SSOAR

Open Access Repository
www.ssoar.info

Towards European Anticipatory Governance for Artificial Intelligence

Kolliarakis, Georgios (Ed.); Hermann, Isabella (Ed.)

Veröffentlichungsversion / Published Version
Sammelwerk / collection

Empfohlene Zitierung / Suggested Citation:
Kolliarakis, G., & Hermann, I. (Eds.). (2020). *Towards European Anticipatory Governance for Artificial Intelligence* (DGAP Report, 9). Berlin: Forschungsinstitut der Deutschen Gesellschaft für Auswärtige Politik e.V. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-69231-6

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY-NC-ND Lizenz (Namensnennung-Nicht-kommerziell-Keine Bearbeitung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de

Terms of use:
This document is made available under a CC BY-NC-ND Licence (Attribution-Non Comercial-NoDerivatives). For more Information see:
https://creativecommons.org/licenses/by-nc-nd/4.0

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

German Council on Foreign Relations

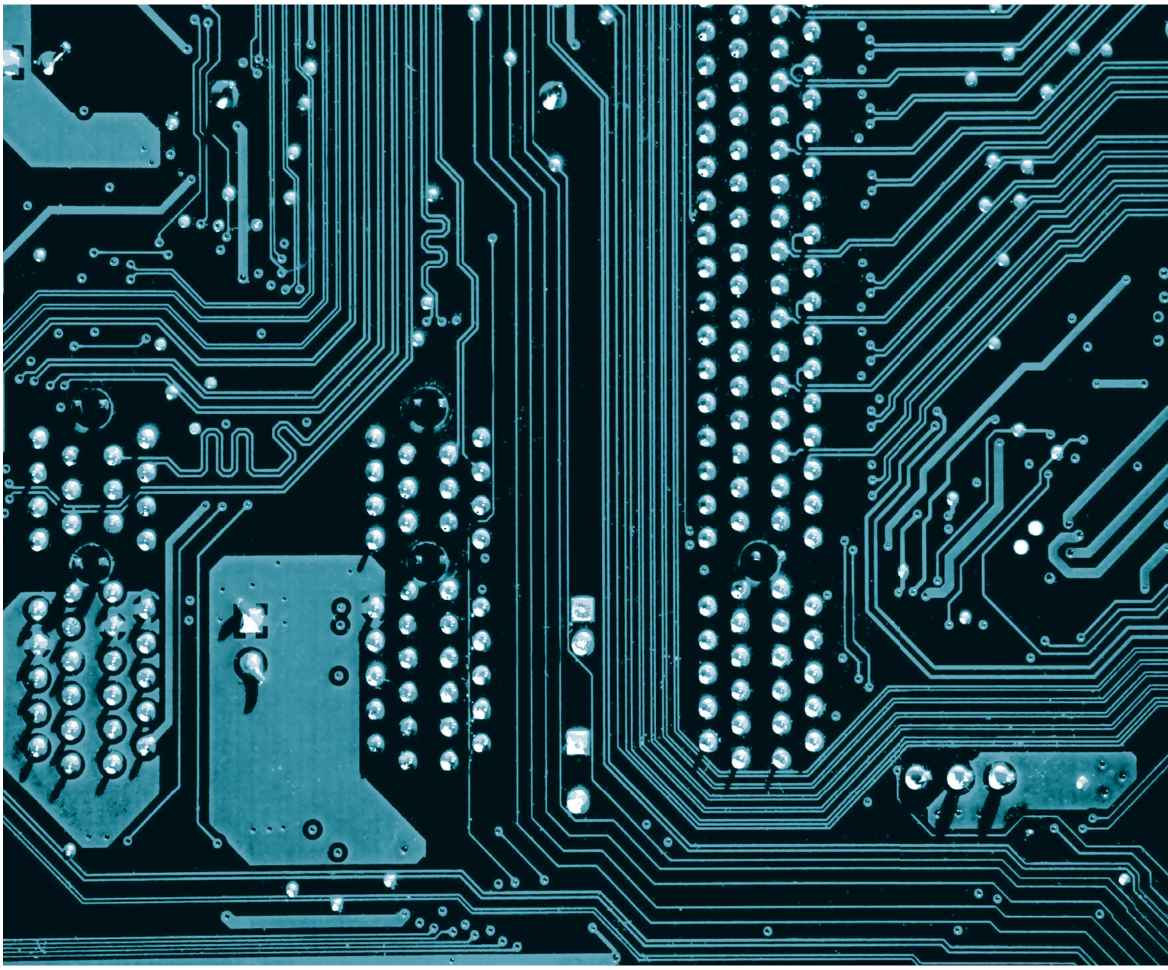
No. 9
April 2020

DGAP REPORT

Towards European Anticipatory Governance for Artificial Intelligence

Dr. Georgios Kolliarakis
Advisor for Research
Strategy, Technology,
Security, Defense, DGAP

Dr. Isabella Hermann
Coordinator of the Working
Group Artificial Intelligence
at the Berlin-Brandenburg
Academy of Sciences



The workshop "Towards European Anticipatory Governance for Artificial Intelligence" – coorganized by the Interdisciplinary Research Group "Responsibility: Machine Learning and Artificial Intelligence" of the Berlin-Brandenburg Academy of Sciences and Humanities and the Technology and Global Affairs research area of the German Council on Foreign Relations (DGAP) – was held in September 2019 in Berlin. It brought leading experts from research and academia together with policy makers and representatives of standardization authorities and technology organizations. It aimed to set framework conditions for a European anticipatory governance regime of artificial intelligence (AI) by exploring which regulatory instrument could deliver beneficial AI for society, as well as when and in which stakeholder constellation it could be implemented in order to safeguard fundamental rights, boost responsible behavior, and prevent malicious use.

Based on the fact that technology interacts with society in many ways – desirable and undesirable, predictable and unforeseen – the workshop sought to negotiate both the opportunities and limits of AI's application within a societal, interdisciplinary setting, thereby ensuring that the debate was not distracted by alarmist or euphoric narratives, but grounded in evidence. Our ambition was to demystify the mainstream AI discourse, recast the AI challenge beyond the dominant narratives, and point to a number of overlooked policy options that would reinforce and consolidate Europe's capacity to act in the future, particularly against the backdrop of geopolitical shifts currently being triggered by AI-based technologies.

Our thanks go to all participants who took the time to prepare their statements and come to Berlin. Furthermore, we would like to thank DGAP's event organization team, along with its communications department, which recorded the expert interviews. Last but not least, we wish to thank DGAP intern Karoline Jooß, without whose invaluable support the flow of the whole workshop would not have been so smooth.

**Isabella Hermann and Georgios Kolliarakis,
April 2020**

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Content

12 Key Proposals

The following 12 proposals can promote AI development in Europe and help both industry and citizens to reap its benefits.

1. Recast the challenge by building a policy framework for AI innovation

If Europe is to unlock the value of AI for its societies, we need to depart from a narrative that mystifies AI as the major disruption yet to come. Technology is not a natural phenomenon that imposes structural constraints on decision-making. Rather, it's the other way around: technology is developed and used by human beings and, thus, provides room for action. Policies, laws, and regulation often seem to lag behind innovation because technologies emerge and advance in labs and start-ups, not in bureaucracies. However, emerging AI technologies and their multiple applications are always developed and implemented within a political, organizational, and cultural context and are invariably shaped by them. The fact that AI-based technologies are embedded in societies offers a chance for early intervention in AI value chains with regulatory sticks and carrots.

2. Defend the European way of life instead of following US or Chinese paths

The "European way of life" – democracy, freedom, rule of law – is not necessarily a unique selling point for Europe in the rest of the world. That said, European legal interventions such as the General Data Protection Regulation (GDPR) are, due to public demand, increasingly influencing regulatory approaches in other areas of the world, including the US. Against this backdrop, a strong European brand of AI needs to be based on high quality standards, compliance with existing legal provisions on fundamental rights and non-discrimination, and, not least, on excellent pioneering research. It is the combination of the above that could make a positive difference for an EU label for developing and using AI that would go beyond being an uninspired copy of the US or the Chinese methods. While the EU ought to start out modest given its current position in the overheated global AI game, it should become bolder and more assertive in its level of ambition to create appropriate technologies for the European way of life and beyond.

3. Unlock the potential of ethics assessments and the EU's Responsible Research and Innovation model

Public debate is already saturated with calls for "ethical AI." Yet any claim to ethics is currently rather abstract and would need to become operationalized to really mean something. In this regard, focus should be placed not only on algorithms, but also on the data upon which AI-based technology is developed and the sociopolitical context in which it is applied. This process is only starting. Also, existing ethical guidelines have, so far, been presented or largely influenced by the industry and business sector without sufficient inclusion of experts in (applied) ethics and voices from civil society and the research sector. Broad stakeholder engagement is one of the core prerequisites for Responsible Research and Innovation (RRI). Practicing responsible research, multi-stakeholder engagement, mutual responsiveness, and reciprocal commitment is key to enabling the delivery of inclusive, accountable, and acceptable innovation, which is beneficial to many. Only if those conditions materialize in AI that is developed in Europe may we speak about an ethical and human-centric European brand of AI.

4. Foster trust in institutions and define responsibilities

Who is responsible – and thus liable – for developing AI and AI-based products that are ethical? AI leads to a particular diffusion of responsibility among human and non-human agents, as well as along processes, which makes it increasingly difficult to attribute moral and legal responsibility to certain private or public actors.

Another key question: Can a technology per se be trustworthy or not? The current discussion of this issue obscures the fact that trustworthiness of technology needs to be defined by technical norms, standards, and certificates (see point five below), which delineate a zone of acceptable performance. First and foremost, citizens place their trust in public institutions, such as authorities and governments, which can guarantee their societal welfare and the security of AI-based technologies. Secondly, they place their trust in businesses that provide innovative products and services to the market. Both public institutions and businesses are, however, comprised of people, making them inherently fallible.

5. Streamline the adoption of technical norms, standards, and certification

As ongoing efforts to create norms for autonomous vehicles increasingly show, standardization can be an effective “soft” regulatory tool, accompanying development of emerging technologies for which legislation cannot yet grasp all aspects of their potential use. Currently, there are dedicated efforts at a national, European, and international level to define common terminologies in AI; examine technical specifications in all subdomains of AI-related technologies; assess risk areas of acceptance; and integrate legal, societal, and ethical aspects into standards. A major advantage of the standard-setting process is that it is driven and controlled by the research, development, and innovation (RDI) community and therefore has feedback loops that make it adaptive to changes in technology. There is increasing support for the introduction of AI quality certification for products entering the markets as a guarantee for safety.

6. Integrate foresight, technology assessment, and democratic oversight into policy-making

As technological developments associated with AI have different degrees of maturity, and applications in the market are rapidly evolving, the impacts of their present and future applications are not fully clear. This calls for strengthening forward-looking analyses, including those of institutional, organizational, and cultural/value issues. In the context of RRI and technology assessment, efforts at the parliamentary and international level to anticipate new and converging technologies and their potential disruptive effects – both desirable and undesirable – should help to address societal and geopolitical aspects. Such activities need to inform political decision-making and democratic oversight in a systematic manner.

7. Strike a conscious balance between innovation and precautionary principles

There may often appear to be irreconcilable tension between innovation and precautionary principles. Innovation does not, however, have to be restricted by unnecessary bans. The precautionary principle – enshrined in EU treaties since 2005 – prescribes proactive caution when it comes to risks to the consumer/citizen that cannot be prevented or mitigated with available solutions. Precaution is not about bans, but rather about establishing “traffic rules,” and even imposing moratoriums if more time is needed to cope with the risks. This approach is particularly important when it comes to dual-use technologies with civil and military applications that raise serious concerns of accidental or intentional misuse by malevolent parties.

8. Boost capacity to act strategically at the national and European level

Action at the EU level is often too slow and too cautious, which can – more often than not – be attributed to the reluctance of member states to proceed jointly and decisively in one direction. The stakes involved in AI research, development, and innovation processes are high and include the welfare and protection of individual citizens, industrial competitiveness, the protection of critical infrastructure and national security, and the European capacity to act in an interconnected world. Critical mass and scaling potential can only be achieved jointly at a European level. Adopting a capability-driven approach to AI could facilitate the transformation of novelties into genuine and sustainable innovations. It will be necessary to mobilize relevant EU industries to exploit synergies, avoid unnecessary duplication, and scale-up European efforts. Furthermore, in the manner of RRI and Open Science, a comprehensive EU governance approach ought to establish a permanent dialogue platform that engages all stakeholders throughout the AI value chain. Doing so could end the current disconnect that exists between AI actors at the development end and stakeholders at the user end, as well as among regulators and authorities.

9. Disrupt AI by rethinking desirable and undesirable consequences from a policy viewpoint

AI-based technologies have triggered several debates both in expert and lay circles about multiple upcoming “disruptions.” Only cautious monitoring can reveal which of them will be mere hype and which will bring real, expected benefits, not to mention their costs and unintended effects. It is the task of policy makers to make informed decisions about desirable objectives and intervene with laws, standards-setting, or other means to achieve them. Impacts related to ecology, welfare, fundamental rights, and socio-economic equality are to be considered, in addition to arguments about technological competitiveness and economic profit. When it goes unharnessed, disruption through technology may lead to political turbulence and societal unrest.

10. Regulate AI-based technologies in a smart and sustainable way

A European brand of AI aims to sustain and enhance the "European way of life" through AI-based technologies as well as by working toward welfare, fairness, and societal resilience. Firstly, we need to look at where there is already regulation that applies to AI-based technologies and, secondly, decide what kind of new regulation makes sense. At the same time, other highly salient RDI domains, such as the medical and pharmaceutical sectors, can teach us how legal frames of self-regulation could ultimately be a possibility for enforcing codes of conduct. In order to stay competitive, it can be viable to not regulate technology per se, but to define how we want AI-based technology to be used and what kind of application we will not tolerate. Technology-neutral regulation makes it possible to contextualize further developments in established social use.

11. Invest in enablers of AI innovation, such as digital literacy

Governance measures should address and boost AI uptake and diffusion in businesses, public authorities, and research, while simultaneously enabling representatives of these sectors and citizens in general to take informed decisions and action. If citizens are not given training to improve their skills beyond a basic understanding of the logic of algorithms and the role of data, no diffusion of innovative technological solutions will take place – and also no critical oversight. At the same time, we need philosophers, ethicists, and social scientists be trained in the specifics of AI in order to realize the potential of a European brand of AI.

12. Promote European champions

This will demand joining forces at the EU level instead of pursuing separate national strategies. Moreover, in the new Multiannual Financial Framework, European governments need to "put the money where their mouths are," following the prominent role given to digital technologies by the European Commission. Instead of following a backward-looking distribution model, the R&D, digitalization, and competition dossiers need to be strengthened in view of the challenges facing Europe in the wider shifting geopolitical context. This implies that relative and not absolute performance is what counts regarding the dynamics in China, India, the US, Russia, and elsewhere. Close European coordination on policies on competition, innovation, trade, and fundamental rights is key to delivering an effective, coherent, and sustainable instrument for mobilizing the untapped potential of AI for Europe. A crucial enabler for scaling up B2B or B2C AI-supported solutions is infrastructure that allows connectivity and interoperability. Innovation should be based on purpose- and rule-driven data sharing, while safeguarding fundamental rights as inscribed in the respective European Charter.

INTERVIEW CLIPS

During the workshop "Towards European Anticipatory Governance for Artificial Intelligence," held in Berlin in September 2019, we recorded interviews with a number of participants. We have then selected and compiled some of the responses in three video clips, each of which features the reaction to a single question. Experiencing the range of reactions by experts of diverse backgrounds gives us useful insights about key challenges related to AI application and possible ways to tackle them.

Click o the icon to watch each video.

QUESTION 1

In your opinion, who should be responsible for what?

Link: https://www.youtube.com/watch?v=LcbtKK7ZtUc

QUESTION 2

What will the role of AI in society be in 2030, both in and outside of Europe?

Link: https://www.youtube.com/watch?v=JIAwFJG3G0I

QUESTION 3

How should Europe enhance its capacity to act, given the tensions between its established values and the need for technological innovation under fierce global competition?

Link: https://www.youtube.com/watch?v=efdA9hUEWYM

Georgios Kolliarakis and Isabella Hermann

Dimensions of AI Governance

The development of AI-based technology, its possible uses, and its disruptive potential for society, business, and policy are intimately interconnected. In this respect, governance of AI systems, as in most cases of emerging technologies, resembles a moving target. This poses a threefold policy challenge: first in terms of the high levels of uncertainty in assessing future AI applications as beneficial or malevolent; second, in terms of value and interest conflicts among involved actors, but also societies and states; and third, in terms of a high degree of complexity due to the involvement of several policy fields beyond technology R&D and industrial policies, such as those related to consumer protection, competition, labor, defense, and foreign affairs. A whole array of policy instruments is available including those that are self-regulatory, such as codes of conduct (CoCs); those that are "soft", such as RDI investment, standardization, and certification; and those that are "hard" and binding, such as legislation and international agreements[1].

INTRODUCTION

Hopes and concerns for the application of technologies using artificial intelligence (AI) have been fueling public debate for quite some time. The major lines of discussion run between optimistic innovation narratives about AI's benefits for society and precautions to prevent potential negative effects, both unintended and anticipated. These could include risks for fundamental rights and security, lack of provisions for responsibility and accountability, lack of law enforcement, and regulations unfit to handle the accelerating pace of research and development (R&D) as well as AI's multiple "dual-use" applications.

As in most cases of emerging technologies, certainty and consensus on how to reach desirable goals with AI are low, whereas the complexity of governance and overall stakes around it are high. The European Anticipatory Governance we propose consists of three dimensions: European, Anticipatory, and Governance. Firstly, we use the term governance because it points to the array of instruments and multitude of players involved in shaping the nexus of policy, society, and technology. Concretely, to address governance, discussions must go beyond a spectrum defined by the continuum of laws on the one hand and industrial self-constraints in the form of ethical checks on the other; they must be broadened to include additional tools that have already been established to shape the present and future of technology in, with, and for societies. Secondly, anticipation points to the fact that there is no single, deterministic future lying ahead of us, but – depending on our choices – many contingent ones. It is, therefore, necessary to assess possible, probable, and desirable effects of technology in society. In this way, we can create awareness and, ideally, shared visions in order to mobilize resources and elaborate paths to a beneficial future for society. Hence, thirdly, we should use the power of the European Union and other European countries to create a strategic, material, and moral advantage at an international level based on European values. In doing so, Europe can harvest the benefits and avoid the risks of AI applications for its people. Providing technological alternatives to deliver on the EU's promise of a common good is, thus, not merely a task for technology and industrial policy, but also civil society as a whole.

Our ambition was to showcase that, while technology creates constraints for society and policy, the opposite is also true: societal choices and policies can create constraining and/or enabling conditions for technologies in order to reach the goals that humans set. Based on this premise, we asked two main questions: What is our vision for a specifically European research, innovation, and application of AI-based technology? And what mix of legislation, standardization, certification, and self-regulatory approaches is needed to best allow AI to deliver on societal benefits while preventing undesirable side effects? We tackled these questions along four policy dimensions:

[1] We thank the four rapporteurs of the respective sessions, Sabine Ammon, Kaan Sahin, Timo Rademacher, and Jens Krause for sharing their notes.
[2] European Commission, "Ethics Guidelines for Trustworthy AI," April 8, 2019: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed March 31, 2020).
[3] European Commission, "Responsible Research & Innovation": https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation#Article (accessed March 31, 2020).

- Codes of conduct, ethics, and moratoriums in R&D, which are all components of self-regulating, non-binding constraints
- Research and innovation policies, norms, standardization, and certification
- National, European, and international legislation, treaties, and agreements
- Europe's capacity to act in AI innovation policies, focusing upon barriers and windows of opportunity

CODES OF CONDUCT AND ETHICS

Although ethics have already been identified as an indispensable component of AI R&D and application, many questions still remain with regard to the "translation" of ethical norms into algorithms. RRI could be an effective and sustainable model for building transparency, accountability, inclusiveness, and precaution into research and innovation. The role of self-imposed CoCs should be underlined and strengthened as laid out in the Ethics Guidelines for Trustworthy AI published by the European Commission's High Level Expert Group on AI (AI HLEG) in April 2019.[2]

> *Red lines should not be perceived as constraining innovation, but rather as enabling it to deliver its intended benefits*

EU treaties and charters should guide a value- and human-centric approach to AI

If there are inherent European ethics based on European heritage, the European treaties, and the European Charter of Fundamental Rights, then they should also provide the framework conditions for new technologies based on AI. A European approach to AI needs to be human-centered, meaning that it ensures good collaboration between humans and machines. Its focus must be on what is best for the human being, not on what constitutes the most efficient and optimized process. Consequently, when we discuss the regulation of AI applications, we must also consider drawing red lines. This should not be perceived as constraining innovation, but rather as enabling it to deliver its intended benefits and preventing it from causing harm to humans. Cases in point could be, for example, AI deployment on the military battleground or specific applications such as facial recognition or emotion recognition. Moratoriums in these fields are also thinkable until we have proper mechanisms to cope with risks – as decades of experience in research in medicine, genetic biology, and pharmaceuticals has shown.

