put forward corresponding legal measures based on different risks in different scenarios.

A Work In Progress: Regulation of Artificial Intelligence in the United States as of June 2020, as seen through a human rights lens

Mathias Risse, Harvard University

I will look at the first set of initiatives to regulate AI in the Obama White House in 2016; the renewed (and modified) efforts to do so in the Trump White House in 2019 (which came in response to competition from China); as well as the ongoing efforts in the 116th Congress as of 2020. As I am a political philosopher and a human rights scholar, rather than a lawyer, I will look at these efforts through a human rights lens.

“But we don’t even have clean Water!” Some Challenges for Global Governance of AI

Mark Coeckelbergh, University of Vienna

The idea of global governance of AI is attractive but faces some important principled and practical challenges. This talk has three aims. First, it argues not only for more cooperation between national states, but also for global governance of AI by means of supranational institutions. The main argument is that AI poses global problems which, if they cannot be dealt with sufficiently at local level, needs global solutions (subsidiarity principle). Second, the talk outlines and discusses some of the challenges this raises, including the question regarding

priorities at a global and regional level (e.g. climate change and priorities of the global south), the issue regarding the possibility of a global ethics and the danger of neocolonialism or imperialism, differences in political culture, for example when it comes to freedom/authoritarianism (e.g. China versus US), and the future of supranationalism in a world dominated by nationalist ideologies and powerful nation states. Third, the talk makes suggestions for how to overcome these challenges and invites the audience to contribute to this project.

Covid-19, Contact tracing, and Data Governance – a South Korean View

Haksoo Ko, University of Seoul

Confronting Covid-19, South Korea deployed a legally mandated contact tracing mechanism. In this mechanism, the Korean Centers for Disease Control and Prevention ("KCDC") serves as a central authority for compilation of relevant data from various sources. They include location data from mobile carriers; immigration records from Immigration Services; closed-circuit television footage from the policy; records for credit, debit and prepaid card transactions from credit card companies; and transit pass records from public transit companies. After compiling the data, the KCDC engages in epidemiological research and also makes public disclosures so that the general public becomes aware of the status of virus spreading in the country. In the whole process, intriguing issues are raised regarding the flow and provenance of data as well as regarding the control of data. These issues will have broader implications on the governance aspect related to AI, in particular regarding data needed for AI.

AI and the Quest for Augmented Science Journalism

Volker Stollorz, SMC (Science Media Center Germany)

Sundar Pichai, Google's boss, has described developments in Artifical Intelligence as "more profound than fire or electricity", the main reason being that AI technologies provide "general-purpose technologies". With vast computing resources and oceans of emerging data, many actors across societies can and will adopt AI based technologies rapidly. One can distinguish two general approaches using AI: Mimesis wants to design machines that mimic and potentially replace human work. Human-machine symbiosis requires humans and machines to work more intimately together, leveraging their distinctive kinds of intelligence to transform work processes and organizations. In business speech this translates into the use of AI-enabled information, tools, and systems to empower, not replace, those who serve.

I see big unmet needs, but also major barriers for journalism to tackle breakthroughs in AI Research and developing technologies. Journalistic investigations into the crucial public issues have to be multidisciplinary where Public Interest Technologists have to work closely together with investigative journalism to be able to dive into highly complex, fast evolving and demanding research environments. I will discuss some Public Issues which make AI-related journalism crucial but in very short supply

- Research in mathematics, informatics and computer sciences develops in arcane fields of knowledge written in highly inaccessible formal languages almost impossible to decipher by nonscientific publics. Trust in research mathematic and related fields is based on absolutely no common knowledge in the public domain.

- AI remains a “black-box” kind of technology, where the core complex statistical models e.g. behind machine learning tools or Natural Language Generation Model applications cannot be easily be explained or a least communicated as a rough draft to lay audiences, mostly not even to experts in related fields of science who are using AI-based systems, e.g. in medicine

- The history of AI-Research contains a lot of hyped science communication driven by vocal academic proponents or private corporations with vested interests. The Economist recently warned of a new winter of AI, because although AI-systems can do certain tasks, such as recognizing images or speech, far more reliably than those programmed the traditional way with hand-crafted rules, they are not yet “intelligent” in the way that most people understand the term. So, not only in the journalistic space there is an established hype pipeline with lots of actors communicating tons of disinformation.