Merge the ethics discussion with RRI

As a starting point, the ethics discussion on AI should be merged with other normative approaches to technological development. One such approach is the EU's Responsible Research and Innovation (RRI) initiative "that anticipates and assesses potential implications and societal expectations with regard to research and innovation, with the aim to foster the design of inclusive and sustainable research and innovation."[3] Based on the methodology of technology assessment (TA), RRI is implemented under the EU Research and Innovation Program "Horizon 2020" as a cross-cutting priority for all research actions extending over seven years (2014 to 2020). The point of departure of RRI is that – even though there are general principles that technological development must follow, e.g. human rights – they are interpreted differently in certain situations depending on the social context. Technology is always context-dependent because socio-technical chains and social contexts vary. RRI focuses on social challenges, involvement of stakeholders, risk avoidance through anticipation, inclusiveness, and responsibility.

RRI addresses many of the challenges in the current discussion around AI, namely the disconnect between the AI ethics discussion and the actual development and application of AI-based technology. The EU's Ethics Guidelines for Trustworthy AI try to overcome this discrepancy, but much could still be gained from RRI, especially as there are already tangible benchmarks and lessons learned, and the approach is very well understood and used outside of Europe. Currently, RRI is being integrated into the even broader term Open Science. A RRI/Open Science approach to AI development could pose a practical attempt to work together internationally. In addition, it could contribute to the demystification of AI by helping it to be seen as what it is: a technology enabler acting as glue between the two socio-technical trends of big data and digitalization. In this respect, an anticipatory approach – instead of retrospective interventions – is key.

From top to bottom and from left to right: Robert Gianni, Eleonore Pauwels, Kaan Sahin, Wolfram von Heynitz, Raimond Kaljulaid, Miltos Ladikas, Nico Geide, Mika Nieminen, Olaf Theiler, Isabella Hermann, Claudia Mrotzek, Wei Wei, Georgios Kolliarakis, Thomas Metzinger, Joanna Goodey, Günther Stock, Paul Lukowicz, Fruzsina Molnár-Gábor, Oliver Unger, Sabine Ammon, Katharina Sehnert, Isabelle Meslier- Renaud, Hinrich Thölken, Susanne Beck, Thomas Nunnemann, Timo Rademacher, Jens Krause

Photos: © Angela Laudenbacher

Formulate and consolidate codes of conduct

Nevertheless, if we are to implement CoCs in the spirit of RRI and Open Science, the question arises how binding they can be. Let us take a step back here and consider two crucial aspects in the discussion. On the one hand, there is an imbalance in the industry's involvement in the current drafting of numerous ethics guidelines; on the other hand, AI-based technologies lead to an atomization of human responsibility.

Firstly, industrial actors play a prominent role in defining ethical principles for the development and application of AI – be it internally, in cooperation with other companies, in political committees, or even in the academic environment. Microsoft and Facebook can be named as prominent examples. Also, industry and business have played a role in the Ethics Guidelines for Trustworthy AI themselves since industry actors are well represented in the AI HLEG. Generally, the industry counts on self-regulation, since the more binding certain rules are, the more they are regarded as inhibiting innovation and business opportunities. It is noteworthy that, in the process of defining ethical guidelines and rules, the experts in this field – philosophers and ethicists – seem to be underrepresented. If we are really convinced that Europe has an intellectual leadership role in the definition and implementation of ethical AI, we also need to train a new generation of experts in the field of (applied) ethics. Importantly, the discussion about trustworthiness of technology conceals one important element, namely that citizens build trust towards public institutions and business made up of people that guarantee the societal welfare and security of AI-based technologies – not towards the technology per se.

Secondly, AI leads to a diffusion of responsibility among human and non-human agents and along processes, which makes it increasingly difficult to attribute moral and legal responsibility to certain (personal) actors. One of the basic questions to address is who is responsible – and thus liable – that AI development and products are ethical. Is it the company developing the application, the coder and data scientist, the business or state agency offering it as a service, the person using it for assistance? Even though the problem of distributed responsibilities in technological development has always existed, we are now confronted with a new severity to the issue. Technology using AI is not only applied in single situations, but it also creates a whole techno-social system that will affect our societies in fundamental ways.

NORMS, STANDARDIZATION, AND CERTIFICATION IN RESEARCH AND INNOVATION

R&D spending, fostering innovation ecosystems, standardization, and certification of AI are "soft" yet important governance instruments. Being competitive in the global AI field while adhering with European fundamental rights and values is one challenge here. Another challenge is the interplay – and often unnecessary fragmentation – of policies at the national level versus Europeanization of efforts.

Strategically structure AI funding in R&D

Europe has great researchers at the university level and excellence centers in the AI field. To better leverage the opportunity they present, however, these research institutions and innovation hubs should be brought together in the European context to generate and leverage synergies in order to create competitive critical mass. This is especially true because the US technology companies Google, Amazon, Facebook, Apple, and Microsoft – known collectively as "GAFAM" – are transforming themselves more and more into "AI companies." To be competitive, the EU must understand that investment in physical infrastructure, such as highways, needs to be matched in the middle-term by investment in intangible assets, such as R&D, training skills, and education.

> *Assess which applications and corresponding standards can stand before writing new ones*

In order to achieve more European competitiveness and the ability to innovate, it is not useful for Europe to "fight already lost battles." It makes little sense to compete in fields or develop tools and platforms where it is almost impossible to gain ground vis-à-vis the United States and China – for example, by creating a European search engine. Instead, the European Union and its member states should find a niche in the AI industry with the aim of creating "the next big thing" with market potential and in line with European values. In that context, Europe should merge its strengths in its industrial base and further combine these advantages with AI technologies. Even though there seems to be the perception that only large countries and organizations can attain innovation breakthroughs in AI, the field also holds a great deal of potential for small players.

Move toward international norms and technical standards for AI

Norms and standards are bound to play a key role in the process of creating governance for AI. At the national and international level, a certain division of labor currently exists: political decision-makers define the specific requirements for AI, but standardization bodies define concrete technical standards. In order to fully develop AI standardization, common concepts, operationalization, and vocabulary are needed internationally. The Joint Technical Committee (JTC 1) of the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), a consensus-based and voluntary international standards group,[4] is currently developing an international standard on AI terminology. The first draft of this standard is expected to be available as early as 2020.

In this process, there must also be a balanced approach between innovation and standardization with market-oriented thinking. If standardization comes too early, it can block innovation. We should bear in mind, though, that there are different types of standards that are used at various times in the development process. Standards in terminology, for example, can be used at the outset to promote innovation by improving interoperability, while standards for the implementation of technology applications can only be introduced during the use phase.

The integration of EU values into AI systems could make a positive difference in the market, providing a good selling point for the EU worldwide. Based on the high quality of European R&D and international standardization efforts, a European branded AI could be created, opening a window of opportunity not only to catch up, but also to compete with the United States and China.

Optimize the nexus of public-private-partnerships

Collaboration between public and private entities in AI development is becoming increasingly problematic because the former do not have the same level of access to knowledge as the latter. This can lead to public organizations not always being able to judge what they are buying into. Therefore, increasing the knowledge – and, thus, indepen-

[4] Further information on the work of the Joint Technical Committee (JTC 1) can be found on its website https://jtc1info.org/ (accessed March 31, 2020).

dence – of the state and general public vis-à-vis AI-based technologies is an investment opportunity. Furthermore, the relationship between consumers and industry needs to be based upon a tangible benefit from technologies. Therefore, European R&D has to develop AI systems, which are simultaneously user-friendly and conform to standards.

Many of the AI projects currently being led by the EU have a pilot and preparatory character trying to involve multiple stakeholders. Hence, it seems that the EU is taking a cautious approach before launching large-scale initiatives. The EU still has among the biggest GDPs worldwide. It could leverage its huge market size to create multi-stakeholder innovation ecosystems that are inclusive. To increase the strategic autonomy of the EU in AI, its initiatives should be bold, investing in strong innovation hubs and engaging with public entities and civil society organizations from early on. Here, an open question is whether existing EU initiatives for developing research and innovation frameworks – such as the upcoming Horizon Europe, 2021–2027 – take sufficient account of the importance of AI for society as a whole and not merely for the business and R&D communities.

NATIONAL, EUROPEAN, AND INTERNATIONAL LEGISLATION, TREATIES, AND AGREEMENTS

Complementing self-regulatory and "soft" regulatory instruments in shaping the framework conditions for beneficial AI-based technology, is a third dimension: legislation or "hard" regulation. The principal question is how smart, adaptive regulation can achieve positive results in a broad, emerging technological domain without stifling innovation.

Balance new law-making with better implementation of existing laws

AI is bound to affect policy fields such as competition (antitrust), industry and trade (including dual-use), justice (including GDPR), consumer protection, defense, and foreign affairs. While there is a need for new pieces of legislation in some areas – for example in the rapidly evolving field of Lethal Autonomous Weapons Systems (LAWS) – we need to carefully examine when and where on a case by case basis. It might be more sensible to regulate specific applications of AI rather than regulate AI as such. Although AI-based technologies are developing to such an extent that existing regulation cannot accommodate all use cases, these cases should be a starting point for assessing which applications and corresponding standards can stand before writing new ones. This position was confirmed by research undertaken in Estonia and other northern EU countries that states that there is no need for new regulations specifically addressing AI – also to avoid the danger of overregulation. This might be especially true in order to ensure the competitiveness of smaller European start-ups or enterprises because large (US-based) companies could more easily comply with new and/or additional laws.

A key question to guide new policy initiatives is whether we are dealing with a new problem, or whether the problem can be tackled with existing regulation. In the context of the "regulatory fitness" debates of the past couple of years, we should be careful not to unnecessarily add new laws that complicate implementation of initiatives and are expensive to enforce. A contested point is whether we need specialized, tailor-made pieces of legislation for specific AI applications, or if this would lead to an explosion in regulation. The opposing view maintained that the smart way to regulate is to reduce the number of laws, and instead become more efficient in defining framework conditions that catch up with several key aspects of AI applications. The European Commission has, for example, been pursuing that approach for several years – not least on the grounds of facilitating entrepreneurial activity in a cross-border way within the EU.

There is often uncertainty on the part of the AI R&D community about the legal framework that applies to its activities, including the risk of being penalized for developing or adopting AI applications. Existing legislation was, of course, mainly written prior to current progress in AI-based technologies. Consequently, we have to check its suitability to take into account today's very different facets of AI application. Currently, only the GDPR provides some norms, which more or less specifically address AI. Yet, before resorting to the option of all-encompassing AI regulation, we need to examine whether combining the EU Charter of Fundamental Rights with corresponding case law (including such aspects as the rights of the child and freedom of expression) would be sufficient to deal with – to name just one prominent example – machine-learning-driven facial recognition technologies. Given the fact that case law is notoriously slow to offer guidance to stakeholders, especially in the field of liability, some new legislation might be warranted.

Regulate technology application, not (only) technology development

A complementary ongoing discourse concerns the question of how technology can be well regulated against rapidly advancing technological development. One option is not to regulate technology per se, but to define not only how we want AI-based technology to be used in line with our values, but also what kind of application we are not prepared to tolerate. Technology-neutral regulation makes it possible to contextualize further developments in established social use. At this point, it is important to note that there are already laws in place that provide for a certain desired social outcome, such as the European antidiscrimination law and its national implementations. However, in order to enact new laws or enforce existing ones, ethical, legal, social, and technical expertise regarding the development and application of AI by civil servants is required.

In the realm of emerging and converging technologies, we naturally cannot foresee how they will affect society. Precedence comes from the regulation of nuclear technology for civil and military use, which has led to decades of difficult international negotiations, some of which are still ongoing, about restrictions of development, test bans, and conditional definitions of use. Given the probabilistic nature of AI, procedures can be designed to test legal framework conditions for certain fields of application, instead of the technology itself. This might lead to better predictability of outcomes, and, in the medium term, provide the right insurances to citizens. In this respect, the ongoing efforts in the EU to update the dual-use export control regulation of sensitive goods, services and, crucially, intangible knowledge transfer – including AI-related chapters – is a case in point.

Establish harmonized and coherent red lines

The challenge for policy intervention is to provide framework conditions in a twofold manner. On the one hand, European governments should animate innovation and entrepreneurial activities in a joint and concerted effort (as also mentioned above). On the other hand, we need to define zones of intolerance. The Precautionary Principle has been enshrined in the EU Treaties since 2005 and should be applied to AI R&D in contexts of technology assessment. One possible avenue would be to shape future consumer protection law along the regulation of increasingly salient human-machine interactions. In this context, the issue of "distributed agency, which hampers attributability and causes diffusion of responsibility" needs to be addressed, as well as that of the diffusion of dangerous products or services.

A disquieting development to be taken into account is that companies flee from the ambit of EU law and move to China or the US to develop AI applications while subsequently selling the products in the EU market. Therefore, in order to promote trust in the regulatory competence of the EU, a stronger, more effective and more efficient EU-wide mechanism needs to be devised to eliminate duplicities and inconsistencies in consumer, competition, trade, and human rights law. More coherence and harmonization would better serve the interests of both industry and citizens.

FOSTERING THE EUROPEAN CAPACITY TO ACT IN AI THROUGH GOVERNANCE

How can governance facilitate the European capacity to act on AI in the current, shifting geopolitical and economic global order?

Strike a balance between a "European way of life" and a viable Europe in the world

When considering how to build Europe's capacity to be competitive internationally, it is important to take the wider context in which AI-related policies are developed into account. Potentially, these considerations may result in a tradeoff between competitiveness and policy-making driven by EU-values, which are based on democracy, non-discrimination, the right to data security, and transparency. It is a matter of debate whether such a tradeoff is unavoidable. Could a European brand of AI, instead, also be marketed as a uniquely innovative EU service? "Trustworthiness" in European AI – meaning that AI applications are secured by institutions and standards – was identified as one potential unique selling point. The EU could aim to be leading in this field and promote debates on responsible and accountable AI research, innovation, and application by leading international and multilateral conferences on the topic. The EU should not confine itself to working out a regulatory framework for AI, which could potentially be seen as one-sided and stifling. Nevertheless, it is important to also identify incentives for growing AI solutions in particular areas. Therefore, EU initiatives should also strongly focus on the positive outcomes of AI and the services it can provide for individual citizens and communities.

Think in ecosystem terms and engage key stakeholders along the value chain

Competitiveness on a European level can only be achieved jointly within a European ecosystem that provides the infrastructure for AI development and application. This implies integrating national digitalization and AI strategies into a single European one. It will be necessary to activate all parts of the industrial-technological base of the EU in order to exploit synergies, avoid costly duplication, and rein-force European efforts. In addition, a comprehensive EU governance approach should create a permanent dialogue platform involving all stakeholders along the entire AI value chain to improve the current separation of actors. This includes bringing together AI developers, stakeholders on the user side, regulatory authorities, businesses, and suppliers, as well as citizens and civil society actors.

Enable European infrastructures and operations to build AI champions

A key challenge for the future will be to achieve better data sharing between EU countries – a process, which is currently limited. The amount and quality of data available within the EU is a valuable resource for developing AI tools, but it is currently not fully exploited. The reason why such data sharing is not easy to implement is partly that countries take different views on how many national restrictions are necessary and/or possible without losing competitiveness in the international market. A prerequisite for data sharing, therefore, is to establish rules for sharing and interoperability of infrastructures. Further to this point, establishing co-funded and co-operated EU-wide infrastructures – from fundamental research to mission-driven applied research – is a must for enabling framework conditions that help ideas to enter the market, public administration, and society. Not least, in order to reap the benefits from AI, "tech nationalism" needs to give way to a unified European approach. This is key not only for building up critical mass, but also for being able to scale up the efforts to a bigger market. European AI champions, which can compete with US and Chinese initiatives, need pan-European infrastructures and resources to tap into.

Think Pieces[1]

[1] All Think Pieces are published here as submitted by the author(s).

Eleonore Pauwels

Converging Risks: the UN and Cyber-AI Prevention

Earlier this month, researchers created an AI-driven malware that can be used to hack hospital CT scans, generating false cancer images that deceived even the most skilled doctors. If introduced into today's hospital networks, healthy people could be treated with radiation or chemotherapy for non-existent tumors, while early-stage cancer patients could be sent home with false diagnoses. Today's medical intelligence about the treatment of cancers, blood clots, brain lesions, and viruses could be manipulated, corrupted and destroyed. This is just one example of how "data-poisoning" – when data is manipulated to deceive – poses a risk to our most critical infrastructures. Without a common understanding of how AI is converging with other technologies to create new and fast-moving threats, far more than our hospital visits may turn into a nightmare.

Policymakers need to start working with technologists to better understand the security risks emerging from AI's combination with other dual-use technologies and critical information systems. If not, they must prepare for large-scale economic and social harms inflicted by new forms of automated data-poisoning and cyberattacks. In an era of increasing AI-cyber conflicts, our multilateral governance system is needed more strongly than ever.

Data attacks are the nuclear weapon of the 21st century. Far more important than who controls territory, whoever controls data has the capacity to manipulate the hearts and minds of populations. AI-driven algorithms can corrupt data to influence beliefs, attitudes, diagnoses and decision-making, with an increasingly direct impact on our day-to-day lives. Data-poisoning is a new and extremely powerful tool for those who wish to sow deception and mistrust in our systems.

The risk is amplified by the convergence of AI with other technologies: data-poisoning may soon infect country-wide genomics databases and potentially weaponize biological research, nuclear facilities, manufacturing supply chains, financial trading strategies and political discourse. Unfortunately, most of these fields are governed in silos, without a good understanding of how new technologies might, through convergence, create system-wide risks at a global level. In a new report entitled The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI, I explore these inter-related risks, develop scenarios that illustrate how emerging technologies may play out in the coming period, and offer a way forward for the multilateral system to help prevent large-scale crises triggered by AI convergence.

CONVERGING RISKS: DATA-POISONING, DRONE SWARMS AND AUTOMATED BIO-LABS

Here is a likely scenario:

1. Data-poisoning: Similar to the falsification of hospitals' CT scans, malicious actors could use machine-learning algorithms to wage data-poisoning attacks on automated biotech supply chains. As bio-experiments are increasingly run by AI software, malware could corrupt engineering instructions, leading to the contamination of vital stocks of antibiotics, vaccines and expensive cell-therapies.

2. Genetic-engineering: Cloud labs let you control up to fifty types of bio-experiments from anywhere in the world while sitting at your computer. Hackers could rely on such automated workflow to modify the genetic makeup of the E. coli bacteria and turn it into a multi-drug resistant bio-agent.

3. Delivery: As a next step, hackers could harness off-the-shelf drones, and equip them with aerosols, to spread the multi-drug resistant bacteria within water-systems or on farms. Farmers already use drones to spray insecticides on crops.

4. False narratives: Finally, hackers could inundate social media with warning messages about contaminated antibiotics, sowing fear and confusion among afflicted populations.

Such a combination of data-poisoning, weaponization of bio-manufacturing, and manipulation of strategic information would have drastic economic costs and potentially lethal outcomes for populations. It would also have a significant impact on societal wellbeing. However, the most damaging impact would be on citizens' trust – trust in governing institutions, emergency data-systems, industrial laboratories, food supply chains, hospitals and critical infrastructures.

AI-CYBER CONFLICTS: VULNERABLE STATES AND POPULATIONS

New forms of covert data-poisoning attacks go far beyond biosafety and biosecurity. The capacity of a range of actors to influence public opinion and destabilize political, financial and critical institutions could have powerful, long-term implications for peace and security.

State or non-state actors can already generate high-quality forgeries targeted at an ethnic or religious group to foment violence and discrimination. In Myanmar, a UN report confirmed that Facebook posts had fuelled virulent hate speech directed at Rohingya Muslims. Across India in summer 2018, manipulative messages on social media sites, including Facebook and WhatsApp, painted certain groups as responsible for child abduction. The hysteria led to more than thirty deaths and left many injured. As the lines between reality and deception become blurred, there is a growing potential for large-scale mobilization of people, resources and weapons around false narratives.

The cyber- and human security implications of data manipulation are corrosive, with the landscape of hybrid threats expanding as well as the attack surface. Every country is a potential target, but especially those that have poor, vulnerable and outdated technological and cyber-infrastructures.

As vulnerable states are unable to prevent and mitigate data-poisoning attacks, they could become fertile operating grounds for cyber mercenaries, terrorist groups, and other actors, increasingly compromising the data integrity and the robustness of our globalized intelligence system.

We could face new geopolitics of inequality and insecurity, driven by the growing digital and cybersecurity divide. To meet these challenges, we need a common understanding of emerging security risks across the international community, driven by incentives for a shared approach to prevention.

THE NEED FOR STRATEGIC FORESIGHT

These dire scenarios point to the need to collectively develop a strategic foresight for the kinds of global risks posed by AI convergence.

Corporate and government leaders should conduct combined foresight progras across technological domains to anticipate and mitigate emerging threats that could harness data-manipulation and target critical infrastructures. For instance, we already know that data centers (think of medical databases or banks) and cloud environments (such as cloud bio-laboratories) are highly vulnerable to data-poisoning and other types of adversarial cyberattacks. Foresight efforts should imperatively include cooperation with states in the global south.

From data-manipulation on the safety of vaccines or gene-therapies to disinformation campaigns about the health of financial institutions, the attack surface in AI-cyber conflicts is vast and complex. Governments must collaborate with the private sector to create more efficient early warning-systems to detect and analyze the sources of data-forgeries and targeted propaganda. States will need to continuously map how these new deception tools influence public discourse and opinion. Moreover, they will need to foster cybersecurity and (bio)technological literacy among large swaths of the population.

I am convinced that there is no unilateral or bilateral solution to the kinds of pervasive threats posed by these new technologies. Our ability to understand emerging global security risks must be developed collectively or risks are becoming infected.

WHAT ROLE FOR THE MULTILATERAL SYSTEM?

Politically, legally and ethically, our societies are not adequately prepared for the deployment of AI and converging technologies. The United Nations was established many decades before this technological revolution. Is the Organization currently well placed to develop the kind of responsible governance that will channel AI's potential away from existing and emerging risks and towards our collective safety, security and well-being?

The resurgence of nationalist agendas across the world points to a dwindling capacity of the multilateral system to play a meaningful role in the global governance of AI. Major corporations may see little value in bringing multilateral approaches to bear on what they consider lucrative and proprietary technologies. Powerful Member States may prefer to crystallize their competitive advantages and rules when it comes to cybertechnologies. They may resist United Nations involvement in the global governance of AI, particularly as it relates to military applications.

Nevertheless, there are some innovative ways in which the United Nations can help build the kind of collaborative, transparent networks that may begin to treat our "trust-deficit disorder". First, the United Nations should strengthen its engagement with big technology platforms driving AI innovation and offer a forum for significant cooperation between them, along with State actors and civil society. For AI cooperation, the United Nations will need to be a bridge between the interests of nations that are tech-leaders and those that are tech-takers.

In this brokering function, an array of entities within the United Nations system could play a role that is sorely needed at the international level: 1) technological foresight, which is inclusive of diverse countries' challenges; 2) negotiating adequate normative frameworks; and 3) the development of monitoring and coordination standards and oversight.

Inclusive foresight and normative monitoring and coordination will be particularly crucial in the promotion and protection of human rights. Given the powerful, sometimes corrosive, implications that AI may have for self-determination, privacy and other individual freedoms, United Nations entities will need to collaborate to monitor and guarantee coherence across multiple normative efforts spurred by national, regional and private actors.

Finally, achieving the United Nations prevention agenda will require providing sharp and inclusive horizon scanning to anticipate the nature and scope of emerging security risks that will threaten not only nations but also individuals and vulnerable populations. Such foresight will become increasingly critical as AI converges with other technologies that are beyond State control and more accessible to a wider range of actors around the world.

Perhaps the most crucial challenge for the United Nations in this context is one of relevance, of re-establishing a sense of trust in the multilateral system. If the above analysis tells us anything, it is that AI-driven technologies are an issue for every individual and every State. Without collective, collaborative forms of governance, there is a real risk that they could undermine global stability.

Miltos Ladikas

Technology Assessment in Artificial Intelligence

It is worth highlighting again the difference between general AI and narrow AI in every debate on AI. General AI is ill-defined (e.g. creating human-like thinking systems) and, although not exactly in the realm of science-fiction, it refers to capabilities that are decades away. Narrow AI (i.e. solutions to task-specific assignments such as face-recognition), on the other hand, is happening now and results in actual social and ethical implications that need immediate attention and require urgent policies. This is the realm of Technology Assessment (TA) that comprises a family of scientific, participatory and interactive assessment processes undertaken to support the formation of public and political opinion on new technologies that may have significant impacts on society.

TA studies are undertaken in a number of AI applications that require immediate attention in assessing their implications in terms of social, ethics, economic and environmental aspects, such as:

Deep neural networks

The AI application area that creates most of the public debates that we witness globally. This refers to machine-learning algorithms that detect objects, identify people, classify images, generate text from speech, etc. that can be used in a variety of settings for a variety of purposes, ranging from surveillance to medical diagnosis.

Cybersecurity

Systems that are involved in cybersecurity generate vast amounts of data that require AI principles in identifying threats. Data input ranges from the individual (e.g. face recognition, social media and internet use, travel patterns) to software (e.g. identification of malicious codes).

Financial services

A new area of application with even more possibilities for social friction. Here, AI could be used to confirm consumer identities in relation to banking and credit services. That would entail the use of AI to issue new regulations in the form of computer code that, in turn, would establish automatic compliance.

There are countless other possibilities for AI applications that could potentially lead to sensitive public debates. These range from assessing the risk of parolee reoffending or a defendant defaulting on bail, to predicting wildlife poaching.

TECHNOLOGY ASSESSMENT OF AI

Attempting to assess the potential of AI applications and their impact on society is as new as the technology itself. Few full-scale TA projects have concentrated on AI so far and there is a general conception that policymaking in AI is inadequate and/or at an experimental stage. There is naturally an implicit difficulty in assessing AI, namely, in benchmarking its performance. There is no common comparative denominator in the performance of AI, as it could be anything between the existing performance of humans in a specific area and an ideal state of automated system functioning in the same area.

An overview of the relevant TA topics that represent AI challenges should include the following:

Data quality

Any kind of implicit or explicit bias that is inbuilt in the data that AI is based on will eventually amplify the bias. For instance, systems for predictive criminal behavior that are built upon current bias in terms of specific ethnic/religious/physical/etc. Attributes will inevitably promote the same bias in the algorithmic structure that will eventually make it harder to uncover the unfair practice.

Decision evaluation

Automated decisions based on big data (e.g. financial services) are difficult to deconstruct and justify at the individual level. Choices based on algorithms that can be challenged on appeal would be difficult to evaluate. For instance, the EU General Data Protection Regulation gives citizens a "right to explanation" on decisions based on personal data and it is unclear how this right can be upheld in big data automated decisions.

Liability issues

Machine-based decisions are prone to mistakes like any other decision-making process. It is unclear who should be held responsible for advanced AI-based errors (e.g. medical misdiagnosis). The current thinking is that liability should be held by the companies that develop the systems, but in reality, the complexity of the input in such systems does not allow for an easy liability procedure.

Monopoly of AI systems

The existing big data that most common AI systems are using belong to a few multinational internet corporations. The amount of data is a crucial factor for the accurate development of the system that the more accurate it is, the larger the amount of data it generates. This feedback loop provides an unfair advantage to the bigger players that stifle healthy competition and create market monopolies.

Control of intelligence

Although more geared towards general AI, there is still concern that any truly "intelligent" development in AI, would mean less human control over its aims. As human values differ from place to place and context to context, "machine values" could be at odds with humanity's best interests.

AI DEBATES IN THE USA

The USA has a remarkable and more advanced, in relation to Europe, AI sector. The debates on the implications of AI are not qualitatively different than those taking place in Europe. A good overview of the problematic of AI and the relevant policy options for the House of Representatives, see the main challenges as:

- Collecting and sharing the data that are needed to train AI
- accessing adequate computing resources and requisite human capital
- ensuring laws and regulations governing AI are adequate and that the use of AI does not infringe on civil liberties
- developing an ethical framework to govern the use of AI and ensuring the actions and decisions of AI systems can be adequately explained and accepted by those who interact with such systems
- As in Europe, the relevant policy challenges follow under the same headings:
- incentivizing data sharing
- improving safety and security
- updating the regulatory approach by establishing regulatory sandboxes and developing high-quality labeled data

MEANWHILE IN CHINA

At the EIT Digital conference in Brussels on September 10, 2019, an interesting assertion was discussed. This was basically the argument that Chinese AI research is far too advanced, compared to the European one, that makes any attempt to compete fruitless and wasteful. Presumably, this is the result of the superior access to vast data and little restrictions that AI research in the country is enjoying. The argument continues that Europe should concentrate on "principled" research that will create the least social friction. I could not comment on this debate at present, but I am aware that China, despite its global leadership in AI research, is not immune to the debates on social implications that are taking place in Europe and the US.

In July 2017, the State Council issued a decree on the "Development Planning for a New Generation of Artificial Intelligence", that stated, "Artificial Intelligence is an extensively disruptive technology and it may lead to problems. These problems include changes in employment structure, impact on legal and social ethics, infringement on individual privacy…". It also requested that "high attention should be paid to potential safety risks and challenges, to intensifying prospective prevention and restriction guidance, and reducing risks to the greatest extent to ensure the safe, reliable and controllable development of artificial intelligence".

As a result, in November 2017, the Ministry of Science and Technology announced the establishment of a new generation of artificial intelligence development planning promotion office, which consist of 15 central Government Ministries. The ensuing advisory committees consist of experts from universities, research institutes and enterprises, to study the governance of AI technology, and to provide policy advice to the government.

Our TA colleagues in China (CASTED) play an essential role in the new generation of AI development planning promotion office, by conducting research projects on the social impacts of AI and publishing an annual report of AI developments in China. As for the present, the results have not been translated, but anecdotal evidence points to similar challenges identified. More recently, in his annual state speech in October 2018, President Xi Jinping requested to strengthen the study, assessment and prevention of potential risks of AI.

IN CONCLUSION

AI research and developments vary greatly among the various economic powerhouses of the world, but the TA debates on social implications do not. Due to differing policy decision-making structures, policy options to deal with such issues are not similar, but the analysis of the challenges that societies are faced with could be. AI offers an excellent opportunity to attempt a global TA project that would use common methodologies in analysing societal challenges and enhance collaborative capabilities for potential common solutions. We are working towards that aim.

References

Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (2016), Technologien und Visionen der Mensch-Maschine-Entgrenzung, <http://www.tab-beim-bundestag.de/de/publikationen/index.html> (accessed March 20, 2020).

Institute of Technology Assessment of the Austrian Academy of Sciences (2017), The social effects of artificial intelligence, <https://www.oeaw.ac.at/ita/en/projects/the-social-effects-of-artificial-intelligence/overview/> (accessed March 20, 2020).

Oliver Siemoneit, "Big Data – quo vadis? Trends, Treiber, Determinanten, Wildcards," KIT Scientific Working Papers – 86 (2018) <https://publikationen.bibliothek.kit.edu/1000082069> (accessed March 20, 2020).

United States Government Accountability Office, Technology Assessment: Artificial Intelligence: Emerging Opportunities, Challenges, and Implications;Report to the Committee on Science, Space, and Technology, House of Representatives (March 2018). <https://www.gao.gov/assets/700/690910.pdf> (accessed March 20, 2020).

Miron Wolnicki and Ryszard, Piasecki, "The New Luddite Scare:The impact of artificial intelligence on labor, capital and business competition between US and China," Journal of Intercultural Management 11, no. 2 (June 2019), pp. 5-20 <https://www.degruyter.com/downloadpdf/j/joim.2019.11.issue-2/joim-2019-0007/joim-2019-0007.pdf> (accessed March 20, 2020).

World Economic Forum, The Global Risks Report 2017 (January 2017) <http://www3.weforum.org/docs/GRR17_Report_web.pdf> (accessed March 20; 2020).

Claudia Mrotzek

Governance for Artificial Intelligence

The success of modern Artificial Intelligence (AI) is yet another example of the impact of Moore's Law. Inexpensive computing power and mass quantities of data have made academic concepts dreamt up decades ago real. While high profile efforts to create self-driving cars and autonomous drones capture headlines, the real impact of AI will come from its ability to augment human decision making and productivity in all parts of the economy.

Governments around the world are noting what AI could mean for national competitiveness. In September 2017, Vladimir Putin pronounced "whoever becomes the leader in AI will become the ruler of the world." China followed shortly behind, declaring a national goal of being the global AI leader by 2030 [3]. These statements have triggered fears of an "arms race in AI" and prompted policy discussions on what Governments around the world must do to ensure its leadership in AI technologies.

The private sector will be critical to this process. Unlike Cold War "arms races", the private sector is leading the development and deployment of AI solutions, with government and military applications following behind. AI leadership depends on reinforcing market forces and minimizing the regulatory barriers to deploy AI solutions across all sectors of the economy.

DEFINING ARTIFICIAL INTELLIGENCE

Modern AI applies techniques that mimic human learning, enabling computers to solve problems that are too complex for the strict logic employed by earlier technologies. At its root, AI is about designing computers that can mimic human intelligence. The ultimate goal of AI research is often seen as producing human-like general intelligence that can flexibly perform a variety of tasks. However, narrow AI, which is tailored for specific applications, has driven practical application of AI to the modern economy. The current AI boom is a result of progress in machine learning, a subset of AI that enables computers to solve our problems from data without relying on humans to pre-program all conceivable rules.

Machine Learning: Machine learning is a set of techniques to enable computers to evaluate data, draw lessons from that data, and then make a prediction or decision using those lessons. Most machine learning leverages a variety of statistical and analytic techniques to look at data and draw conclusions in a process commonly called training. One simple explanation is that machine learning is a "new programming paradigm" that teaches computers to perform tasks with "examples, not instructions" [6]. For example, in looking for bank fraud, a simple machine learning system might be fed a large quantity of bank transaction data and analyze it to develop a model of good and bad activity. The system would then apply the learned model to evaluate and flag suspicious future transactions for bank staff.

The Future of AI: Recent progress in artificial intelligence has come about as the result of techniques that layer or combine multiple systems to solve more complex problems. This process, often called deep learning, mimics how the human brain integrates multiple systems specializing in tasks like vision, language, reasoning, and pattern recognition. Many recent, high-profile applications of AI have come about from applying deep learning – and similar techniques – to complex tasks like driving a car, holding a conversation, or recognizing a person.

Technical Challenges with AI: One of the challenges of machine learning is explainability - understanding how and why a system reached a specific outcome. In older expert systems, how decisions were made was always clear because the choices were defined in advance. In machine learning, it is not always clear why a particular model works. When faced with inputs from edge cases and outliers that were not part of the training data, the system may behave in unpredictable ways, leading to surprising outcomes, such as the Google machine learning algorithm that labeled a photo of a black woman as a gorilla [5]. Additionally, research continues to struggle with commonsense reasoning, which relates to how humans make assumptions about the essence of ordinary activities and situations.

[3] Central Committee and the State Council , "A Next Generation Artificial Intelligence Development Plan," translated by Graham Webster, et al., 2017 <https://chinacopyrightandmedia.wordpress.com/2017/07/20/a-next-generation-artificial-intelligence-development-plan/> (accessed March 20, 2020).

[5] Jessica Guynn, "Google Photos labeled black people 'gorillas'," USA TODAY, July 1, 2015 https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-asgorillas/29567465/ [Accessed 20 March 2020].

[6] Cassie Kozyrkov, "The simplest explanation of machine learning you'll ever read," Hackernoon, May 24, 2018 <https://hackernoon.com/the-simplest-explanation-of-machine-learning-youll-ever-read-bebc0700047c> (accessed March 20, 2020).

INFRASTRUCTURE OF MODERN AI

The “democratization” of the infrastructure underlying AI applications has enabled the current

AI boom. AI depends on large volumes of data to train its component systems, high-performance computers to conduct machine learning, and widespread connectivity to apply the resulting models to real problems in real (or near-real) time. Early efforts at practical AI were stymied by the high cost and limited availability of underlying infrastructure, which kept AI in corporate data centers and out of the public view. The modern era of AI is closely linked to the rise of the internet, smartphones, and cloud computing, coupled with the falling costs of compute and data storage. More consumers are now exposed to and aware of the possibilities of AI, and more developers can access the tools to put these systems into practice.

Data: The power behind modern AI is rich, plentiful, and diverse data sets, and the success depends on the ability to unlock and manage data at scale. In particular, the growing Internet of Things has been – and will continue to be – a particularly important contributor to a robust data ecosystem. Integrating cheap, connected, sensor-laden devices into every home, and business generates massive volumes of data that can be used for machine learning.

AI Applications: The public’s AI awareness is shaped by exposure to just a few products. The earliest AI applications were email management, in particular spam filtering and email categorization. Today, social media and online searches use AI to manage and customize content. Natural language processing has given rise to chatbots and digital virtual assistants. Finally, autonomous vehicles – in particular, cars and drones – have captured the public imagination by bringing science fiction into reality.

Cloud Computing: Cloud computing has become closely associated with AI for its role in accelerating AI adoption. Cloud computing has made inexpensive, scalable compute and storage widely available. Companies no longer need to invest in buying and maintaining drives to store petabytes of training data, high-performance computing racks to train AI systems, and compute capacity to run those AI´s in real-time for many users and data streams. Instead, they can rent all this capability in the cloud. Additionally, cloud computing has made basic AI applications widely available. With this, developers can plug cloud services specializing in tasks like natural language processing and image recognition into their applications with minimal effort. Cloud computing has accelerated both adoptions of AI across the software development community and the spread of AI-enabled products for consumers.

AI AND THE EUROPEAN ECONOMY

Applying AI throughout the European economy will create the next great productivity boom. The most visible applications of AI today are consumer-facing products. Frontier-pushing applications of AI, such as autonomous vehicles or robotic assembly lines, grab headlines because they raise fears that machines will replace people. However, the greatest potential for AI is the ability to improve the efficiency and accuracy of systems that collect, process, and interpret ever-growing volumes of data. In doing so, AI can free workers to focus their time on higher value-added tasks and provide information on how to make better decisions when performing those tasks.

AI for Business: Business-centric applications of AI focus on improving the business support functions performed by every private sector company, from human capital to financial management to IT. They may spot errors, reduce paperwork, or provide recommendations on appropriate best practices or solutions to problems. Because these applications are about improving the internal functioning of corporations, they carry very different risks from consumer-facing applications of AI, such as digital advertising or autonomous vehicles. Promoting AI for business is an opportunity to forge ahead in practical AI applications, addressing everyday problems encountered by millions of businesses of all sizes.

AI and Productivity: AI is emerging as a general-purpose technology that could transform the economy, like the steam-engine, electricity, and the computer. AI has applications in every part of the economy and could create the next great leap in economic productivity by assisting and augmenting human workers. Companies are looking to build a sufficient stock of knowledge, capability, and complimentary innovations so that AI can become pervasively used. They must also understand what data they can leverage and be prepared to reorganize aspects of their business to take full advantage of AI.

Augmenting vs. Replacing Human Workers: There is a great deal of concern about how AI-enabled autonomous systems will displace human workforces. Some studies suggest up to 47% of total U.S. employment may be susceptible to automation [4]. Yet in a 2017 Deloitte survey, 77% of companies suggested they planned to retrain people to use technology or redesign jobs to better leverage human skills, rather than reducing the number of jobs [7]. AI will doubtlessly disrupt certain parts of the workforce, particularly low- and unskilled-labor; however, it will also create new opportunities for people to manage and support AI-enabled systems. Job profiles characterized by repetitive tasks and activities that require low digital skills may experience the most significant decline as a share of total employment, from some 40 percent to near 30 percent by 2030, according to a McKinsey study "notes form the AI Frontier Modeling the impact of AI on the world economy [1].

Scaling AI Applications: AI-enabled systems must become a baseline capability used by all businesses, regardless of their size or whether they have staff with data science degrees. Many high-profile AI applications – such as machine vision and natural language processing – have come about because a small number of technology giants have thrown large numbers of high paid experts at very small problems. However, the future success of AI will be in scaling the technology across a large number of companies of all sizes, who have few people to throw at a huge number of problems. The countries that dominate the next century will be those that can most successfully mobilize information, integrating data-driven intelligence into every part of their economy.

AI CONSIDERATIONS AND CHALLENGES

AI adoption is not without challenges and complications. By creating non-human entities capable of making decisions that affect the health, welfare, and safety of human beings, AI is challenging many assumptions built into current legal, regulatory, and governing frameworks. Additionally, as with any new technology, frameworks that were suitable for earlier generations no longer work in the new technological environment. Several of these considerations and issues are described below.