- We live in times of unbounded scientific knowledge. Non state actors or even nonmembers of the traditional academic scientific community can develop easy to use AI tools for benign or malicious purposes. As the resources required and the number of domain specific expertise needed decrease, the power to circulate disinformation or to do harm with AI-based information systems in unprepared societies increase enormously. So far, journalism is mostly unprepared for these challenges

- Advanced research on Artificial General Intelligence is nowadays taking place at huge private corporations, so independent actors in traditional academic institutions are less able to compete with the knowledge base and the computer resources required to make progress and deeply understand risks unforeseen in the field of General Artificial Intelligence Research

- Most of the Risks of AI Systems are emergent properties arising from the arbitrary complexity of information systems in general. Take the NLP-Model developed by Open AI as an example. The research institute was set up by Elon Musk to “ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity”. The Research Team delayed disclosing a complex OPEN AI Model of NLP in early 2019, only to then being shaken up by a deal with Microsoft, where the researchers now are developing a “Turing Natural Language Generation (T-NLG)” which is a “17 billion parameter language model by Microsoft that outperforms the state of the art on many downstream NLP tasks.”

- There is a long silent history of deploying AI Research in the military domain which is highly secretive and mostly inaccessible for the press and public domain scientists. The new Joint Artificial Intelligence Center at the US Department of Defense proposed recently to accelerate the adoption of AI by fostering “a culture of experimentation and calculated risk taking”. At the same time advanced AGI Research contains Ruin Risk type of collective action problems, where the risk of unintentional or intentional consequences for humanity is low with non-zero probability, but could result in unrecoverable losses if realized.

• We don´t have international standards or codes that are accepted by all states or researchers for the containment of risks or misuse in the field of artificial general intelligence. Since research is progressing rapidly, all the states and researchers particularly affected should urgently work together more closely, also internationally, in order to agree on meaningful and proportionate regulatory standards.

What does this complex situation means for journalism and an informed public? To be able to detect rare signals, investigate, report and discuss potential benefits or risks of AGI in the domain of journalism in the public interest, we urgently need many more “Public Interest Technologist” (Bruce Schneier) helping journalists figure out what happens in the research on artificial general intelligence and its potential use cases. We also need a type of augmented science journalism, which is AI savy and works more closely with AI-Scientists to better understand the power of AGI, its limitations and the risks to humanity.

What kind of AI do we need for a Good Living and a Healthy Planet?

Anna Christmann, Member of the German Parliament

AI can hold great potential for multiple sectors such as mobility, health, agriculture, logistics, construction and many more. It can help doctors identify cancer, help farmers to till their land more sustainably and with it government can provide better services to the public. At the same time, AI will fundamentally change the way some parts of society are structured and it is essential to make sure that these changes are for the better and not the worse. On a policy level, we need to make sure that undue discrimination, surveillance and profiling mechanisms will not be part of our future but that we will create spaces that foster innovation to improve our living and sustain the environment. In order to have technology that is based on European values, we need to be able to develop "AI made in Europe". For this, we need European research networks, strong AI-development ecosystems and incentives for scientists and developers to stay in Europe.

Artificial Intelligence and the Right to Data Protection

Ralf Poscher, Max Planck Institute for the Study of Crime, Security and Law

One way in which the law is often related to new technological developments is as an external restriction. Lawyers are frequently asked whether a new technology is compatible with the law. This implies an asymmetry between technology and the law. Technology appears dynamic, the law stable. We know, however, that this image of the relationship between technology and the law is skewed. The right to data protection itself is an innovative reaction of the law from the early days of mass computing and automated data processing. In my presentation, I want to explore how an essential aspect of AI-technologies, their lack of transparency, might support a different understanding of the right to data protection. From this different perspective, the right to data protection is not regarded as a fundamental right of its own, but rather as a doctrinal enhancement of each and every fundamental right against the abstract dangers that come with digital data collection and processing. This understanding

of the right to data protection shifts the perspective from the individual data processing operation to the data processing system and the abstract dangers that are connected with it. The systems would not be measured by how they can avoid or justify the processing of some personal data, but by the effectiveness of the mechanisms employed to avert the abstract dangers associated with a specific system. This shift in perspective should also allow an assessment of AI-systems despite their lack of transparency.