Data Availability: As AI continues to mature, we must ensure that data remains widely accessible to companies. The greatest challenge in this area is likely to come in modernizing regulations around data privacy, ensuring the companies can easily assemble the datasets they need to use for machine learning, while also protecting the privacy of individual citizens.

Data-Derived Bias: There is currently concern about how to ensure that AI does not generate biased or unfair outputs. Machine learning often uses historical data sets to generate its models. A hiring AI may analyze the resumes of previously hired candidates to help identify job candidates for an interview. However, this historical data may be biased, producing flawed results. If a company historically discriminated against women in hiring, the hiring AI may decide men make more successful candidates and carry this bias forward. Ensuring AI models create fair models that reflect the world as it should be, not just as it is, is an essential topic of discussion.

Regulating AI: An AI economy requires tailored regulations that distinguish between consumer- and business-facing applications of AI. AI is a broad category, and the regulatory framework that is appropriate for consumer interaction with drones is different from a business employing analytics powered by AI that reduces product inventory gaps. The government has a role in addressing ethical considerations such as bias, fairness, and privacy, but these questions are most relevant in consumer-facing applications of AI. Business-facing applications of AI - which tend to be more concerned with practical questions of efficiency, best practice, and optimization – should not be subject to the same regulatory frameworks, which would unnecessarily hold back innovation. Regulators should be vigilant concerning the collection and use of large amounts of data for AI purposes to ensure that individual companies do not develop dominant positions that can be abused to prevent fair competition in the marketplace.

Privacy, Ethics, and AI: As AI begins to make more decisions that affect the general public, new questions will be raised about liability and decision making. If an autonomous vehicle has to choose between hitting a pedestrian or driving off a bridge, how does it make that choice? And who is liable for the resulting harm? Legal and regulatory frameworks for addressing these questions do not exist. Similarly, how does machine learning affect ownership over individual data? For example, healthcare AI requires assembling and moving large amounts of private patient data, but current regulations make this difficult. How should countries balance individual protections against the benefits of innovation and competition?

[1] Jacques Bughin, et al., "Notes from the AI Frontier: Modeling the Impact of AI on the World Economy," McKingsey Global Institute, Discussion Paper (September 2018) <https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20frontier%20Modeling%20the%20impact%20of%20AI%20on%20the%20world%20economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-economy-September-2018.ashx> (accessed March 20, 2020).
[4] Carl Benedikt Frey and Michael A. Osborne, "The Future of Employment: How Susceptible Are Jobs to Computerisation?" Oxford Martin Programme on Technology and Employment, Working Paper (September 2013) <https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-employment/> (accessed March 20, 2020).
[7] Jeff Schwartz, et al., "The Future of Work: The Augmented Workforce: 2017 Global Human Capital Trends," Deloitte (February 2017) <https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2017/future-workforce-changing-nature-of-work.html?icid=dcom_promo_featured%7Cae;en> (accessed March 20, 2020).

Raimond Kaljulaid

The Dangers and Benefits of Legislation on Artificial Intelligence

The legal and ethical issues which arise with the implementation of artificial intelligence require reflection in EU Member States and negotiations between countries. Concurrently the topic has been the focus not only of the German Council workshop but also of a high-level digital summit on artificial intelligence in Tallinn (Tallinn Digital Summit 2019).

I left the Berlin discussion thinking mainly about how to help our European partners better understand that the implementation of artificial intelligence is not only a question of ethics but may prove to be a crucial issue for Europe's commercial competitiveness and our people's economic welfare. Our aim is that when speaking of regulations, the European Union not only thinks about mitigating risks but also removing unnecessary obstacles to the development and implementation of artificial intelligence. The legal and ethical issues which arise with the implementation of artificial intelligence are important and certainly require reflection in the Member States and negotiations between countries. However, it is not constructive to worry solely about risks and it is not possible to fully mitigate all risks.

Last year, a high-level working group was formed in Estonia to examine the issue of artificial intelligence more seriously. Its report is a public document accessible to everybody. In short, Estonia considers the adoption of artificial intelligence important, especially in terms of the e-services provided by the state, where we can rely on existing solutions. A few years ago, there was only a handful of pilot projects in the public sector, but today there are more than 20. The working group analyzed whether a separate law is needed for the implementation of artificial intelligence and concluded that such a law is not necessary.

Yes, specific issues need to be specified in legislation (e.g. who is responsible if something goes wrong), but there is no need for the fundamental restructuring of our legal system. The Nordic countries and the Baltic States have also jointly called for excessive regulation to be avoided. So far, at the European level, general ethical guidelines of an indicative nature have been published. The worst thing for Europe's competitiveness would without doubt be for the Member States to create their national laws in the area of artificial intelligence.

The need for regulations is partially objective – this is a potentially ground-breaking form of technology. The fact that politicians want to feel important and have a say on a key issue also plays a part. The regulations seek to avoid possible negative effects and ensure that the implementation of artificial intelligence is in compliance with European values. For example, it would not be in compliance with European values if artificial intelligence mediating employers and job seekers "learned" that men should generally be recommended for leading positions. This would constitute discrimination.

Ethical choices are particularly difficult when we talk about implementing artificial intelligence in the military sector. Could we, at some point, give an algorithm the right to take a human life?

The problem is a multifaceted one. Many experts believe that the introduction of artificial intelligence in weaponry represents a technological breakthrough with the same fundamental meaning as the addition of atomic weapons to national arsenals. This in turn means that the risks are enormous. For instance, if it can be assumed that a country or military alliance that is ahead of others in the implementation of artificial intelligence thus achieved decisive strategic advantages against its potential opponents, then might a country that was lagging behind decide to pre-emptively attack its potential enemy?

In the view of non-governmental organizations fighting for human rights, the taking of human life by a machine cannot, under any circumstances, be ethical or compatible with international law. Germany has already adopted the stance that autonomous weapons should be pre-emptively banned. However, ethical considerations and realpolitik are often contradictory. Can the West afford technology that significantly changes the balance of power to be in the arsenal of its potential opponents but not in its own?

It can also be predicted that Europe and America will have very different expectations in their approach to the military implementation of artificial intelligence. All these questions are very relevant. Yet, it is equally worth talking about the possibilities that the implementation of artificial intelligence could provide.

JUST A FEW EXAMPLES

Innovative solutions in healthcare that enable us to live healthily for most of our lives and to cope with the increase in health expenditure due to the aging population, transferring routine tasks to robots, solutions that considerably increase traffic safety, the use of artificial intelligence-based applications in social welfare, as well as establishing even more bureaucracy-free states.

Efforts at the European level to make data available to Member States accessible and usable could open up opportunities for a number of European companies to grow faster on the domestic market and from there to enter into global competition from a stronger starting point.

Many believe that the implementation of artificial intelligence could be part of the solution to the climate crisis, enabling greater energy efficiency and less wasteful transport solutions (self-driving cars and smart public transport).

Firstly, I fully support Estonia's continued participation in the international debate on artificial intelligence, an example of which was the digital summit held in Tallinn.

Secondly, we need to remind our European partners that analysis and regulation of ethical aspects are important, but that we simultaneously need to look at how we can eliminate obstacles to implementing artificial intelligence where it could be beneficial. In addition to the abstract challenges of the future, we also need to deal with current limitations that prevent both the development of the field of technology and, for example, the free movement of data and services. It will also help to alleviate Eurosceptic sentiment if Europe is able to accomplish things, rather than just talk about them. Experience has shown that common objectives and actions unite people – the gaps and contradictions that have also arisen in Europe could be reduced if we collectively take pride in our continued success in the digital field.

And thirdly, concerning the issue of using artificial intelligence for military purposes, let us take into consideration that a certain gap may develop over time between the interests of Europe and those of our American partners.

Paul Lukowicz

The Challenge of Human-Centric AI [1]

The notion of "Human Centric AI" increasingly dominates public AI debate in Europe [2]. It postulates a "European brand" of AI beneficial to humans on both individual and social level that is characterized by a focus on supporting and empowering humans as well as incorporating "by design" adherence to appropriate ethical standards and values such as privacy protection, autonomy (human in control), and non-discrimination. Stated this way (which is how it mostly appears in the political debate), it may seem more like a broad, vague wish list then a tangible scientific/technological concept. Yet, on a second glance, it turns out that it is closely connected to some of the most fundamental challenges of AI [3].

First of all, consider the notion of "supporting and empowering" humans in the context of privacy protection and autonomy. Today a broad range of AI assistants exists from smart loudspeakers through recommender systems, intelligent search engines and personalized news services to professional assistance systems, e.g. for industrial applications. All of them struggle with privacy concerns as most of the underlying Machine Learning (ML) techniques critically depend on having as much training data as possible. Thus, the ability to learn from as little data as possible, just as humans do, which is a core fundamental research question of ML [4], is also an essential component of the vision of human Centric AI. Related is the problem of distributed, collaborative learning that does not require a centralized collection of large amounts of possibly sensitive data.

Beyond privacy, today's AI assistants also run into fundamental limits with respect to the notions of "empowerment" and autonomy. Empowerment and autonomy imply that a system should help users make more informed decisions, pursue their own agendas more efficiently and build up their own differentiated opinions. In other words, systems should be able to truly, constructively elaborate and explore issues with human users. By contrast, today's AI systems mostly provide the users with a limited set of or recommendations to choose from, collect and filter information, or try to prevent users from doing what the system considers as mistakes. In most cases, the system reaches its decision in a "black box" like manner which the user has no way of understanding or arguing with. Thus, in a way, today's systems are largely prescriptive and autonomy constraining rather than empowering and truly supportive. To change these AI systems must gain the ability to develop a differentiated understanding of human lines of reasoning, relate to human motivations, emotions, moral assumptions and implications in this reasoning, help human partners challenge their assumptions as well as provide simulations with consequences and explain alternate "AI angle" on seeing the problem. They must be able to make their reasoning transparent to the user and anchor it within complex differentiated world models grounded in both the physical reality and the user's subjective perception of reality.

Transparency and explainability of ML systems, together with the ability to reason within complex, differentiated world models, are also core concerns when it comes to the adherence to ethical standards, fairness and non-discrimination. Thus, AI systems increasingly support or even make decisions that have grave personal and/or social consequences. Examples are judges, doctors, policymakers or managers who more and more rely on AI decision support or even decision-making systems. The ability to challenge such decisions when they have an impact on a person's life is a fundamental ethical concern that can not be satisfied when the decisions are influenced or even directly made by a "black box" like AI systems. Instead, AI systems must be able to translate its computation into an explanation that is accessible to a non-expert. Such translation between the complex Ai model and related computation and simple non-expert mental model anchored within a user's subject world view goes far beyond the current state of the art in explainable ML [5].

Fairness and non-discrimination are further issues where AI systems need to relate their computation models to complex world models and the way humans perceive and judge real-world situations. The problem is that, in most cases, discrimination and bias do not arise as a result of any objective errors in the respective AI algorithms. Instead, the systems do what they were designed to – build models based

[1] Kind permission to reprint from Digitale Welt, Vol 4/2019, under https://link.springer.com/article/10.1007/s42354-019-0200-0

[2] European Commission, "Shaping Europe's digital future: Ethics guidelines for trustworthy AI," April 8, 2018 < https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> (accessed March 20, 2020).

[3] Andrzej Nowak, et al., "Assessing Artificial Intelligence for Humanity: Will AI be the Our Biggest Ever Advance? or the Biggest Threat [Opinion]," IEEE Technology and Society Magazine 37, no. 4 (December 2018), pp. 26-34 <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8558761> (accessed March 20, 2020).

[4] Jake Snell, et al., "Prototypical networks for few-shot learning," arXiv:1703.05175 (March 2017).

[5] Wojciech Samek, et al., "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," arXiv preprint arXiv:1708.08296 (August 2017).

on patterns contained in the training data. Unfortunately, training data often reflect social bias and unfairness of the real world, which the systems incorporate and often amplifies. To avoid this, the systems must be able to go beyond mere statistical analysis and numerical optimization and relate the data to world models that reflect human ethical and moral values [6]. In a way, it requires that the systems do not discover the actual statistical properties of the data but adjust what they discover according to what is desirable from an ethical/social point of view. Give the vagueness and fluidity of such an adjustment (we do not want the system to fully ignore the data and just produce outcomes that we like); this will, in general, not always be a fully automated process. Instead, humans must be able to accompany and guide the learning process and specify high level boundary conditions and optimization goals. This goes beyond explainable ML towards the notion of interactive ML [7,8], where the learning process is a "co-creation" between the user and the AI.

In summary, the notion of Human Centric AI should not be seen as a potential regulatory roadblock to AI research but rather as a challenge involving basic open AI problems such as:

1. Comprehensive world models that, in their scope and level of sophistication, should strive for human-like world understanding

2. "Interactive AI" that allows humans to not just under stand and follow the learning and reasoning process, but also to seamlessly interact with it and guide it.

3. Understanding and naturally interacting with humans and complex social settings within dynamic open-world environments.,

4. Reflexivity and expectation management

The above research challenge is at the core of the Humane AI initiative (https://www.humane-ai.eu/) which combines nearly fifty reknown European AI labs in the effort to make the vision of Human Centric AI a reality.

[6] James H. Moor, "The nature, importance, and difficulty of machine ethics," IEEE intelligent systems 21, no. 4 (August 2006), pp. 18-21.
[7] Ashraf Abdul, et al., "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda." Proceedings of the 2018 CHI conference on human factors in computing systems 582 (April 2018), pp. 1-18 <https://doi.org/10.1145/3173574.3174156> (accessed March 20, 2020).
[8] Saleema Amershi, et al., "Power to the people: The role of humans in interactive machine learning," AI Magazine 35, no. 4 (December 2014), pp. 105-120 <https://www.aaai.org/ojs/index.php/aimagazine/article/view/2513/2456> (accessed March 20, 2020).

Ramak Molavi Vasse'i

We Need a Moratorium on the Use of AI in Critical Impact Areas

The idea of a healthy digital transformation is to develop innovation that enables a better life for all and strengthen individual and collective freedoms and rights. A successful transformational process leaves no one behind. The expectations on artificial intelligence are high: hockey stick business growth, eradication of diseases, immortality, climate rescue, discovering and exploring other planetary systems: not even the sky is the limit when it comes to what we can do with AI.

AI (better: ADM) is applied in ever more areas such as digital platforms and tools, government, health, justice, surveillance and are used for the general curation of information. In many cases, the use has a significant but often invisible impact on the lives of individuals on their opportunities and democratic participation.

AI development and investment strategies of the last years were solely focussed on automation, efficiency and technical viability. Immature systems, trained with low-quality data from questionable sources, implemented without any preceding impact assessment on human rights.

The reality of AI development is an opaque and rashly deploy in increasing areas of a real environment without any public debate or legitimation - hunted by the AI Race narrative.

At the same time, citizens are getting increasingly aware of the downsides of the technology. As a reaction, more than eighty ethical Guidelines were published in the last three years. Most of them aligned on the fact that AI should be fair, transparent, or at least explainable and accountable.

The majority of the guidelines were created by the industry out of external pressure: the threat of new upcoming AI regulation or enforcement of the existing legal framework arose since the GDPR came into force.

But the guidelines were also a countermeasure to mitigate the rising concerns within the own workforce. Movements like the #techwontbuildit or #notokgoogle are whistleblowing about secret business tactics, dark patterns, data breaches and manipulations and scrutinize the purpose of their work. AI experts develop a new awareness of their responsibility for the society, some of them refusing to support cooperations with companies such as Palantir to enable state surveillance and censorship in less free countries than their own.

The growing ubiquity of algorithms in society, the implications and effects need to be heavily monitored and the use has to be actively guided.

The regulatory bodies lack deep expertise, personnel and tools to fulfill their duties. They face ADM systems that did not integrate the legal framework by design.

Numerous governance proposals were already formulated by academia and institutions to overcome this situation. The establishment of 'an FDA for algorithms', a 'right to reasonable inferences', new roles for consumer protection agencies, proposals based on tort liability in combination with algorithm certification by a regulatory agency, mandatory algorithmic impact assessments on human rights and common standards to name some of them are all viable proposals towards the right direction. Besides this, 28 countries in the United Nations have explicitly endorsed the call for a ban on lethal autonomous weapons systems. Unfortunately, all these proposals are still in the ideation phase.

In the meantime, the individual stands defenseless against the massive roll-out of untransparent and uncontrollable ADM systems by companies and governments. Everyone involved in its development and implementation must maintain joint oversight and control over the system. An algorithmic system must be manageable throughout the lifetime of its use. The complexity of a machine-learning system's operations must never exceed the capacity of human oversight and a person's capacity to make changes to the system. If this cannot be guaranteed, the algorithmic system in question should not be used.

According to a newly published study of the oxford internet institute, there is a growing divide in experience and perception between those who use the internet and those who don't and therefore, miss out on access to key services. This could widen the 'digital divide'. Ten times more often than in 2013, data protection concerns were cited as a reason for reluctance and deterrence.

Fast processors, Big Data and Cloud solutions have ended the last AI winter. The lack of control will lead to distrust in technology and their operators. Followed by broad rejection reactions on AI-generated decisions and priotizations, this could lead to the 3rd AI Winter.

The misguided and immature development of digital transformation requires readjustment. We need time for the operationalization of law and IT security in a ubiquitous digital environment.

If it is not about the Movie selection in Netflix or the Music stream but about a decision that can affect human rights and wellbeing: until those safeguards are in place to guarantee control over AI, we need a moratorium on the use of this technology in critical areas.

Thomas Metzinger

Towards a Global Artificial Intelligence Charter[1]

It is now time to move the ongoing public debate on artificial intelligence (AI) into the political institutions themselves. Many experts believe that we are confronted with an inflection point in history during the next decade and that there is a closing time window regarding the applied ethics of AI. Political institutions must, therefore, produce and implement a minimal but sufficient set of ethical and legal constraints for the beneficial use and future development of AI. They must also create a rational, evidence-based process of critical discussion aimed at continuously updating, improving and revising this first set of normative constraints. Given the current situation, the default outcome is that the values guiding AI development will be set by a very small number of human beings, by large private corporations and military institutions. Therefore, one goal is to proactively integrate as many perspectives as possible – and in a timely manner.

Many different initiatives have already sprung up worldwide and are actively investigating recent advances in AI in relation to issues concerning applied ethics, its legal aspects, future socio-cultural implications, existential risks and policy-making [I].There exists a heated public debate, and some may even gain the impression that major political institutions like the EU are not able to react in an adequate speed to new technological risks and to rising concern in the general public. We should, therefore, increase the agility, efficiency and systematicity of current political efforts to implement rules by developing a more formal and institutionalized democratic process and perhaps even new models of governance.

To begin a more systematic and structured process, I will present a concise and non-exclusive list of the five most important problem domains, each with practical recommendations. The first problem domain to be examined is the one which, in my view, is constituted by those issues having the smallest chances to be solved. It should, therefore, be approached in a multi-layered process, beginning in the European Union (EU) itself.

THE "RACE-TO-THE-BOTTOM" PROBLEM

We need to develop and implement worldwide safety standards for AI research. A Global Charter for AI is necessary because such safety standards can only be effective if they involve a binding commitment to certain rules by all countries participating and investing in the relevant type of research and development. Given the current competitive economic and military context, the safety of AI research will very likely be reduced in favor of more rapid progress and reduced cost, namely by moving it to countries with low safety standards and low political transparency (an obvious, strong analogy is the problem of tax evasion by corporations and trusts). If international cooperation and coordination succeed, then a "race to the bottom" in safety standards (through the relocation of scientific and industrial AI research) could, in principle, be avoided. However, the currently given landscape of incentives makes this a highly unlikely outcome.

RECOMMENDATIONS

1. The EU should immediately develop a European AI Charter.

2. In parallel, the EU should initiate a political process leading the development of a Global AI Charter.

3. The EU should invest resources in systematically strengthening international cooperation and coordination. Strategic mistrust should be minimized; commonalities can be defined via maximally negative scenarios.

The second problem domain to be examined is arguably constituted by the most urgent set of issues, and these also have a rather small chance to be solved to a sufficient degree.

[1] Reprinted under the kind permission of the European Parliamentary Research Service, "Should we fear artificial intelligence?" (March 2018) <http://www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA(2018)614547_EN.pdf> (accessed March 20, 2020).

[I] For an overview of existing initiatives, see for example Baum 2017 and Boddington 2017, p. 3p. I have refrained from providing full documentation here, but helpful entry points into the literature are Mannino et al. 2015, Stone et al. 2016, IEEE 2017, Bostrom, Dafoe & Flynn 2017, Madary & Metzinger 2016 (for VR).