How to better Understand AI (in order to assess it)

Christoph Durt, University of Vienna

In my presentation, I propose a novel view on the relation between AI and humans. Traditionally and up to today, this relation is usually conceived as an external confrontation. AI is either thought to be an object, such as a tool, or directly compared to humans, as if it were an alien being. Already the Turing Test directly compares human and machine output, and early visionaries believed AI could soon simulate all of human behavior. Today, a popular question is not whether but when AI becomes more intelligent than humans, such as in speculations about singularity. Many believe there is no alternative way to conceive of AI.

I argue that this conception of AI is fundamentally mistaken, and that it leads to grave misconceptions of the chances and dangers of AI technology. I propose a radically different understanding of the relation between AI and humans. I argue that AI is neither a mere tool, nor a subject. Rather, AI is in embedded in the world of human experience and understanding, in a unique way. Constitutive for artificial "intelligence" is the interrelation of AI hardware, humans, lifeworld, and digital representations or data. I show the specificity of each of these interrelations. According to the novel view, the relation of AI and humans is not one between two given entities, but constitutive for AI. Understanding AI in this way allows a reassessment of the chances and dangers of AI.

Programming Self-driving Cars for Moral Dilemmas

Tatjana Hörnle, Max Planck Institute for the Study of Crime, Security and Law

Self-driving cars need to respond, like a human driver would, to traffic constellations that involve an unsolvable moral dilemma, for instance: continuation of the vehicle's course would kill several persons, changing the course would result in the death of one person. Should the car change its course? Such moral dilemmas are a common topic for discussions in moral philosophy (the famous "trolley problem") and criminal law. With regard to automated cars, many authors assume that the rules that will guide how the car moves could be modelled according to the rules that were developed for human beings in moral dilemmas. I will present a different view: solutions developed for human agents (such as: it is better to stay passive rather than act with deadly consequences) do not make sense for the programming of cars way before an actual accident might happen

Autonomous Weapons: A Scientist's View

Toby Walsh, UNSW Sydney

The world faces some wicked problems that threaten the progress made in the last century to make the world a better place for more of its citizens. All of us are living through the global pandemic. But there are many other problems like the climate emergency. My adopted country, Australia burnt for months on end. And then there is the increasing inequality fueling division within many societies. One problem that keeps me awake at night but that might not be on everyone's radar yet is machines killing us.

I'm an AI researcher. I've spent nearly 40 years exploring how to make computers smarter. And I'm very fearful of what the technologies I and my colleagues have been building will do in the wrong hands. Will we hand over killing to machines? Build what the diplomats prosaically call "fully autonomous weapons" but the media more evocatively call "killer robots"?

There is a growing political movement against killer robots. 30 nations have called on the UN to ban them pre-emptively. Discussions are ongoing at the UN but moving slowly. The UN Secretary General, Antonio Guterres has thrown the weight of his office behind the talks, offering this stark warning to the planet: "Let's call it as it is. The prospect of machines with the discretion and power to take human life is morally repugnant."

There is also a growing movement within civil society against such weapons. The Campaign to Stop Killer Robots, for instance, now numbers over 100 non-governmental organizations such as Human Rights. A recent IPSOS poll of 26 countries show that six out of every ten people opposed the use of autonomous weapons. But despite all this opposition to killer robots, we face a critical choice today. The technology to build autonomous weapons is about to cross out of the research lab and into the battlefield. Earlier this year, Turkey deployed autonomous kamikaze drones on its border with Syria. These drones can use face recognition software to identify, track and kill.