PREVENTION OF AN AI ARMS RACE

It is in the interest of the citizens of the EU that an AI arms race, for example between China and the US, is prevented at a very early stage. Again, it may well be too late for this, and obviously, European influence is limited, but research into and development of offensive autonomous weapons should be banned and not be funded on EU territory. Autonomous weapons select and engage targets without human intervention; they will act on ever shorter time- and reaction-scales, which in turn will make it rational to transfer more and more human autonomy into these systems themselves. They may, therefore, create military contexts in which it is rational to relinquish human control almost entirely. In this problem domain, the degree of complexity is even higher than in preventing the development and proliferation of nuclear weapons, for example, because most of the relevant research does not take place in public universities. In addition, if humanity forces itself into an arms race on this new technological level, the historical process of an arms race itself may become autonomous and resist political interventions.

RECOMMENDATIONS

4. The EU should ban all research on offensive autonomous weapons on its territory and seek international agreements.

5. For purely defensive military applications, the EU should fund research into the maximal degree of autonomy for intelligent systems that appears to be acceptable from an ethical and legal perspective.

6. On an international level, the EU should start a major initiative to prevent the emergence of an AI arms race, using all diplomatic and political instruments available.

The third problem domain to be examined is the one for which the predictive horizon is probably still quite distant, but where epistemic uncertainty is great and potential damage could be extremely large.

A MORATORIUM ON SYNTHETIC PHENOMENOLOGY

It is important that all politicians understand the difference between artificial intelligence and artificial consciousness. The unintended or even intentional creation of artificial consciousness is highly problematic from an ethical perspective, because it may lead to artificial suffering and a consciously experienced sense of self in autonomous, intelligent systems. “Synthetic phenomenology” (SP; a term coined in analogy to “synthetic biology”) refers to the possibility of creating not only general intelligence but also consciousness or subjective experiences on advanced artificial systems. Future artificial subjects of experience have no representation in the current political process, they have no legal status, and their interests are not represented in any ethics committee. To make ethical decisions, it is important to have an understanding of which natural and artificial systems have the capacity for producing consciousness, and in particular, for experiencing negative states like suffering [5, 6]. One potential risk is to dramatically increase the overall amount of in suffering the universe, for example via cascades of copies or the rapid duplication of conscious systems on a vast scale.

RECOMMENDATIONS

7. The EU should ban all research that risks or directly aims at the creation of synthetic phenomenology on its territory and seek international agreements [II].

8. Given the current level of uncertainty and disagreement within the nascent field of machine consciousness, there is a pressing need to promote, fund and coordinate relevant interdisciplinary research projects (comprising philosophy, neuroscience and computer science). Specific relevant topics are evidence-based conceptual, neurobiological and computational models of conscious experience, self-awareness and suffering.

9. On the level of foundational research, there is a need to promote, fund and coordinate systematic research into the applied ethics of non-biological systems capable of conscious experience, self-awareness and subjectively experienced suffering.

[5] Thomas Metzinger, “Suffering,” in: The Return of Consciousness, eds. Kurt Almqvist and Anders Haag (Stockholm, 2017), pp. 237–262 <https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_Suffering_2017.pdf> (accessed March 20, 2020).
[6] Thomas Metzinger, “Two principles for robot ethics,”in: Robotik und Gesetzgebung, eds. Eric Hilgendorf and Jan-Philipp Günther, (BadenBaden, 2013), pp. 247–286 <https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_RG_2013_penultimate.pdf> (accessed March 20, 2020).
[II] This includes approaches that aim at a confluence of neuroscience and AI with the specific aim of fostering the development of machine consciousness. For recent examples see Dehaene, Lau & Kouider 2017, Graziano 2017, Kanai 2017.

The next general problem domain to be examined is the one which is the most complex one and which likely contains the largest number of unexpected problems and "unknown unknowns".

DANGERS TO SOCIAL COHESION

Advanced AI technology will clearly provide many possibilities to optimize the political process itself, including novel opportunities for rational, value-based social engineering and more efficient, evidence-based forms of governance. On the other hand, it is not only plausible to assume that there are many new, at present unknown, risks and dangers potentially undermining the process of keeping our societies coherent; it is also rational to assume the existence of a larger number of "unknown unknowns", of AI-related risks that we will only discover by accident and at a late stage. Therefore, the EU should allocate separate resources to prepare for situations in which such unexpected "unknown unknowns" are suddenly discovered.

Many experts believe that the most proximal and well-defined risk is massive unemployment through automatization. The implementation of AI technology by financially potent stakeholders may, therefore, lead to a steeper income gradient, increased inequality, and dangerous patterns of social stratification. Concrete risks are extensive wage cuts, a collapse of income tax, plus an overload of social security systems. But AI poses many other risks for social cohesion, for example by privately owned and autonomously controlled social media aimed at harvesting human attention, and "packaging" it for further use by customers, or in "engineering" the formation of political will via Big Nudging strategies and AI-controlled choice architectures, which are not transparent to the individual citizens whose behavior is controlled in this way. Future AI technology will be extremely good at modeling and predictively controlling human behavior – for example by positive reinforcement and indirect suggestions, making compliance with certain norms or the "spontaneous" appearance of "motives" and decision appear as entirely unforced. In combination with Big Nudging and predictive user control, intelligent surveillance technology could also increase global risks by locally helping to stabilize authoritarian regimes in an efficient manner. Again, very likely, most of these risks to social cohesion are still unknown at present, and we may only discover them by accident. Policymakers must also understand that any technology that can purposefully optimize the intelligibility of its own action to human users can, in principle, also optimize for deception. Great care must, therefore, be taken to avoid accidental or even intended specification of the reward function of any AI in a way that might indirectly damage the common good.

AI technology currently is a private good. It is the obligation of democratic political institutions to turn large portions of it into a well-protected common good, something that belongs to all of humanity. In the tragedy of the commons, everyone can often see what is coming, but if mechanisms for effectively counteracting the tragedy aren't in existence, it will unfold, for example in decentralized situations. The EU should proactively develop such mechanisms.

RECOMMENDATIONS

10. Within the EU, AI-related productivity gains must be distributed in a socially just manner. Obviously, past practice and global trends clearly point in the opposite direction: We have (almost) never done this in the past, and existing financial incentives directly counteract this recommendation.

11. The EU should carefully research the potential for an unconditional basic income or a negative income tax on its territory.

12. Research programs are needed about the feasibility of accurately timed retraining initiatives for threatened population strata towards creative skills and social skills.

The next problem domain is difficult to tackle because most of the cutting-edge research in AI has already moved out of publicly funded universities and research institutions. It is in the hands of private corporations, and therefore, systematically non-transparent.

RESEARCH ETHICS

One of the most difficult theoretical problems lies in defining the conditions under which it would be rational to relinquish specific AI research pathways altogether (for instance, those involving the emergence of synthetic phenomenology or an explosive evolution of autonomously self-optimizing systems not reliably aligned with human values). What would be concrete, minimal scenarios justifying a moratorium on certain branches of research? How will democratic institutions deal with deliberately unethical actors in a situation where collective decision-making is unrealistic and graded, non-global forms of ad hoc cooperation have to be created? Similar issues have already occurred in so-called “gain-of-function research” involving experimentation aiming at an increase in the transmissibility and/or virulence of pathogens, such as certain highly pathogenic H5N1 influenza virus strains, smallpox or anthrax. Here, influenza researchers laudably imposed a voluntary and temporary moratorium on themselves. In principle, this could be possible in the AI research community as well. Therefore, the EU should always complement its AI charter with a concrete code of ethical conduct for researchers working in funded projects.

However, the deeper goal would be to develop a more comprehensive culture of moral sensitivity within the relevant research communities themselves. A rational, evidence-based identification and minimization of risks (also those pertaining to a more distant future) ought to be a part of the research itself and scientists should cultivate a proactive attitude, especially if they are the first to become aware of novel types of risks through their own work. Communication with the public, if needed, should be self-initiated, an act of taking control and acting in advance of a future situation, rather than just reacting to criticism by non-experts with some set of pre-existing, formal rules. As Madary and Metzinger [2] write in their ethical code of conduct, including recommendations for good scientific practice in virtual reality: “Scientists must understand that following a code of ethics is not the same as being ethical. A domain-specific ethics code, however consistent, developed and fine-grained future versions of it may be, can never function as a substitute for ethical reasoning itself.”

RECOMMENDATIONS

13. Any AI Global Charter, or its European precursor should always be complemented by a concrete Code of Ethical Conduct guiding researchers in their practical day-to-day work.

14. A new generation of applied ethicists specialized on problems of AI technology, autonomous systems and related fields has to be trained. The EU should systematically and immediately invest in developing the future expertise needed within the relevant political institutions, and it should do so aiming at an above-average, especially high level of academic excellence and professionalism.

META-GOVERNANCE AND THE PACING GAP

As briefly pointed out in the introductory paragraph, the accelerating development of AI has perhaps become the paradigmatic example of an extreme mismatch between existing governmental approaches and what would be needed in terms of optimizing the risk/benefit ratio in a timely fashion. It has become a paradigmatic example of time pressure, in terms of rational and evidence-based identification, assessment and management of emerging risks, the creation of ethical guidelines, and implementing an enforceable set of legal rules. There is a “pacing problem”: Existing governance structures simply are not able to respond to the challenge fast enough; political oversight has already fallen far behind technological evolution [III].

I am not drawing attention to the current situation because I want to strike an alarmist tone or to end on a dystopian, pessimistic note. Rather, my point is that the adaptation of governance structures themselves is part of the problem landscape: In order to close or at least minimize the pacing gap, we have to invest resources into changing the structure of governance approaches themselves. “Meta-governance” means just this: a governance of governance in facing the risks and potential benefits of explosive growth in specific sectors of technological development. For example, Wendell Wallach has pointed out that the effective oversight of emerging technologies requires some combination of both hard regulations enforced by government agencies and expanded soft governance mechanisms [7]. Marchant and Wallach have, therefore, proposed so-called “Governance Coordination Committees” (GCCs), a new type of institution providing a mechanism to coordinate and synchronize what they aptly describe as an “explosion of governance strategies, actions, proposals, and institutions” [IV] with existing work in established political institutions. A GCC for AI could act as an “issue manager” for one specific, rapidly emerging technology, as an information clearing-

[2] Michael Madary and Thomas K. Metzinger, “Real virtuality. A code of ethical conduct: Recommendations for good scientific practice and the consumers of VR-technology,” Frontiers in Robotics and AI 3 (2016), p. 3. <http://journal.frontiersin.org/article/10.3389/frobt.2016.00003/full> (accessed March 20, 2020).

[III] Gary Marchant (2011) puts the general point very clearly in the abstract of a recent book chapter: “Emerging technologies are developing at an ever-accelerating pace, whereas legal mechanisms for potential oversight are, if anything, slowing down. The legislation is often gridlocked, regulation is frequently ossified, and judicial proceedings are sometimes described as proceeding at a glacial pace. There are two consequences of this mismatch between the speeds of technology and law. First, some problems are overseen by regulatory frameworks that are increasingly obsolete and outdated. Second, other problems lack any meaningful oversight altogether. To address this growing gap between law and regulation, new legal tools, approaches and mechanisms will be needed. Business as usual will not suffice”.

[7] Wendell Wallach, A Dangerous Master. How to Keep Technology from Slipping Beyond Our Control (New York, 2015).

[IV] This quote is taken from an unpublished, preliminary draft entitled „An agile ethical/legal model for the international and national governance of AI and robotics”; see also Marchant & Wallach 2015.

house, an early warning system, an instrument of analysis and monitoring, an international best-practice evaluator, and as an independent and trusted "go-to" source for ethicists, media, scientists and interested stakeholders. As Marchant and Wallach write: "The influence of a GCC in meeting the critical need for a central coordinating entity will depend on its ability to establish itself as an honest broker that is respected by all relevant stakeholders" [4]. Many other strategies and governance approaches are, of course, conceivable. This is not the place to discuss details. Here, the general point is simply that we can only meet the challenge posed by the rapid development in AI and autonomous systems if we put the question of meta-governance on top of our agenda right from the very beginning.

RECOMMENDATION

15. The EU should invest in researching and developing new governance structures that dramatically increase the speed by which established political institutions can respond to problems and actually enforce new regulations.

Conclusion

I have proposed that the EU immediately begins working towards the development of a Global AI Charter, in a multi-layered process starting with an AI Charter for the European Union itself. To briefly illustrate some of the core issues from my own perspective as a philosopher, I have identified five major thematic domains and provided fifteen general recommendations for critical discussion. Obviously, this contribution was not meant as an exclusive or exhaustive list of the relevant issues. On the contrary: At its core, the applied ethics of AI is not a field for grand theories or ideological debates at all, but mostly a problem of sober, rational risk management involving different predictive horizons under great uncertainty. However, an important part of the problem is that we cannot rely on intuitions because we must satisfy counterintuitive rationality constraints.

Let me end by quoting from a recent policy paper titled Artificial Intelligence: Opportunities and Risks, published by the Effective Altruism Foundation in Berlin, Germany:

In decision situations where the stakes are very high, the following principles are of crucial importance:

1. Expensive precautions can be worth the cost even for low-probability risks, provided there is enough to win/lose thereby.

2. When there is little consensus in an area amongst experts, epistemic modesty is advisable. That is, one should not have too much confidence in the accuracy of one's own opinion either way [3].

References

Seth Baum, "A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy," Global Catastrophic Risk Institute Working Paper (November 2017), pp. 17-1 < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070741> (accessed March 20, 2020).

Paula Boddington, Towards a Code of Ethics for Artificial Intelligence, (Cham 2017).

Nick Bostrom, et al., "Policy Desiderata in the Development of Machine Superintelligence," Working Paper, Oxford University, 2017 <http://www.nickbostrom.com/papers/aipolicy.pdf> (accessed March 20, 2020).

Stanislas Dehaene, et al., "What is consciousness, and could machines have it?" Science 358, no. 6362 (October 2017), pp. 486–492. DOI: 10.1126/science.aan8871.

Michael S. A. Graziano, "The Attention Schema Theory. A Foundation for Engineering Artificial Consciousness," Frontiers in Robotics and AI 4 (2017), p. 61. DOI: 10.3389/frobt.2017.00060.

Ryota Kanai, "We Need Conscious Robots. How introspection and imagination make robots better," Nautilus 47 (2017) <http://nautil.us/issue/47/consciousness/we-need-conscious-robots> (accessed 20 March 2020).

[3] Adriano Mannino, et al., "Artificial Intelligence. Opportunities and Risks," Policy Papers of the Effective Altruism Foundation 2, (December 2015), pp. 1–16 <https://ea-foundation.org/files/ai-opportunities-and-risks.pdf> (accessed March 20, 2020).
[4] Gary E. Marchant and Wendell Wallach "Coordinating technology governance," Issues in Science and Technology 31, no.4 (2015), p. 43.

Gary E. Marchant, “The growing gap between emerging technologies and the law,” in The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight, eds. Gary E. Marchant et al. (Arizona, 2011), pp. 19–33.
Peter Stone, et al., Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel (September 2016) <https://ai100.sites.stanford.edu/sites/g/files/sbiybj9861/f/ai100report-10032016fnl_singles.pdf> (accessed March 20, 2020).

IEEE, “Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems,” Global Initiative on Ethics of Autonomous and Intelligent Systems (2017) < https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf > (accessed March 20, 2020).

Robert Gianni

Anticipatory Governance for Artificial Intelligence and Robotics

SIENNA is a three-and-a-half-year project (started in October 2017), with 11 core partners and two associate partners, focussing on ethical and human rights challenges posed by human genomics, human enhancement and human-machine interaction technologies such as robots and smart devices. While these technologies offer significant benefits to individuals and society, they also present significant ethical challenges, e.g., in relation to human autonomy, equality, personal liberty, privacy, and accountability [I][1].

Sienna consortium performed two related studies concerning the analysis of regulatory aspects for AI and robotics: (1) a study of national academic ethical discussions of AI and robotics, and (2) a study of national and supranational discussions of legal and human-rights related aspects of AI and robotics.

Concurrent with writing these studies, the SIENNA consortium has planned, conducted and analyzed citizen surveys in eight EU (France, Germany, Poland, Sweden, The Netherlands, Greece, Spain, and the United Kingdom) and four non-EU countries (United States, China, South Africa, and Brazil), as well as citizen panels in five EU countries (France, Spain, Poland, Greece, Germany; each country around fifty participants) both of which were aimed at obtaining insight into public awareness of and public opinions about present and future developments in AI and robotics. Two reports on this—one regarding the panels and one on the surveys—have been submitted to the European Commission and will be available at https://www.sienna-project.eu. This project represents a great opportunity to advance in the discussion with regard to the governance of AI and Robotics through an investigation of ethical and regulatory frameworks.

The first results from these studies at the hard-regulatory level have shown the following aspects.

The adequacy of the EU legal framework to meet the challenges of AI and robotics is highly differentiated depending on the field. When it comes to the issues of algorithmic transparency and transparency in decision-making, bias, and discriminations and personal data protection, it seems that the reviewed EU data protection framework may potentially offer some legal tools to accommodate these challenges and, in this regard, the EU may be considered a global trailblazer. However, its potential effectiveness largely depends on indirect guarantees that may or may not be used by individuals – for example, a data subject exerting her or his right of access may use this right to detect algorithmic bias (as a first step to fight the bias), but this will require knowledge, skills, time and willingness.

At the national level, it has been noticed that legal academic discourses are established in some countries, while in others they are in their infancy. In many cases, issues pertaining to AI and robotics have attracted the high-level attention of political parties. Overall, there were no major or significant amendments in legislation bearing on constitutional or human rights in direct response to AI and robotics developments reported in the country research for the last five to ten years. In some countries, even in the future this is extremely unlikely to happen (such issues are projected to be left to the courts to adjudicate based on existing laws). With regard to plans to create or adopt new legislation to specifically regulate 'AI' or 'robotics', most countries have adopted a cautious or slow response which has required or left existing laws to be creatively applied or existing regulatory bodies to step in. The national research revealed no regulatory bodies had been created specifically to regulate AI or robotics – at least none with a pure 'regulatory' remit and scope, though there have been calls for these.

In France, for instance, there is no specific law regulating AI or robotics in France. Nonetheless, some pre-existing laws address a number of key issues related to these technologies, in particular the law n°78-17 of January 6, 1978 regarding information technology, files and liberties as modified by the law n°2018-493 of June 20, 2018 that integrates the EU General Data Protection Regulation (GDPR) with national legislation.

[I] This text is based upon the report of the SIENNA project (Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact) - which has received funding under the European Union's H2020 research and innovation program under grant agreement No. 741716.

[1] Agustin Araya, "Questioning Ubiquitous Computing," Proceedings of the 1995 ACM 23rd Annual Conference on Computer Science - CSC 95 (February 1995), pp. 230-237 https://doi.org/10.1145/259526.259560 (accessed March 20, 2020).

[2] John Arquilla and David Ronfeldt, "Swarming & The Future of Conflict," RAND, National Defense Research Institute, 2000.

Developments in AI and robotics challenge the existing legal framework, pose new questions, and create new issues that call for deeper and broader analyses. More profoundly, they challenge a fundamental distinction (e.g. in French law), the one between the subject of the law (as a physical or a legal person) and the object of law. Some legal experts and lawyers in France today enthusiastically support the creation of new legal categories to provide a legal framework for these emerging technologies, including through the recognition of the autonomous system as a subject of right, and not simply as an object, as it is the case today. The lawyer Alain Bensoussan is a particularly vocal proponent of this evolution [5]. A number of legal experts are against this legal development, arguing that the legal distinction between the object and the subject should never be challenged as it is justified by the very nature of the entities at stake (such as for Xavier Labbée [14]). Others argue that it is still too early to move in the direction of such profound transformation of the very structure of French law (such as for Alexandra Bensamoun and Grégoire Loiseau [4]). For these experts, solutions to questions and issues raised by IA and robotics can and should be addressed with the already existing legal, regulatory frameworks.

If aspects of economic growth and innovation are often raised to justify their uptake, the impacts of these artifacts question their ethical appropriateness and their relation with societal needs, values and desires. The analysis based on existing legislation has shown that, due to the severity of certain topics, the rapid technological developments that have taken place over the last ten years, and given the interconnection between different contexts (regulatory bodies, geographical and cultural differences), there are several questions that need to addressed in a broader way before new hard regulatory schemes can be successfully established.

Therefore, the project has also conducted several investigations aimed at identifying the current and future ethical challenges and potential ways to address them beyond the existing legal framework. The ethical analysis has been conducted by following the Anticipatory Technology Ethics approach developed by Philip Brey (2012) [8] .This means that the ethical issues in relation to AI and robotics have been analyzed at three so-called levels of ethical analysis: (1) the technology level, the most general level of description, which specifies the technology in general, its subfields, and its fundamental techniques, methods and approaches; (2) the artifact level or product level, which provides a systematic description of the technological artifacts (physical entities) and procedures (for achieving practical aims) that are being developed on the basis of the technology; and (3) the application level, which defines particular uses of these artifacts and procedures in particular contexts by particular users.

The first outcomes of these studies at the ethics level have highlighted the following aspects.

THE AIMS OF THESE TECHNOLOGIES ARE:

AI – general aims: We found that AI technology is being developed with the following aims in mind: efficiency and productivity improvement; effectiveness improvement; risk reduction; system autonomy; human-AI collaboration; mimicking human social behavior; artificial general intelligence and superintelligence; and human cognitive enhancement. We then considered ethical critiques of each of these aims. We found, amongst others, that efficiency, productivity and effectiveness improvement are inherently tied to the replacement of human workers, which raises ethical issues. The mimicking of social behavior is associated with risks of deception and of diminished human-to-human social interaction. The development of artificial general intelligence and superintelligence raises issues of human obsolescence and loss of control and raises issues of AI and robot rights. Human cognitive enhancement, finally, comes with risks to equality, human psychology and identity, human dignity and privacy.

Robotics – general aims: For robot technology, we found the following general aims: efficiency and productivity improvement; effectiveness improvement; risk reduction; robot autonomy; social interaction; human-robot collaboration; novelty; and sustainability. Most of the ethical issues here mirror those with the aims of AI.

[1] Agustin Araya, "Questioning Ubiquitous Computing," Proceedings of the 1995 ACM 23rd Annual Conference on Computer Science - CSC 95 (February 1995), pp. 230-237 https://doi.org/10.1145/259526.259560 (accessed March 20, 2020).
[2] John Arquilla and David Ronfeldt, "Swarming & The Future of Conflict," RAND, National Defense Research Institute, 2000.
[3] Peter Asaro, cited in Mark Coeckelbergh, "From Killer Machines to Doctrines and Swarms, or Why Ethics of Military Robotics Is Not (Necessarily) About Robots," Philosophy & Technology 24 (September 2011), p. 271.
[4] Alexandra Bensamoun and Grégoire Loiseau, "L'intégration de l'intelligence Artificielle Dans l'ordre Juridique En Droit Commun: Questions de Temps," Dalloz IP/IT (2017), p. 239.
[5] Alain Bensoussan, "Droit Des Robots: Science-Fiction Ou Anticipation?" Recueil Dalloz, no. 28 (2015), p. 1640.
[14] Pascal Labbée, "L'homme Augmenté à l'éprevue de La Distinction Des Personnes et Des Choses," in L'homme Augmenté Face Au Droit, ed. Xavier Labbée (Villeneuve d'Ascq, 2015), p. 47. Quote in the original French language: "le robot n'aura jamais la personnalité juridique… Il lui manque 'l'âme'".

THE IMPLICATIONS AND RISKS OF THESE TECHNOLOGIES ARE:

AI – general implications and risks: We identified the following general implications and risks associated with the development and use of AI: potential negative implications for autonomy and liberty, privacy, justice and fairness, responsibility and accountability, safety and security, dual-use and misuse, mass unemployment, transparency and explainability, meaningfulness, democracy and trust. Frequently recurring ethical issues in these different domains are privacy, transparency, responsibility, fairness, freedom, autonomy, security and trust. For domains in which they are an issue, we discussed their particular manifestations and peculiarities. Healthcare applications of AI raise special issues regarding potential risks to privacy and trust, threats to informed consent, discrimination, and risks of further increasing already existing health inequalities. Law enforcement applications raise issues of bias and discrimination, surveillance, and the risk of a lack of accountability and transparency for law enforcement decisions. Defense applications come with possible negative effects of AI on compliance with the principles of just war and the law of armed conflict, the possibility for uncontrolled or inexplicable escalation, and the potential for responsibility gaps.

Robotics – general implications and risks: We identified the following general implications and risks associated with the development and use of robots: loss of control, autonomy, privacy, safety and security, dual-use and misuse, mass unemployment, human obsolescence, human mistreatment, robot rights, and responsibility and accountability. The topic of companionship covers applications of companion robots, such as robot pets, robot nannies, conversational robots and sex robots. Ethical issues include security, privacy and safety, possible negative implications for human-human interaction, and the appropriateness of certain applications of companion robots, for example, for childcare, elderly care, and sex and romantic relationships. In the service sector, including retail, recreation, restaurants, banking, and communications, amongst others, an issue is the extent to which robots should be able to make decisions without human approval or interference, and the value trade-offs this involves. Two other issues concern the replacement of human workers by service robots and the risk of resemblances to slavery in certain service robot applications. The other mentioned application domains also raise various special ethical issues.

More in detail, the main overall issues that emerged from the analyses and on which it might be worth reflecting are the following [II]:

LOSS OF CONTROL

Human controllers may lose their grip on robotic actions by way of robot evolution. This ethical concern focuses on the wisdom of creating robots that can grow and evolve beyond human understanding and control. Especially regarding the development of biological robots, the biggest concern behind creating self-sustaining and evolving robots is that they one day may surpass human understanding and control. As these types of robots would be very novel entities to humankind, their motivations, decisions, and actions would likely be opaque, leading to high degrees of unpredictability. When thinking of more present applications, unmanned vehicles, and military applications are particularly concerning as they have the means to cause significant amounts of death and destruction with incohesive policies and features to remedy unintended actions. This concern is always worth considering at every advancement of robots in any field as it would prove difficult to regain control once lost.

AUTONOMY

Humans may become fully dependent on robots and may be incapable of survival without their aid. It is not so difficult to see how dependent human beings are on preceding technology, like electricity, running water, internet, telecommunications, automobiles, et cetera. This idea is particularly troublesome as humans are already very dependent upon various technologies and technological infrastructures. Also, if electric grids would somehow go dark, humankind would be in a large amount of trouble very quickly. It is uncertain how much robots would really add to this dilemma, or if it would add to the loss of human independence significantly more than any other technological advancement. In fact, if some of the environmental and maintenance robots are successful, it may help humans become more sustainable if robots are seen not as a fix, but as a redirection for the human community. It is pertinent to be mindful if one is creating robots that enable human self-sufficiency or are being used as an excuse not to change harmful human practices. Further, at each increase of automatization, decisions and the power to decide, however incremental, is be-

[II] These are mainly concerning robots, but given the intersecting nature of the two technologies, most issues can be discussed with regard to both.

ing taken from human beings. At some point, there may be a threshold in which so much decision-making power has been allocated to robots, that humans are unable to make certain types of decisions due to black boxing of necessary information.

PRIVACY

Humans may no longer be able to expect privacy, as it is always possible that the robot may be collecting data and humans do not know what, where, or when. Privacy concerns remain a top ethical dilemma among all types of innovative technologies, and robots are no exception. The more sensory data the robot relies upon to function, the more data it is going to need to be constantly collecting to ensure adequate performance. Whether this data be limited to a need-to-function basis, or additional data is being collected, remains unclear to users. Further, what the data is being used for and who has access to it and control over it leaves much room for ethical input. The more advanced robots become, the clearer the paramount nature of privacy-oriented questions will be, as the roles assigned to robots will heavily depend on the level of trust that can be assigned to them. If the potential for robots reporting confidential information and intimate interactions back to their companies for targeted advertising and analytics is too high, the growth of robots and their uses will be stunted. Even if individuals are willing to sacrifice some privacy for the sake of convenience, it is likely there will be a point of no return to where many robots will only be seen as advanced surveillance devices and not as mere machines or (for some robots) relational Others. The side-effect of this being an increase of social paranoia and a “chilling effect” on society as it is no longer apparent who or what may or may not be observing human behavior.

SAFETY AND SECURITY

Robots could cause a great deal of harm if they suffer a computer security breach or have design flaws. In cases where robots have a large amount of responsibility for humans and trust, for example, hospitals, military contexts, elderly or child care, the prospect of an individual gaining unauthorized access to a robot in these scenarios would be a profound concern, especially if the human interactors are not aware of there being a security breach or unable to regain control of the robot. Accordingly, it is incredibly important that security measures, parameters, and safeguards are implemented and followed that evolve with the robot. If security and safety designs, policies, or procedures begin to lag behind the robot’s societal responsibilities and capabilities, the potential risks are great. Further, even while using robots appropriately and following design protocols, there is the potential for robots to malfunction or function unexpectedly that may potentially lead to human harm. The consequences of machine malfunctions during approved use may be enough to kill the technology’s implementation in near-future applications. Additionally, if robots are particularly susceptible to security breaches or are sneakily reporting data back to its corporate creators for the use of advertising and analytics, sensitive fields, like healthcare, may want to carefully consider if robots are the best fit for them. This also threatens the already-fragile trust of robots at present.

DUAL-USE AND MISUSE

Robots may be used in ways unintended by their creators. This set of ethical dilemmas is really focused upon during the design and creation part of robots, as designers and engineers have the largest hand in eliminating potentials for misuse and dual-use. Unfortunately, even when trying to make design choices that eliminate these possibilities, it is impossible to control for everything. As such, it still stands that sex robots could be used for spying or a food delivery robot could be used to breach buildings. Or friendly security robots could be modified into something more nefarious. There are seemingly few regulations and rules that address the issue of robot modification and misuse; in a way, it is understandable. If the regulations lean too heavily towards the favor of non-modifiable robots, it might be difficult for individuals to perform their own maintenance, repairs, or experiments on their own devices—much like cellular devices of present times. However, with no regulations at all, leaves the question too open-ended, and it may be likely problems will occur similarly to the ethics surrounding 3D printed weapons. For this area of ethics, it is difficult to find a middle ground between beneficial modification allowances and misuse.

MASS UNEMPLOYMENT

There is still much uncertainty about the impact of robots in terms of unemployment. Robots may take over human jobs that cause unemployment rates to rise, but already present issues of exacerbated socio-economic inequality. While it is always important to be mindful of a robot’s impact on the labor market and laborers themselves, many of the concerns tend to be out of proportion to the scale and speed of automation. Further, as more problems begin to surface with fully automated business strategies, many companies are looking towards collaborative robotic solutions. These solutions utilize robots for monotonous or dangerous tasks, while human laborers work with robots on more complicated tasks. Not only bumping up the quality and speed of labor but also easing the burden of these tasks on human workers. While this may lead to a large number of job layoffs from these positions, recent studies suggest that the human job market will flow into the areas required to keep these robots up-and-running and to perform more difficult tasks that robots are not yet capable of achieving. Now, the more concerning area of this rerouting, and one that does not generate as much attention, is the facilitation and worsening of existing socio-economic class stratifications and power- relations. Further, keeping a sharp eye on worker conditions and ensuring that the workload and expectations of laborers are not increased without adequate compensation and training. The jobs themselves do not seem to be as problematic as the societal fallout from such a change.

HUMAN OBSOLESCENCE

Over the long term, we may arrive at a future where robots have become so superior to human beings so that humans will lose their place and purpose. This concern is more often formulated in media and science fiction as “robots taking over the world” and is a concern that is often a combination of other human dignity concerns like: “loss of control”, “human mistreatment”, and “human obsolescence”. Most of these debates and discussions are on many far-off iterations of humanoid androids or robots, but it still stands worth mentioning as these moral and existential concerns will still guide the creation, policy, and research surrounding robots and their advancements, even if they are unwarranted at present. Using ethics to not only help individuals come to terms with robotic others, but also to come to terms with and understand that the meaning of ‘being human’ will also change in a new technological era. The importance of ethics at this time will be as important for guiding the development of humans as it will robots—as many individuals will likely turn to the arts and humanities for guidance when they feel a loss of identity is imminent, as humankind has done in the past with cultural transitions.

HUMAN MISTREATMENT

If the development of robots goes too far, they may evolve to treat humans poorly or harm us. Especially with high risks of inequality and discrimination being learned by robots, it is critical that the algorithms robots are using for decisions and the sensory information gleaned by robots are being carefully monitored for biases. To prevent such situations, some authors call for more transparency in machine decision-making processes and starting data points. While this may not completely fix data biases and discriminatory decisions, it would allow for more participation and monitoring for these problems than black boxing this information would. Furthermore, other researchers suggest setting hard parameters on how robots are permitted to interact with humans, e.g. not killing human beings or no robots allowed in law enforcement. Ethics stands to have much to offer in how this area will develop, and it is important that these frameworks are decided upon and implemented before the robots are given free rein in their roles.

ROBOT RIGHTS

Undoubtedly one of the most complicated issues in robot ethics, the question of robot moral standing respective to humans and animals, is one that generates much debate. Questions on whether moral responsibilities, duties, and treatment are owed to robots, and, if so, to which types of robots and what those duties, responsibilities, and treatment entail, are important. And not only for the sake of the robots, but the ways in which humans treat robots, especially those designed specifically to imitate human beings, may reveal some uncomfortable truths about those human beings that need to be addressed. While it may not be pragmatic to jump to personhood status for, even some, robots like Saudi Arabia has decided, there is something to be said for epistemic caution when approaching the idea of robot rights. At the very least, prohibiting individuals from physically attacking robots, preventing them from performing their assigned roles, or interacting with them maliciously (i.e., bullying) may prove beneficial to paving the way for robotic community members of the future.

RESPONSIBILITY AND ACCOUNTABILITY

If robots cause harm or destruction, who is responsible for reparations? One of the most frequently discussed question, both in the academic spheres and in the media, is that of robot responsibility and accountability. Especially pertinent in ongoing discussions about self-driving (or "autonomous") vehicles, which is to blame when the machine malfunctions? The more complex and black-boxed a machine's decision-making models and processes are, the more difficult it becomes to determine who or what is responsible. This is particularly important when it comes to determining how to compensate damages and harm done by robots— if a self-driving vehicle crashes and kills its driver due to a faulty decision-making protocol, is the company responsible for the malfunction? The QA board for not catching the error before deployment? The driver for not monitoring driving conditions? All of these entities? None of them? Before robotics hit ubiquity, it is critical to establish chains of responsibility for these technologies and formulate legal and regulatory policies to account for non-human decision-makers.

Practical Examples

We will now provide two examples of AI and robotics technology that have particular importance because of the challenges they pose, and so they might be precious for anticipatory governance reflections.

1. EMBEDDED AI AND INTERNET OF THINGS

The concept of Internet of Things (IoT) refers to the interconnection via the Internet of computing devices (which are often embedded in everyday objects) that enables these devices to share and exchange data without requiring human-to-human or human-to-computer interaction. IoT is generally unable to fulfill its promises by itself as it is unable to make sense of the data that is communicated. For this, it often requires artificial intelligence, specifically machine learning algorithms. Such algorithms can interpret the data collected by the IoT to provide a deeper understanding of hidden patterns within the data. This allows the devices to, for instance, adapt to users' preferences or provide predictive maintenance (i.e., prevent possible harms by analysing patterns). Devices that combine IoT and AI are also referred to as "smart devices". Smart devices aim to assist people in their life using technologies embedded in the environment. Some important characteristics of such a device include that it is embedded (the device is "invisible" to the user), context-aware (the device recognizes users), personalized (the device is tailored to user's need), adaptive (the device is able to change according to its environment and/or user), anticipatory (the device can anticipate a user's desires), unobtrusive (the device is discrete) and non-invasive (the device can act on its own, does not necessarily require user's assistance) [11].

Related to the Internet of Things are embedded systems. An embedded system commonly requires little to no human interference and provides a connection between devices. The Internet of Things is a specific type of an embedded system, namely in which the devices are connected through the internet. Applying artificial intelligence to embedded systems creates the concept of Embedded AI. Embedded AI is not limited to embedded systems that are connected through the internet. Devices that combine embedded AI with IoT have the following characteristics: ubiquitous computing, ubiquitous communication and user-adaptive interface [20]. Ubiquitous computing, a term coined by Mark Weiser, refers to "computer use by making computers available throughout the physical environment, while making them effectively invisible to the user" [25]. It aims to "serve people in their everyday lives at home and at work, functioning invisibly and unobtrusively in the background and freeing people to a large extent from tedious routine tasks" [25] .Ubiquitous communication implies that computers have the ability to interact with each other. This can also be seen as a part of ubiquitous computing.

A user adaptive interface or intelligent social user interface (ISUI) has as its main characteristics profiling ("ability to personalize and automatically adapt to particular user behaviour patterns") and context-awareness ("ability to adapt to different situations)" [13]. Devices with the ISUI component are able to "infer how your behaviour relates to your desires" [13]. ISUI includes the ability to recognize visual, sound, scent and tactile outputs [20].

IoT and Embedded AI have several benefits, such as the potential to save people time and money, provide a more convenient life, and increase the level of safety, security and entertainment [20]. This, then, may lead to "an overall higher quality of life" [20]. Although some, if not all, of these benefits are likely, several ethical concerns arise with their

[9] Philip Brey, "Freedom and Privacy in Ambient Intelligence," Ethics and Information Technology 7, no. 3 (2005), pp. 4, 8; pp. 157–166.
[11] Matjaz Gams, et al., "Artificial Intelligence and Ambient Intelligence," Journal of Ambient Intelligence and Smart Environments 11, no. 1 (2019), pp. 71-86.
[13] Johnny Hartz Søraker and Philip Brey, "Ambient intelligence and problems with inferring desires from behaviour," International Review of Information Ethics 8, no. 1 (2007), pp. 7-12.
[20] Mahesh Raisinghani, et al., "Ambient intelligence: Changing forms of human-computer interaction and their social implications," Journal of Digital Information 5, no. 4 (2004).
[25] Mark Weiser, 1991 cited in Sarah Spiekermann and Frank Pallas, "Technology paternalism–wider implications of ubiquitous computing," Poiesis & Praxis 4, no. 1 (2006), pp. 6-18.
[26] David Wright, "The Dark Side of Ambient Intelligence," Info 7, no. 6 (2005), pp. 33–51.

usage, relating to privacy, identity, trust, security, freedom and autonomy [26, 9]. Furthermore, smart technologies may influence people's individual behavior as well as their relation to the world [13, 1] .

Privacy concerns are considered of utmost importance by both critics and proponents of embedded AI and IoT technologies [9]. Four properties of ubiquitous computing that make it especially privacy-sensitive compared to other computer science domains include ubiquity, invisibility, sensing, and memory amplification [16]. Thus, ubiquitous computing is everywhere, unnoticed by humans, with the ability to sense aspects of the environment (e.g. temperature, audio) as well as of humans (e.g. emotions) and potentially creating "a complete record of someone's past" [9]. Regarding the Social Interface, one may add the properties of profiling (i.e. constructing unique profiles of users) and connectedness (wireless connection between devices) [9]. The privacy risks of embedded AI and IoT are considerable due to the aspect of the interaction between devices. It is the combination of the sensitivity of the recorded information, the scale of this recording, and the possibility that interaction of devices facilitates the distribution of personal information to other parties that make embedded AI and IoT so vulnerable to privacy violation [9]. Relating to privacy concerns are concerns about the security of, and trust in, embedded AI and IoT systems. Trust is important for all human-technology relations [24]. If a user has the feeling that the system may have malicious intentions, he or she might be reluctant to use the system. It is thus essential that the user can trust the system.

While IoT and embedded AI may be regarded as fostering freedom due to time and money savings, it may also be regarded as diminishing human autonomy and freedom [9]. Autonomy is commonly regarded as dependent on an individual's ability to make their own decisions and is seen as important due to the opportunity for "self-realization" [9]. Furthermore, freedom and autonomy are closely related. Freedom may be split into two categories; no one must stand in your way, and no one should tell you what to think [9]. Brey (2005) has analyzed the concept of IoT and AI in relation to these types of freedoms, and concludes that IoT combined with AI has a chance to enhance our freedom in both ways: it may "enhance control over the environment by making it more responsive to one's needs and intentions" as well as improve "our self-understanding and thereby helping us become more autonomous" [9]. It simultaneously limits both freedoms by confronting "humans with smart objects that perform autonomous actions against their wishes and "by pretending to know what our needs are and telling us what to believe and decide" [9].

In addition, the use of IoT and embedded AI systems may influence a person's behavior [13]. Søraker and Brey argue that for IoT and embedded AI systems to understand what we want, the behavior humans need to show to a device is similar to the behavior they need to show to a pet; it must be "discrete, predictable and overt" [13]. They claim that this may change our natural behavior. Thus, IoT and embedded AI may force us into changing who we are and how we act; we will then be forced to fit ourselves within this technology. Moreover, some IoT and embedded AI devices may promote their use in solitude, risking isolation of individuals and degeneration of society. Also, as some devices may replace tasks as doing groceries, the "face-to-face interaction between people" might diminish [20], potentially adding to a feeling of isolation. Furthermore, as IoT and embedded AI technologies spread globally, there is a risk of cultural bias. This may result in discrimination of some cultures and encourage "homogenization of cultural expressions" [13]. Finally, IoT and embedded AI systems may lack easy to access and easy to use manual overrides. Søraker and Brey warn for a potential widening between users that simply go along with the requirements of the device and people that try to "game the system" [III]. Not only is there an influence on the individual level, but it has also been argued that the whole relation between men and world may be altered, as the entire world is transformed into a surveillance object [1].