The world will be a much worse place if we don't stop this. I and thousands of my colleagues, other researchers in AI have warned of these dangerous developments. We've been joined by Nobel Peace Laureates, church leaders, politicians and many members of the public. (At this point, it may remind you of the debate around the climate emergency where experts and moral leaders of our society weren't listened to for years)

There is a strong moral argument against killer robots. We give up an essential part of our humanity if we hand over the decision of whether someone should live to a machine. Machines have no emotions, compassion or empathy. Machines are not fit to decide who lives and who dies. Beyond the moral arguments, there are many technical and legal reasons to be concerned about killer robots. Autonomous weapons will be perfect weapons of terror. Can you imagine how terrifying it will be to be chased by a swarm of killer drones? They will be an ideal weapon with which to suppress a civilian population. Unlike humans, they will not hesitate to commit atrocities, even genocide.

You may be surprised but not everyone is on board with the idea that the world would be a better place with a ban on killer robots. “Robots will be better at war than humans,” they say. “Let robot fight robot and keep humans out of it.” These arguments don’t stand up to scrutiny. Robots won’t be more ethical. We don’t know how to program ethics. Facebook and other tech giants have been failing at this miserably for years. And we won’t simply have robots fighting robots. Wars are fought in amongst civilian populations. Indeed, war are frequently fought against civilian populations. Robots won’t reduce civilian casualties. They’ll be causing more civilian casualties.

AI and robotics can be used for many great purposes. They will make our lives healthier, wealthier and happier. We stand at a crossroads on this issue. It must be seen as morally unacceptable for machines to decide who lives and who dies. In this way, we may be able to save ourselves and our children from this terrible future. In July 2015, I helped organise an open letter to the UN calling for action that was signed by thousands of my colleagues, other AI researchers. Sadly the concerns we raised in this letter have yet to be addressed. Indeed, they have only become more urgent. I urge others to join the global campaign to make the world a better place by banning such weapons.

Autonomous Weapon Systems: Where Have We Come From, Possible Future Pathways

Markus Wagner, University of Wollongong

Autonomous Weapon Systems (AWS) are distinct from unmanned systems, such as Unmanned Aerial Vehicles (UAVs) which have been in use by many militaries as well as non-state actors for the past two decades. Current unmanned systems are either controlled remotely by human operators or operate automatically. In contrast, AWS do not require human operators for direct operational control and therefore humans are ‘not in the loop’ when it comes to decisions by AWS to use lethal force, though current rules require that humans are ‘on the loop’.

Regardless of the weapon employed and under the current rules of International Humanitarian Law (IHL), military force must be used in a way that distinguishes between military targets and civilian persons and objects, and is proportionate to military requirements. AWS raise a number of novel challenges for IHL as software-based algorithms will be making decisions on what constitutes a legitimate military target and what level of force is proportionate. Moreover, the use of AWS also raises important questions concerning ‘command responsibility’: if an AWS uses force unlawfully, who is held accountable?

The presentation outlines potential future pathways for regulating the use of AWS, ranging from proposed but arguably unrealistic outright bans to a laissez-faire attitude, with more suggestions for regulations of various forms occupying the middle ground.

Ethical Issues in the Autonomous Weapons Debate

Alex Leveringhaus, University of Surrey

In this presentation, I examine some of the ethical issues in the debate on Autonomous Weapons Systems (AWS). Ethical perspectives are important here, for two reasons. First, within contemporary just war theory, as well as practical ethics more generally, issues relating to weapons and weapons research are largely underrepresented, which is surprising. The debate on AWS thus offers a lens through which to view the contentious subject of weapons development. Second, in the debate on AWS, in particular, ethical issues have, alongside legal issues, played a surprisingly prominent role. Many objections to AWS are ethical, rather than legal, in character. There is a sense, at least among opponents of AWS, that there is something deeply morally offensive about utilising machine autonomy in armed conflicts. In the presentation, I outline some of these arguments and provide a critical assessment of them. These are the Argument from Dignity, the Argument from Responsibility, and the Argument from Distance. In addition, I shall outline some of the conceptual difficulties in grasping AWS. Whether one rejects, on moral grounds, AWS, or whether one accepts them, largely depends on how one defines AWS in the first place. I conclude the presentation by pointing to two issues that present a fruitful way forward in the debate on AWS (as well as the nascent debates on weapons research and military technology more generally), (1) the issue of trade-offs and (2) the issue of levels of (acceptable) risk in armed conflict.