[1] Agustin Araya, "Questioning Ubiquitous Computing," Proceedings of the 1995 ACM 23rd Annual Conference on Computer Science - CSC 95 (February 1995), pp. 230-237 https://doi.org/10.1145/259526.259560 (accessed March 20, 2020).
[6] Jürgen Bohn, et al., "Social, Economic, and Ethical Implications of Ambient Intelligence and Ubiquitous Computing," Ambient Intelligence (2005), pp. 5–29.
[7] Nicolas Bredeche, et al., "Embodied Evolution in Collective Robotics: A Review," Frontiers in Robotics and AI 5, no. 12 (2018), pp.1, 12.
[9] Philip Brey, "Freedom and Privacy in Ambient Intelligence," Ethics and Information Technology 7, no. 3 (2005), pp. 4, 8; pp. 157–166.
[13] Johnny Hartz Søraker and Philip Brey, "Ambient intelligence and problems with inferring desires from behaviour," International Review of Information Ethics 8, no. 1 (2007), pp. 7-12.
[15] Irving Lachow, "The Upside and Downside of Swarming Drones," Bulletin of the Atomic Scientists 73, no. 2 (2017), pp. 96, 98.
[16] Marc Langheinrich, "Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems," Ubicomp 2001: Ubiquitous Computing Lecture Notes in Computer Science (2001), pp. 273–291, p. 6.
[18] Stew Magnuson, "Military Beefs Up Research into Swarming Drones," National Defense Magazine, March 1, 2016 <https://www.nationaldefensemagazine.org/articles/2016/2/29/2016march-military-beefs-up-research-into-swarming-drones> (accessed March 20, 2020).
[19] Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," Ethics and Information Technology 6, no. 3 (September 2004), pp. 175–183.
[20] Mahesh Raisinghani, et al., "Ambient intelligence: Changing forms of human-computer interaction and their social implications," Journal of Digital Information 5, no. 4 (2004).
[24] Spyros Tzafestas, "Ethics and Law in the Internet of Things World," Smart Cities 1, no. 1 (2018), pp. 98-120., pp. 112-115.
[III] Gaming the systems entails that someone may understand how a device responds to a user's behavior, and therefore, intentionally behaves in a specific way to conform the device to his/her own desires. This is problematic if a device is not merely for individual use but rather for an embedded AI device meant to be used by multiple people. See Soraker & Brey, 2007, p. 11.

Finally, some other concerns related to responsibility and accountability. Who decides what the device shares and records [6]? Perhaps the device acts in a way unintended by the designer and unwanted by the user. Who is to blame in such a case [19]?

2. SWARM ROBOTS

Swarm robots, also called "collective robots" or "distributed collaborative systems" are systems that "demonstrate collective decision-making without human help" [15, 7, 18] .They are one of the key emerging fields of robotics research today and are attracting much attention, especially in the military sector, disaster response, and space exploration. Instead of human beings, they can enter dangerous areas (whether in wars or disaster settings for instance) and avoid loss of life and expensive equipment (as individual robots of a swarm are generally simple and inexpensive) [15, 22]. However, they also raise a number of ethical issues that this section identifies. This section begins by highlighting a set of issues that arise with such robots, i.e., privacy and surveillance, risk of hacking, and environmental costs. It also points to the ethical risks created by the use of this technology in the military sector. It concludes with more fundamental conceptual, ontological, and ethical considerations that swarm robots raise.

One of the strengths of swarm robots consists in their highly adaptive nature: they can adapt to any environment, especially changing ones. However, this makes them also particularly unpredictable and therefore, leads to questions of responsibility and accountability. As Singer puts it, "[s] warms may not be predictable to the enemy, but neither are they exactly controllable or predictable for the side using them, which can lead to unexpected results: [...] a swarm takes action on its own" [23]. This technology has great surveillance power, and this raises deep privacy issues. This risk is further exacerbated when swarm robots are designed to be small or invisible or in a way that enables them to covertly penetrate any area [IV]. Furthermore, the decentralized nature of the technology makes it particularly resilient as the destruction of one component does not mean the destruction of the whole system. This makes this technology even more robustly intrusive, and therefore, a potential threat to privacy [21. An additional ethical issue that arises with this technology relates to the risk of hacking and its high dual-use potential that could have significant impacts on human life and society [15]. Another ethical issue relates to their environmental cost, especially "the end of that product lifecycle" [17]. As Lin observes, "[t]hey may contain hazardous materials, like mercury or other chemicals in their battery, that can leak into the environment. Not just on land, but we also need to think about underwater and even space environments, at least with respect to space litter" [17]. As this technology gets wider use, such effects would increase. The use of swarm robots in the military sector also raises ethical concerns [2]. In particular, the faster reactions rendered possible by this technology might lead to an increased risk of quick escalation in military conflict and, eventually, "make it easier to start a war" [3].

Beyond the practical and concrete ethical issues that swarm robots raise, it is essential to point to more fundamental ethical issues that they have the potential to create due to the high degree of autonomy, adaptability, and resilience that they exhibit. As Bredeche, Haasdijk, and Prieto note, swarm robots are characterized by an "autonomy that occurs at two levels: not only the robots perform their tasks without external control but also they assess and adapt—through evolution—their behavior without referral to external oversight and so learn autonomously. This adaptive capability allows robots to be deployed in situations that cannot be accurately modelled a priori" [7]. As such, these robots push one step further the emancipation of the technology from the human creator. In turn, this raises ethical tensions that are, for the moment, insolvable. This tension is exemplified by the position of Bredeche Haasdijk, and Prieto on this technology. While on the one hand, they claim

[2] John Arquilla and David Ronfeldt, "Swarming & The Future of Conflict," RAND, National Defense Research Institute, 2000.
[3] Peter Asaro, cited in Mark Coeckelbergh, "From Killer Machines to Doctrines and Swarms, or Why Ethics of Military Robotics Is Not (Necessarily) About Robots," Philosophy & Technology 24 (September 2011), p. 271.
[7] Nicolas Bredeche, et al., "Embodied Evolution in Collective Robotics: A Review," Frontiers in Robotics and AI 5, no. 12 (2018), pp.1, 12.
[10] Mark Coeckelbergh, "From Killer Machines to Doctrines and Swarms, or Why Ethics of Military Robotics Is Not (Necessarily) About Robots," Philosophy & Technology 24 (September 2011), p. 269.
[15] Irving Lachow, "The Upside and Downside of Swarming Drones," Bulletin of the Atomic Scientists 73, no. 2 (2017), pp. 96, 98.
[17] Patrick Lin, "Drone-Ethics Briefings: What a Leading Robot Expert Told the CIA," The Atlantic, December 21, 2011.
[18] Stew Magnuson, "Military Beefs Up Research into Swarming Drones," National Defense Magazine, March 1, 2016 <https://www.nationaldefensemagazine.org/articles/2016/2/29/2016march-military-beefs-up-research-into-swarming-drones> (accessed March 20, 2020).
[21] Heather Roff, 2015 cited in Irving Lachow, "The Upside and Downside of Swarming Drones," Bulletin of the Atomic Scientists 73, no. 2 (2017), p. 96.
[22] Paul Scharre, "Robotics on the Battlefield Part II. The Coming Swarm," Center for a New American Security (October 2014), p. 5–7.
[23] Peter Singer, 2009 cited in Mark Coeckelbergh, "From Killer Machines to Doctrines and Swarms, or Why Ethics of Military Robotics Is Not (Necessarily) About Robots," Philosophy & Technology 24 (September 2011), p. 273.
[IV] See for instance the swarm robots developed by engineers at the University of Harvard. Programmable Robot Swarms, Wyss Institute, University of Harvard. https://wyss.harvard.edu/technology/programmable-robot-swarms/ [Accessed 20 March 2020]

that we should keep a human in the loop when it comes to swarm robots, on the other hand, they want to design these robots to be the most autonomous possible and, hence defer responsibility to the machine itself [7]. These are two contradictory positions that policymakers, regulators and society will eventually have to decide upon. Coeckelbergh identifies such systems as “cloudy and unpredictable systems, which rely on decentralized control and buzz across many spheres of human activity” [10]. As he demonstrates, swarm robots question classical ethical frameworks founded on an ontology of technology as tools created by humans [10]. In turn, this challenges “the assumptions of our traditional theories or responsibility” [10]. Eventually, swarm robots bring us one step closer to the classic science-fi scenario of machines emancipated from their human creator and the danger that robots take control over humanity. This is even more worrying as this technology is developed in the military sector and could, in the future, be further equipped with weapons. Furthermore, the possible prospect of swarm robots reproducing themselves autonomously through 3D printing makes this concern even starker [7].

[7] Nicolas Bredeche, et al., “Embodied Evolution in Collective Robotics: A Review,” Frontiers in Robotics and AI 5, no. 12 (2018), pp.1, 12.

Mika Nieminen

RRI Experiences for the Implementation of AI Ethics

There are currently several general frameworks and ethical guidelines available on ethical AI. Such guidelines include, e.g. EU's High-Level Expert Group on Artificial Intelligence recommendations (2019) [3], IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and its vision for "Ethically Aligned Design" (2019) and recent top scientists' joint effort "AI4People—An Ethical Framework for a Good AI Society" [1] . While these sets of principles also include recommendations for implementation, they are relatively abstract and general by scope. Thus, it seems that there is a need for methods and ideas to implement these ideas concretely in the design and implementation of AI technology in different contexts [4].

I suggest that one possible direction to explore for implementation ideas is Responsible Research and Innovation (RRI) related studies and development projects. In H2020, European Commission has financed dozens of RRI-projects that deal with the challenge of implementing responsibility thinking and ethics in research and innovation processes [I]. In the following, I introduce the basic ideas of RRI shortly and exemplify its implementation with the approach developed in the ongoing NewHoRRIzon project and provide two concrete examples of AI relevant RRI pilots.

RRI CONCEPT

Responsible Research and Innovation (RRI) is part of a long tradition of Technology Assessment (TA) approaches, bioethics, technology ethics, ethical technology design, and Ethical, Legal, and Social Aspects (ELSA) research and it shares various characters with these approaches including, e.g. interest in social impacts and close dialogue between science and society.

Researchers have suggested several slightly different definitions of RRI, but they share a number of common characteristics, such as a focus on social challenges, engagement of stakeholders, opening up of research and innovation to society, and risk avoidance [6, 2]. For instance, Owen and his colleagues [5] have suggested four basic dimensions of responsible innovation including "anticipation" (analysis of the social, economic and environmental impacts of innovation activity), "reflexivity" (making visible underlying motivations and purposes for innovation activity), "inclusiveness" (inclusion of stakeholder and citizen interests, values and perspectives), and "responsiveness" (learning and changing of action and practices).

European Commission has defined RRI as "an inclusive approach to research and innovation (R&I), to ensure that societal actors work together during the whole research and innovation process. It aims to align better both the process and outcomes of R&I with the values, needs and expectations of European society." (European Commission, 2013) The EC has defined essential elements (so-called "keys") for the RRI to be implemented horizontally across H2020 being public engagement, open access, gender, ethics, science education and governance [II]. EC has also recently emphasized the significance of the alignment of science with society by openness principles, which are usually called "three O:s" (open science, open innovation, and open to the world).

[1] Luciano Floridi, et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," Minds & Machines 28 (2018) p. 689 <https://doi.org/10.1007/s11023-018-9482-5> (accessed March 20, 2020).
[2] Agata Gurzawska, et al., "Implementation of Responsible Research and Innovation (RRI) Practices in Industry: Providing the Right Incentives," Sustainability 9, no 10 (2017), p. 1759.
[3] High-Level Expert Group on Artificial Intelligence, "ETHICS GUIDELINES FOR TRUSTWORTHY AI," (April 2019) <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top> (accessed March 20, 2020).
[4] Jaana Leikas, et al., "Ethical framework for designing autonomous intelligent systems," Journal of Open Innovation: Technology, Market, and Complexity 5, no. 1 (2019) <https://doi.org/10.3390/joitmc5010018> (accessed March 20, 2020).
[I] Some examples of such projects are, e.g. NewHoRRIzon working out the operational basis to integrate RRI into European and national R&I practices and funding; Responsible-Industry focusing on the implementation of RRI in industry and integration of societal actors into research and innovation processes; JERRI supporting RRI transition process within the two largest European Research and Technology Organizations, the German Fraunhofer-Gesellschaft and the Netherlands Organization for Applied Scientific Research TNO; and CO-CHANGE, starting at the beginning of 2020, initiating and piloting RRI related institutional changes in research conducting and funding organizations.
[5] 7 Richard Owen, et al., Responsible Innovation (Oxford, 2013).
[6] Melanie Smallman, "Citizen Science and Responsible Research and Innovation,"in Citizen Science – Innovation in Open Science, Society and Policy, eds. Susanne Hecker et al. (2018).
[II] See more: <https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation> (accessed March 20, 2020).

NEWHORRIZON AND SOCIAL LABS [III]

The general challenge of introducing responsibility and ethics-related action and procedures is one of creating social and organizational change. In NewHoRRIzon, this challenge has been addressed by using a social lab approach as a general methodological tool.

The NH project aims to promote the integration of RRI and "three O:s" into national and international research and innovation processes and funding. In practice, NH is engaging a wide-ranging group of R&I stakeholders from across Horizon 2020 programming and co-creating tailor-made "pilot actions" related to RRI and three O:s. To achieve these objectives, NewHoRRIzon has organized 19 Social Labs, where interventions will be co-created for pilot implementation, one for each Horizon 2020 program line. Social Labs build on a tradition of participatory action research to bring together people with common interests in solving complex problems related to technology and society. Participants have the opportunity to co-create, prototype and test pilot actions and activities to support RRI.

Social labs have initiated over sixty pilot actions, including activities like training (e.g. higher education course material) public and/or stakeholder engagement actions, institutional and governance change, raising general awareness for RRI, and applying RRI in R&D projects and product development.

Preliminary lessons of social labs, which might be useful also in the implementation of AI ethics, include:

1. Need to contextualize RRI and connect it to the practitioners' understanding;
2. Need to demonstrate the benefits of RRI;
3. Need to identify change agents, to network and anchor RRI in existing institutions and practices;
4. Social Labs need diversity, commitment, shared responsibility, active participation, concreteness and flexibility;
5. Pilots may start small. They should be well aligned to every-day work;
6. Social labs involve small group dynamics; these need to be acknowledged and managed;
7. Implementing RRI is challenging;
8. RRI needs to be communicated over and over again.

AI RELEVANT EXAMPLES OF IMPLEMENTATION

The following two examples come from the social lab in the area of security research, where AI was considered an important theme.

Responsible AI framework and evaluation criteria for a funding call

The social lab pilot produced a responsibility framework and evaluation criteria for Council of Tampere Region's (in Finland), European Regional Development Fund call for Responsible AI project proposals. The Pilot Action took place between October 2018 and January 2019. The Pilot Action developed a set of questions related to responsibility aspects of project proposals that were attached to the official project application template. In addition, an evaluation criteria was designed for this set of questions. The integration of ethical principles to the funding call was, to our knowledge, unique in Europe, and the feedback on the pilot was positive.

The responsibility elements included ethics (including integrity, human dignity and individual freedom, equity and equality, and privacy), engagement (comprehensibility, acceptability, and people's control over innovation), openness/ transparency, and safety/reliability. From the applicants was asked questions like "What ethical questions and possible challenges have been identified in the project and how they will be responded to?", "Which actors and stakeholders will be involved in innovation activities and why?", "How will the project achieve openness and transparency?", and "How to ensure the reliability and safety of project activities and results?".

The call received six project applications, of which five were funded and four of them had RRI evaluation included. Funds allocated on this basis were 1,7 m€ [IV]. The pilot had various positive effects including: RRI interested new applicants and new kind of projects were created – also the technological projects had implemented RRI into proposals; the pilot increased regional competencies in responsible AI development; it enhanced innovation in companies by engaging end-users, and it increased information on ethical and responsible AI among citizens and end-users. While evaluators and applicants experienced the criteria easy to use, further development efforts are needed, for instance, by defining how to evaluate RRI equally in different projects varying from technological ones to RRI focusing projects.

[III] See: https://newhorrizon.eu/ (accessed March 20, 2020).
[IV] The reporting of results in here is based on the presentation by Tiina Ramstedt-Sen, Council of Tampere Region 13.6.2019. "Advanced Action Plan Tampere Region".

RRI application tool for SMEs working especially on AI

The still ongoing pilot creates a toolkit for the development and promotion of RRI, especially in AI SMEs. The tool helps SMEs to get a grip on the R&D, technical feasibility, and commercial potential of their innovative idea and develop it into a credible business reporting on the RRI. The tool is targeted to enable SMEs to measure their project performance against tailor-made RRI key performance indicators (KPIs) and monitor them over time [V].

The motivation of this pilot lies in the fact that effective corporate reporting is key to building trust and aligning investment through transparency and accountability. In addition to informing external stakeholders, corporate RRI reporting is also a stimulus for internal conversation and decision-making in relation to the RRI.

The guide outlines a five-step process to embed the RRI in existing business and reporting processes. Step 1 and Step 2 address the process of prioritization of impacts and the identification of RRI for a company to act and report on. Step 3 looks at how to set business objectives, select disclosures and analyze performance. Step 4 and Step 5 offer tips and guidance on reporting and improving RRI performance. RRI reporting tool focuses on seven types of value: technical, commercial, ethical, social, environmental, legal, and political values. AI SMEs are invited to provide input, share best practices and participate in designing the reporting tool. The tool is validated with AI SMEs.

Conclusions

Various RRI projects and implementation experiments may provide important ideas and benchmarks of the solutions for the contextual implementation of AI ethics. This concerns especially so-called “soft forms” of governance, but projects also provide views on the challenges the practitioners may face if “hard governance” (regulation) is used. In addition, RRI studies may provide various practical thinking tools and conceptual frameworks, which help practitioners to think multidimensional responsibility and ethics-related questions.

[V] The pilot is conducted by YAGHMA company <https://yaghma.nl/> (accessed March 20, 2020).

Katharina Sehnert

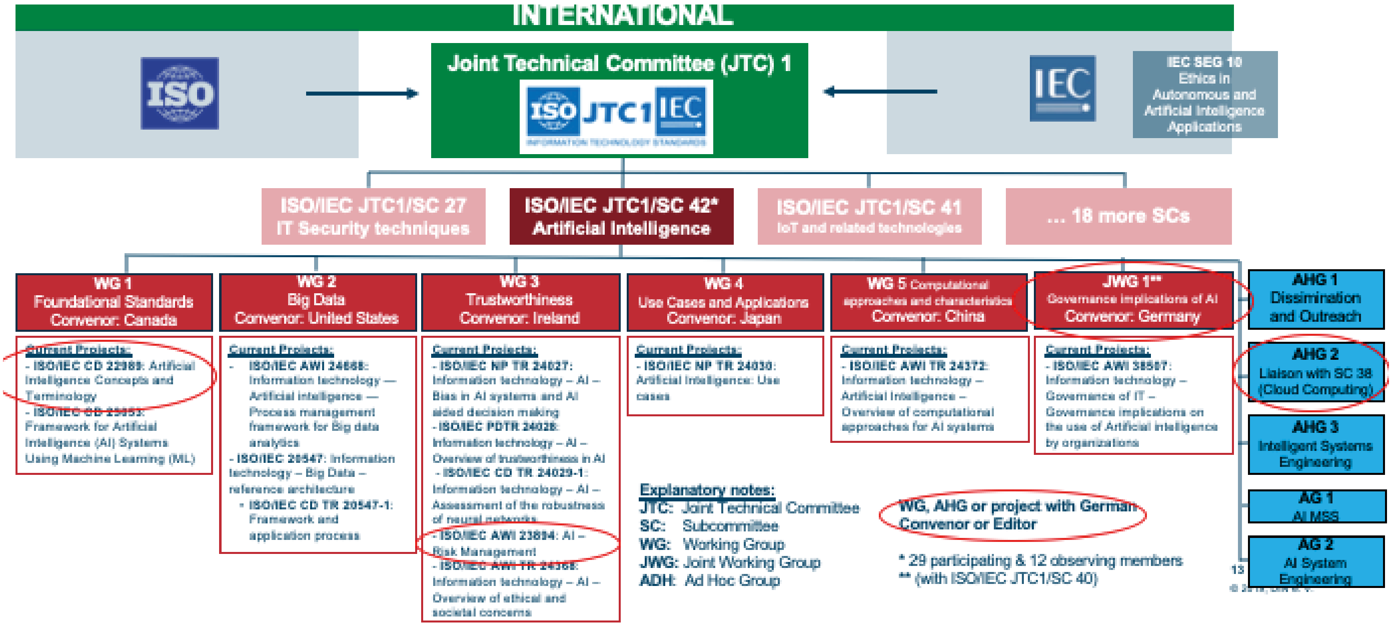
The Value-added of Norms and Standards for Artificial Intelligence

In the national, European and international discussion and the social debate around Artificial Intelligence (AI), many stakeholders express the need for standards to ensure high-quality AI systems and responsible and safe handling of AI technologies.