AI and National Security Law

Ebrahim Afsah, University of Vienna

The purpose and chief comparative advantage of artificial intelligence is the collection and analysis of vast amounts of information with a view to detect patterns human can't see. Particularly the ability to fuse information from different sources and databases creates powerful capabilities to interact with complex dynamic systems, including for surveillance, social control and defence. Primary advantages are speed, precision and pattern recognition, but these come at considerable risks, both practical and ethical. Chief among these are the drastic reduction of decision-making time (note the 1983 Petrov incident), the unavailability of some crucial information, interface problems and the national security imperative, all of which have the unintended consequence of escalation bias. This is will become particularly pronounced with so-called Super AI, that in conjunction with nuclear capabilities conjures fears of being 'man's last invention.' The aim of responsible defence planning must therefore be the use of AI to augment human capabilities, rather than fully autonomous systems.

A related factor is not the consequence of AI itself, but its designation as a national security asset that by virtue of its socio-economic and defence importance will become increasingly a field for heightened competition for talent and capabilities between states. This has led to calls for concerted efforts akin to the early nuclear race (Baruch Plan), recalling earlier arms control efforts.

International Legal Responsibility concerning AI in Armed Conflict: A Framework aimed at Discerning General Concepts and Specific Attributes

Dustin Lewis, Harvard University

Artificial-intelligence tools and techniques might be applied in armed conflicts to a vast — and growing — range of conduct and decision-making tasks. These technologies implicate weapons and the conduct of hostilities (including the "targeting cycle") as well as detention, humanitarian services, warships, naval mines, and the provision of legal advice. In this talk, I will discuss why attention might be placed on discerning certain elements that may be necessary to apply international legal responsibility in this area. In particular, I will raise a handful of general concepts and specific attributes that lawyers, technologists, policymakers, and others may focus on in seeking to preserve and apply legal responsibility in respect of AI-related technologies in armed conflict.

Brain Data and Consumer Neurotechnology

Marcello Ienca, ETH Zurich

Due to converging advances in neurotechnology, Artificial Intelligence and ubiquitous computing, the human mind has increasingly become integral part of the digital transformation. This socio-technical transformation raises novel and complex challenges for ethics and policy. This short talk will outline three core ethical challenges emerging out of the convergence of technological innovation and the human mind. For each of these challenges, possible solutions in terms of responsible innovation, ethics and regulation will be presented and critically discussed.

Governance of AI and Neurotechnologies

Ricardo Chavarriaga, EPFL (École polytechnique fédérale de Lausanne)

Neurotechnologies rely on multiple enabling technologies at emerging stages, notably artificial intelligence. In consequence, their impact at technical, ethical, and societal levels is still uncertain. Devising appropriate governance for neurotechnology should benefit from the current efforts focusing on its enabling technologies. Nonetheless, it is important to clearly identify what are the inherent characteristics of systems that interact with the neural system and ensure they are properly addressed. Failing to do so would lead to inadequate frameworks that either too restrictive or provide loopholes to avoid proper oversight,

A considerable effort has been devoted in the last years on identifying principles and recommendations for responsible AI. However, these principles should be complemented by consistent governance approaches at other levels including regulation and legal frameworks, as well as good-practices and technical standards. Initial activities in the field of

Neurotechnology are being led by institutions like OECD, IEEE Standards Association (IEEE SA), CTA and ISO.

Nonetheless, there is a risk that uncoordinated efforts will fail to align interest of multiple parties and yield effective outcome. Therefore, it is of outmost importance to foster a truly multi-stakeholder engagement in the development of technology and its governance mechanisms. Last that not least, discussions should be framed in terms of what is probable, plausible and possible so as to identify the suitable instruments for promoting responsible research and innovation in neurotechnologies.