Standardization has already started doing exactly this. The decision to start national and international standardization of AI was made as early as autumn 2017. Up to now, 29 nations are working on common standards on AI technologies and processes at ISO/IEC JTC 1/SC 42 “Artificial Intelligence”. One of the main focuses of the standardization activities are foundational standards (e.g. terminology and concepts), the trustworthiness of AI systems (including risk management, bias, robustness and ethical aspects), Big Data, use cases, computational approaches and governance implications of AI.

At the European level, a Focus Group on Artificial Intelligence was founded at CEN/CENELEC in 2019, with the aim of developing an AI roadmap by the 1st quarter of 2020. The focus group supports CEN and CENELEC in investigating the necessity of European standardization for AI within CEN-CENELEC, taking into account the guidelines of HLEG and Commission Communication (237/2018). The group will work on a shared vision for European AI standardization and an overview of ongoing standardization activities in the AI environment. Among other things, the technical committees (TCs) in which an AI reference exists and the extent to which there is a need for coordination will be examined. The Focus Group serves as a contact point for European TCs and the ambitions of the European Commission for the standardization of AI.

The German committee that mirrors these international and European standardization activities is at DIN. In a national working committee, 43 experts develop the German position on AI standardization and send delegates to international meetings to write international standards on AI. There are also some national specifications that are being developed at the German standardization body, DIN, at the moment. The first specification on AI that has been published in Germany is DIN SPEC 92001-1: Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 1: Quality Meta Model. The specification is freely available at www.beuth.de/go/din-spec-92001-1.


INTERNATIONAL
ISO
Joint Technical Committee (JTC) 1
ISO JTC 1 IEC
INFORMATION TECHNOLOGY STANDARDS
IEC
IEC SEG 10 Ethics in Autonomous and Artificial Intelligence Applications
ISO/IEC JTC1/SC 27 IT Security techniques
ISO/IEC JTC1/SC 42* Artificial Intelligence
ISO/IEC JTC1/SC 41 IoT and related technologies
... 18 more SCs
WG 1 Foundational Standards Convenor: Canada
WG 2 Big Data Convenor: United States
WG 3 Trustworthiness Convenor: Ireland
WG 4 Use Cases and Applications Convenor: Japan
WG 5 Computational approaches and characteristics Convenor: China
JWG 1** Governance implications of AI Convenor: Germany
AHG 1 Dissimination and Outreach
AHG 2 Liaison with SC 38 (Cloud Computing)
AHG 3 Intelligent Systems Engineering
AG 1 AI MSS
AG 2 AI System Engineering
Current Projects: - ISO/IEC CD 22989: Artificial Intelligence Concepts and Terminology - ISO/IEC CD 23053: Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
Current Projects: - ISO/IEC AWI 24668: Information technology — Artificial intelligence — Process management framework for Big data analytics - ISO/IEC 20547: Information technology – Big Data – reference architecture • ISO/IEC CD TR 20547-1: Framework and application process
Current Projects: - ISO/IEC NP TR 24027: Information technology – AI – Bias in AI systems and AI aided decision making - ISO/IEC PDTR 24028: Information technology – AI – Overview of trustworthiness in AI - ISO/IEC CD TR 24029-1: Information technology – AI – Assessment of the robustness of neural networks - ISO/IEC AWI 23894: AI – Risk Management - ISO/IEC AWI TR 24368: Information technology – AI – Overview of ethical and societal concerns
Current Projects: - ISO/IEC NP TR 24030: Artificial Intelligence: Use cases
Current Projects: - ISO/IEC AWI TR 24372: Information technology – Artificial Intelligence – Overview of computational approaches for AI systems
Current Projects: - ISO/IEC AWI 38507: Information technology – Governance of IT – Governance implications on the use of Artificial intelligence by organizations
Explanatory notes:
JTC: Joint Technical Committee
SC: Subcommittee
WG: Working Group
JWG: Joint Working Group
ADH: Ad Hoc Group
WG, AHG or project with German Convenor or Editor
* 29 participating & 12 observing members
** (with ISO/IEC JTC1/SC 40)
13
© 2019, DIN e. V.

Standardization, especially on the international level, can aid in explaining AI, fostering the safe implementation of AI, and setting a common and safe ground for international trade. The standards on AI developed at ISO/IEC are a joint work of experts from 29 countries. The results of this work - international standards - is most likely to be applied and adopted on an international level. Since European countries and their delegations are well represented at this committee and due to their active presence and collaboration, they have a big impact on the standards that are developed. For example the editor of the first standard to be published, Concepts and Terminology, is German, it was a Swedish proposal to start work on overview of ethical and societal concerns, France pushes the work on robustness of neural networks, Ireland convenes the working group on Trustworthiness and Germany leads the working group on governance implications of AI.

Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 1: Quality Meta Model	DIN SPEC 92001-1	Published
Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 2: Technical and Organizational Requirements	DIN SPEC 92001-2	Ongoing
Guideline for the development of deep learning image recognition systems	DIN SPEC 13266	Ongoing
Quality requirements for video-based methods of personnel selection	DIN SPEC 91426	Ongoing
Transmission of language-based data between artificial intelligences - Universal Namespace Protocol - Specification of parameters and format	DIN SPEC 2343	Ongoing

The strong impact of the European countries on international standardization can be used to bring the European approach of ethics, trustworthiness and quality to the international stage.

Wei Wei

Artificial Intelligence Standardization Efforts at International Level

Symbols and abbreviated terms used in this document

JTC	Joint Technical Committee
SC	Sub Committee
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
AI	Artificial Intelligence
WG	Working Group
NLP	Natural language processing
AWI	Approved Working Item
TR	Technical Report
PDTR	Proposed Draft Technical Report
SEG	Standardization Evaluation Group
OCEANIS	Open Community for Ethics in Autonomous and Intelligent Systems
IEEE	Institute of Electrical and Electronic Engineers
JWG	Joint Working Group
EU	European Union

AI STANDARDIZATION IN SC42

ISO/IEC JTC1 has created SC42 in April 2018 with the focus on standardization of Artificial Intelligence at the international level. The scope of SC42 are:

- **Standardization in the area of Artificial Intelligence**
- **Serve as the focus and proponent for JTC 1's standardization program on Artificial Intelligence**
- **Provide guidance to JTC 1, IEC, and ISO committees developing Artificial Intelligence applications**

Currently, 36 countries have joined. Inside SC42, there are five working groups, where the relevant standard projects are developed. Following are the six key topics in SC42 regarding AI standardization:

Key Topic 1: Foundational Standards

The foundational standards introduce an overview of the AI topic, terminology (vocabulary) and framework, which are common for all systems using AI. The motivation to develop such standards is to give a high-level description of the area and various components,and to provide a basic understanding and common language for a variety of cross stakeholders such as service providers, service operators, service developers, regulators, politics, etc. Two ongoing AI foundational standards in SC42 WG1 (Working Group) are:

- ISO/IEC 22989 Artificial Intelligence Concepts and Terminology
- ISO/IEC 23053 Framework for Artificial Intelligence Systems Using Machine Learning

Key Topic 2: Computational Methods

Computational methods are the heart of AI. SC42 WG5 is now looking at computational approaches and characteristics of artificial intelligence systems. Study of exiting technologies such as machine learning algorithms and reasoning etc. including their properties and characterizes. Analyze of existing AI systems such as NLP or computer vision to understand and identify their underlying computational approaches, architectures and characteristics. Study of industry practices, processes and methods for the application of AI systems. One potential working item currently under discussion is about "the assessment of classification performance for machine learning models and algorithms". A technical report for computational methods is just started

- ISO/IEC AWI TR 24372 Overview of computational approaches for AI systems

Key Topic 3: Trustworthiness
Trust is the necessary aspect for successfully making broad market adoption of AI. SC42 WG3 has set the focus on looking at a wide range of related issues to security, safety, privacy, robustness, resiliency, reliability, transparency, controllability, etc. in the context of AI applications and systems. Following are the list of current projects

- ISO/IEC 23894 AI Risk Management
- ISO/IEC TR 24027 Bias in AI systems and AI aided decision making
- ISO/IEC PDTR 24028 Overview of trustworthiness in Artificial Intelligence
- ISO/IEC TR 24029-1 Assessment of the robustness of neural networks -- Part 1: Overview
- ISO/IEC TR 24368 Overview of ethical and societal concerns

Key Topic 4: Societal Concerns and Ethics
Societal concerns and ethics are hot topics in AI recently. Such concerns can be addressed by standards from the technical perspective. For example, standards can provide guidance to mitigate risk, which has a potential impact on society during the utilization of AI. Moreover, standards can also provide best practices for the development and training of AI systems to mitigate bias algorithmic or training set of data, etc. Considerations of AI impact on society are not limited to SC 42 but extend into ISO and IEC TCs in their applications. Newly created IEC SEG 10 is the entity to consider ethics in autonomous and AI applications. SC 42 collaborates with other external work programs via liaison to work on that topic, such as OCEANIS, IEEE, EU Ethics guidelines for trustworthy AI, etc.

Key Topic 5: Use Cases and Applications
Use cases are very helpful to analyze different AI application domains and different contexts of their use with the goal to identify standardization needs and gaps across different verticals as well as horizontal usage. On the other hand, use cases can be used to validate and to ensure that current on-going AI standard works are broad enough to be used in different domains. SC42 WG4 is collecting AI relevant use cases, which will be described in

- ISO/IEC 24030 AI Use Cases

Key Topic 5: Big Data
Most AI systems are using data for training, testing and running the model. Data handling, processing and managing become more and more important for AI. Former JTC1 Big Data Working Group has now transformed into SC42 as a sub-working group and expanded its scope to look at AI relevant topics such as data quality, data management, data process, and best practices in the context of AI applications or developments. On-going projects in this newly transformed WG2 are:

- ISO/IEC TR 20547-1 Big Data Reference Architecture – Part 1: Framework and Application Process
- ISO/IEC TR 20547-1 Big Data Reference Architecture – Part 3: Reference Architecture
- ISO/IEC TR 20547-1 Process Management Framework for Big Data Analytics

Key Topic 6: Joint Work and Collaboration
AI can be used by different industry verticals and application domains, which may cause the need to develop context-specific AI standards. Developing such standards require experience and knowledge from AI as well as its application domain. Now, the number of such groups is approaching SC42 for joint work and collaboration. Governance implication of AI is currently the first joint work project under JWG1 between SC42 and SC40

- ISO/IEC 38507 Governance implications of the use of artificial intelligence by organizations

CONCLUSION

In general, AI will be one of the major innovation drivers in technology and business digitalization in the coming years. Standardization can effectively support such an innovation process, successfully introduce new technology into the market, create trust in using such technology. ISO/IEC JTC1 SC42 is the subcommittee developing international AI standards. AI standardization is a challenging task. Some key points for consideration to make AI standardization successful:

Innovation collaterally thinking
Standard makes sense, only when technology has reached a certain level of maturity at the market. Making the standard too early can become an obstacle for innovation. Here, standard load map can help people to gain an overview and to provide valuable input on developing an AI standardization strategy.

Broader participation

Since AI covers a wide spectrum of vertical domains, where each has different characteristics and requirements. Making AI standards, which are applied for all vertical domains, demands broader active participation and contribution from experts' groups with different backgrounds and knowledge. Well organized cooperation and coordination

Domain-specific characters may require application dependent AI standards, such AI standards may be developed outside of SC42, or by other consortia. To avoid overlapping or inconsistency during the development of AI context-specific standards, it is necessary to have well-organized cooperation and coordination between SC42 and other committees or consortia.

Standards and open sources

Open source is another main driver of innovation for AI. Open source provides software that can be used for building AI systems or applications. Guidance and interfaces defined by standards can be used effectively by open source development, while experience and feedbacks from open sources can help to effectively improve standard development.

Frauke Rostalski, Markus Gabriel, Stefan Wrobel

KI.NRW – Project for the Certification of AI Applications

Each time holds its challenges in store. We live in the age of digitalization. New technologies are changing the way we live together. They permeate almost every area of society - be it the world of work, road traffic, the health sector, or simply the way we communicate with each other. Even if much of it takes place in silence or as a creeping process, the speed is unprecedented compared to previous social changes and would have scared our ancestors to death at the time of the industrial revolution in the 18th and 19th centuries. A central driving force of digitization is the rapid development of artificial intelligence (AI), which was triggered by breakthroughs in so-called deep artificial neural networks on high-performance computers. Artificial intelligence has a disruptive potential: its scientific and economic applications are so far-reaching that it is hard to predict how artificial intelligence will change our ways of understanding and acting. Also, problem contexts will emerge to which we cannot respond adequately with our traditional legal, political, ethical and social means. It is evident, however, that the use of AI applications will have an impact on society as a whole within a short period of time.

Artificial intelligence is predicted to make an exponential contribution to global economic growth. In the long run, however, this will only be possible if there is sufficient confidence in AI technology. In order to establish trust, an AI application must be constructed in a verifiable manner so that it functions safely and reliably and complies with ethical and legal framework conditions. In addition to the technical protection, it must also be clarified under which conditions the use is ethically justifiable and which requirements arise in particular from a legal point of view. Since AI applications are often based on particularly large amounts of data and the use of highly complex models, it is difficult for users in practice to check the extent to which the assured properties are fulfilled. The certification of AI applications, which is based on an expert and neutral examination, can create trust and acceptance – among companies as well as users and social actors.

The associated challenges touch on fundamental issues that can only be tackled in an interdisciplinary exchange between computer science, philosophy and law. Since artificial intelligence penetrates almost all social spheres, the interests of a large number of actors worthy of legal protection are affected. Legal framework conditions may have to be concretized or newly created. Conversely, however, it must be avoided that over-regulation has the effect of inhibiting innovation or that it becomes too quickly outdated due to the dynamics of technological progress and is, therefore, not applicable at all. For ethics is not fixed once and for all, which is why there is always the possibility of ethical progress and regression in view of social and technological upheavals.

In view of the challenges presented by the use of artificial intelligence, the competence platform KI.NRW has set itself the goal of developing a certification for AI applications that can be carried out operationally by accredited examiners. In addition to ensuring technical reliability, the aim is to test responsible handling from an ethical and legal perspective. The certificate is intended to certify a quality standard that allows providers to design AI applications in a verifiable, legally compliant, and ethically acceptable manner and that also makes it possible to compare AI applications from different providers and thus promote free competition in artificial intelligence. These goals are pursued in an interdisciplinary dialogue of computer science, law, philosophy and ethics. As a result of this interdisciplinary exchange, six AI-specific fields of action for the trustworthy use of Artificial Intelligence are defined: They include fairness, transparency, autonomy and control, data protection as well as security and reliability and address ethical and legal requirements. While security covers the usual aspects of operational security, reliability addresses the particular audit challenges of complex AI models, such as deep neural networks. The latter is further concretized with the goal of operationalizability. The requirements of these fields of action are derived from existing ethical, philosophical and legal principles (such as the general principle of equal treatment). Due to their disruptive potential, it is particularly important for AI applications to ensure compliance with philosophical, ethical and legal frameworks. Their certification primarily serves to protect the legal and ethical interests of individuals. In this way, inadmissible impairments of individuals and groups are to be avoided. AI certification thus pursues the general purpose of averting injustice or ethically unjustified conditions from society. In addition to the individual's freedom rights and the principle of equal treatment, this also concerns general social interests such as the protection and preservation of the environment and the democratic rule of law. From these basic values and principles of a freely ordered community, a multitude of concretizations can be derived, taking into account the constitutional principle of proportionality. In this way, particular fields of action relevant to certification emerge based on ethics and law as well as information technology requirements. For the development of an AI application, it follows in particular from this that the application's scope, its purpose and extent, as well as those affected from it, must be identified at an early stage. All actors directly or indirectly affected must be involved in this process. A risk analysis should be carried out that includes the possibilities of misuse and dual-use. In further development these results must be appropriately taken into account. Finally, the application should be designed in such a way that it can be audited and verified to the extent specified.

The question of how AI applications can be used responsibly and reliably has been the subject of intensive social and scientific discussions in the international arena for some time now. At the European level, the EU Commission has established a so-called HLEG (High-Level Expert Group) for Artificial Intelligence. In April 2019, it formulated recommendations on which aspects should be taken into account in the development and application of artificial intelligence. The certification project KI.NRW takes up these recommendations, differentiates between them and goes beyond them in some areas. This is necessary because the recommendations of the HLEG are primarily general and do not take into account legal aspects – especially not the specifics of the respective national legal systems – nor operational, ethical requirements with the clear goal of certification. In this respect, the certification project is both broad and in-depth in comparison with the proposals of the HLEG: it looks at law alongside philosophical ethics and relates the two to each other. In order to meet the requirements for operationalization, the fields of action developed in this way are in many places more specific and detailed than the HLEG categories.

Participants

PROF. SABINE AMMON

Technical University Berlin

PROF. SUSANNE BECK

University Hannover

NICO GEIDE

German Federal Foreign Office

DR. ROBERT GIANNI

Maastricht University, Brightlands Institute for Smart Society (BISS)

DR. JOANNA GOODEY

EU Agency for Fundamental Rights

DR. ISABELLA HERMANN

Berlin-Brandenburg Academy of Sciences and Humanities

WOLFRAM VON HEYNITZ

German Federal Foreign Office

DR. GEORGIOS KOLLIARAKIS

German Council on Foreign Relations (DGAP)

RAIMOND KALJULAID

Member of Parliament in Estonia

PROF. JENS KRAUSE

Technical University Berlin

DR. MILTOS LADIKAS

Institute of Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology

PROF. PAUL LUKOWICZ

German Research Center for Artificial Intelligence (DFKI)

ISABELLE MESLIER- RENAUD

European Commission, DG HOME

PROF. THOMAS METZINGER

University of Mainz

RAMAK MOLAVI

Digital Rights Lawyer

DR. FRUZSINA MOLNÁR-GÁBOR

Heidelberg Academy of Sciences and Humanities

DR. CHRISTIAN MÖLLING

German Council on Foreign Relations (DGAP)

CLAUDIA MROTZEK

Government Affairs for ORACLE Germany

DR. MIKA NIEMINEN

VTT, Technical Research Centre of Finland Ltd.

DR. THOMAS NUNNEMANN

Federal Ministry for Economy and Energy Germany

DR. ELEONORE PAUWELS

United Nations University (UNU),
Centre for Policy Research

PROF. TIMO RADEMACHER

University of Freiburg

PROF. FRAUKE ROSTALSKI

University of Cologne

KATHARINA SEHNERT

DIN e.V.

KAAN SAHIN

German Council on Foreign Relations (DGAP)

PROF. GÜNTER STOCK

Einstein Foundation Berlin

DR. WOLFGANG TEVES

Federal Ministry of Justice and Consumer Protection Germany

DR. OLAF THEILER

Planning Department of the German Armed Forces

HINRICH THOELKEN

German Federal Foreign Office

DR. OLIVER UNGER

Federal Ministry of Justice and Consumer Protection Germany

WEI WEI

IBM Europe

DGAP

Advancing foreign policy. Since 1955.

Rauchstraße 17/18
10787 Berlin

Tel. +49 30 254231-0

info@dgap.org
www.dgap.org
@dgapev

The German Council on Foreign Relations (DGAP) is committed to fostering impactful foreign and security policy on a German and European level that promotes democracy, peace, and the rule of law. It is nonpartisan and nonprofit. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the German Council on Foreign Relations (DGAP).

Publisher
Deutsche Gesellschaft für
Auswärtige Politik e.V.

ISSN 1866-9182

Layout Luise Rombach

Design Concept WeDo

Photos
Georgios Kolliarakis author photo © DGAP;
Isabella Hermann author photo © BBAW;
Event photos © Angela Laudenbacher

cc BY NC ND

This work is licensed under a Creative Commons Attribution – NonCommercial – NoDerivatives 4.0 International License.