Sociocultural Perspectives on NeuroRight

Karen Herrera-Ferrá, Mexican Association of Neuroethics

The constant and growing development and globalization of advanced (neuro)technology should include-the many times underestimated-cross-cultural dimension. Especially, because there are some issues and conceptualizations which are ethically unsound to assume as universal or global. For instance, relevant issues related to the brain and mind such as emotions, cognitions, behaviours, consciousness, self, free will, autonomy, personal identity, empathy, morality and decision-making, among others. These brain-mind issues also are characteristic traits related to the human persons' concept and essence, which underlie fundamental and unique culturally-shaped perceptions, interpretations and meanings; and as a result, the given value, importance and sacredness of these traits, may variate among cultures. Accordingly, alternative sociocultural perspectives are expected regarding for example, the replication of these traits in the form of artificial intelligence (AI)and its perceived dimensional impact and potential threats to the human condition. In this sense, it is important to consider specific contextual economic, legal and ethnocultural variables in order to achieve a responsible, prudent, pertinent and culturally valid transnational use of advanced technology such as AI. Therefore, and diligent with the NeuroRights Initiative, any ethical and/or legal national and international regulatory framework should aim for the protection and safety of the brain and mind within respect and inclusiveness of neurocognitive cultural diversity, human rights and fundamental freedoms.

The NeuroRights Initiative: Human Rights Guidelines for Neurotechnology and AI

Rafael Yuste, *Columbia University*

In my talk I will review the proposal that was made by the Morningside Group in 2017 to introduce five new Human Rights into the Universal Declaration of Human Rights (1). These rights ("NeuroRights) will protect mental privacy, personal identity, personal agency, equal access to cognitive augmentation and protection from algorithmic biases. I will also review our earlier proposal to follow a medical model, introducing a "Technocratic Oath" as a deontology in the neurotech and data industry and using existing societal mechanisms similar to those already implemented in the medical industry to regulate future development of

Neurotech and AI (2). Finally, I will discuss our current advocacy efforts for NeuroRights in the US and different countries, coordinated by the NeuroRights Initiative.

Yuste, R., Goering, S. and the Morningside Alliance Group (2017). Four ethical priorities for neurotechnologies and artificial intelligence. *Nature* 551, 159–163; 2017.

Goering, S. and Yuste, R. (2016). On the Necessity of Ethical Guidelines for Novel Neurotechnologies. *Cell* 167: 882-885.

Three Types of Arguments for a Global Moratorium on Synthetic Phenomenology

Thomas Metzinger, University of Mainz

There are at least three ways one might argue for a global 30-year moratorium on all research that *risks* or *directly aims* at the creation of artificial consciousness. First, under pathocentrism and negative utilitarianism: We could create conscious suffering on non-biological carrier-systems, possibly in a very large number of individuals. Second, a deontological approach: Given the right kind of phenomenal self-model, certain classes of systems could develop moral relations *to themselves* and evolve recognitional self-respect to themselves as rational entities capable of autonomous moral agency. An artificial system could assert its own dignity, and this fact could impose moral obligations on *us*. Third, rational egoism under a purely instrumental theory of rationality: We *non-morally* ought not to perform certain actions, if and only if, and because, performing that action would damage the satisfaction of our preferences. For example, given self-conscious artificial moral agents, we might slide into an uncontrollable dialectic. I will very briefly sketch the first kind of argument.

Discrimination by Algorithm - Does Data Protection provide any Answers?

Antje von Ungern-Sternberg, Trier University

Algorithms are increasingly used to assess risks, to judge people, to disseminate information, or to distribute goods. Apart from being more efficient than humans in processing huge amounts of data, algorithms – which are free of human prejudices and stereotypes – would also prevent discriminatory decisions, or so the story goes. However, many studies show that the use of algorithms can lead to discriminatory outcomes. My talk analyses different causes for these algorithm-based discriminations and outlines how two legal regimes, i.e. antidiscrimination law and data protection law, can cope with the issue of discrimination by algorithm. My central claim is that existing norms and concepts of antidiscrimination law can be used to identify illegal (or unwanted) forms of discrimination, and that data protection law can help to detect and to combat them.

Issues around AI in Medicine

Fruszina Molnar-Gabor, Heidelberg Academy of Sciences and Humanities

The presentation considers specific aspects of how the application of AI-based systems in medical contexts under international standards may be guided.

After a brief introduction to the definition, development and impact of AI in medicine, the relevant international frameworks for governance of the subject matter will be briefly sketched.

Among the frameworks presented, the World Medical Association's activity appears particularly promising as a guide for other standardization processes. The organization has already unified the application of medical expertise to a certain extent worldwide and its guidance is anchored in the rules of various legal systems forming the laws of the medical profession. This very standardized application of professional expertise might provide the basis for a certain level of conformity of acceptance and implementation of new guidelines within national rules and regulations, such as those on new technology applications within the AI field.

The next section consists of a close analysis of potential guidance for medical AI applications via a WMA declaration. In order to develop a draft declaration, I will sketch out the potential applications of AI and its effects on the doctor-patient relationship in terms of information, consent, diagnosis, treatment, aftercare and education. This will include a brief examination of what guidance is necessary when taking into account the four ethical principles of medicine.

Finally, there follows a short assessment of how further activities of the WMA in this field might affect national rules, using the example of Germany.

AI as a Challenge for Data Protection - and vice versa

Boris Paal, University of Freiburg

AI-scenarios are mainly driven and determined by data as a valuable resource. In other words, AI goes hand in hand with what may be referred to as an enormous "appetite for data". While AI is highly dependent on the access to large amounts of data (i.e. *big data*), this access is subject to substantial limits imposed by the provisions of data protection law. These restrictions mainly apply to scenarios concerning personal (instead of non-personal) data and primarily stem from the EU-General Data Protection Regulation (GDPR). One of the most important issues with regard to AI and big data is referred to as "small privacy". It describes the inherent conflict between two objectives pursued by data protection law, *i.e.* the strict protection of privacy on the one hand and the implementation of a competitive data economy on the other.

Risk Imposition by Artificial Agents: The Moral Proxy Problem

Johanna Thoma, LSE

The ambition for the design of autonomous artificial agents is that they can make decisions at least as good as, or better than those humans would make in the relevant decision context. Human agents tend to have inconsistent risk attitudes to small stakes and large stakes gambles. While expected utility theory, the theory of rational choice designers of artificial agents ideally aim to implement in the context of risk, condemns this as irrational, it does not identify which attitudes need adjusting. I argue that this creates a dilemma for regulating the programming of artificial agents that impose risks: Whether they should be programmed to be risk averse at all, and if so just how risk averse, depends on whether we take them to be moral proxies for individual users, or for those in a position to control the aggregate choices made by many artificial agents, such as the companies programming the artificial agents, or regulators representing society at large. There are problems for both options.

Against Rationale Explanations

Kate Vredenburgh, Stanford University

In this talk, I will argue against one popular socio-technical solution to the problem of opacity: rationale explanations. Such explanations are taken to achieve many important political values enabled by explanations, such as trust, recourse, respect, and accountability, as well as to enable decision-makers to comply with GDPR. For example, by providing data subjects with a single or small set of counterfactuals that link specific inputs to a desired output value, rationale explanations based on counterfactuals enable individuals to achieve recourse. However, I argue that rationale explanations are often not the best means to enable the relevant political values. The general insight is that rationale based explanations combine two different types of politically important explanations: causal explanations, that give individuals the information needed to adjust their behavior, and normative explanations, which justify the use of the system or a particular application of it.

Contact:

Prof. Wolfram Burgard (Robotics) – Dr. Philipp Kellmeyer (Neurology & Neuroethics) –

Prof. Oliver Müller (Philosophy) – Prof. Dr. Silja Vöneky (Law & Ethics of Law)

c/o

Team Saltus Responsible AI

FRIAS

Freiburg University

Germany

Emails: ai2020@frias.uni-freiburg.de

responsibleAI@jura.uni-freiburg.de