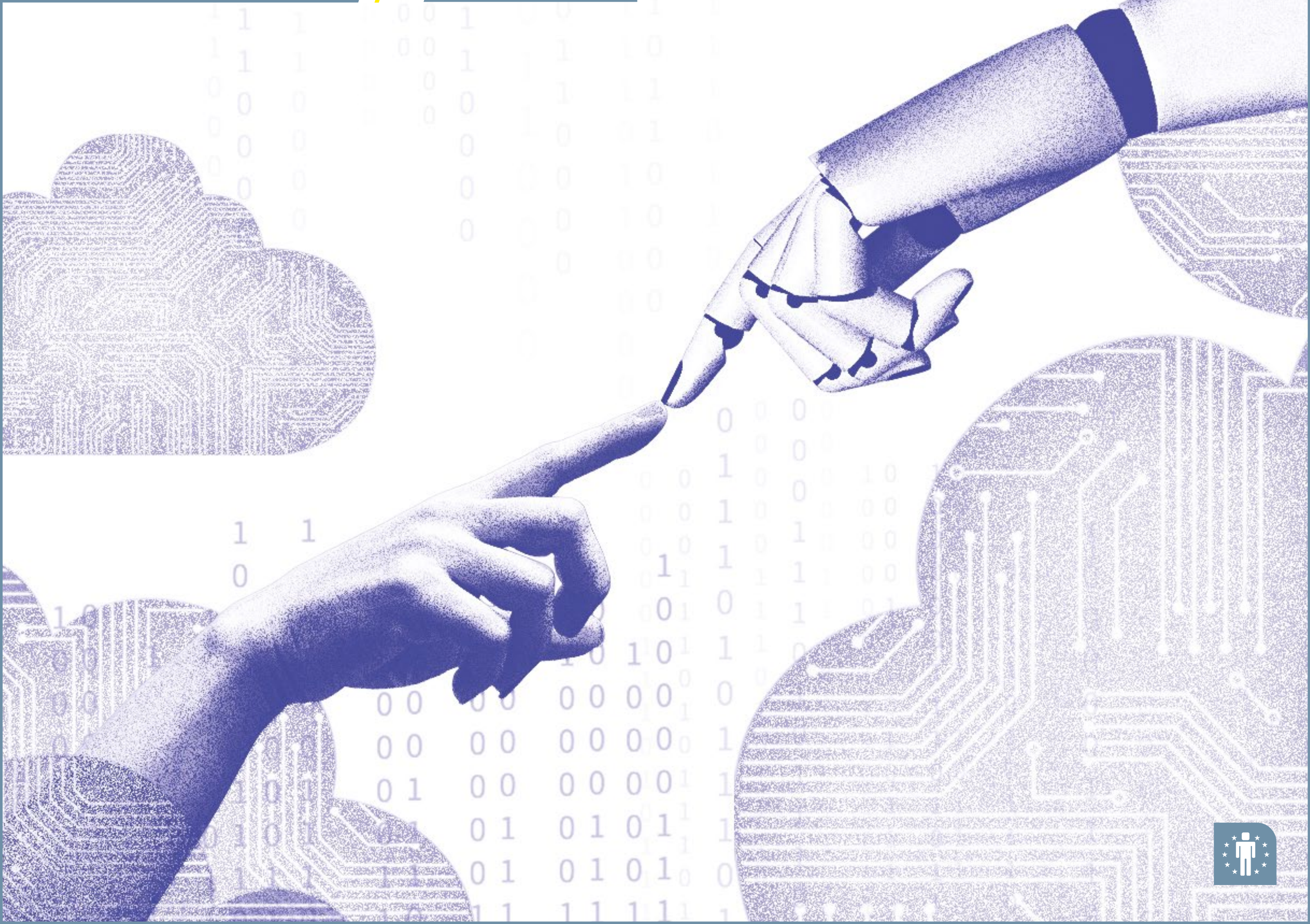
Wilfried
Martens Centre
for European Studies

Artificial Intelligence and Governance

Going Beyond Ethics

Michał Boni



Wilfried
Martens Centre
for European Studies

Artificial Intelligence and Governance

Going Beyond Ethics

Michał Boni

Credits

Wilfried Martens Centre for European Studies
Rue du Commerce 20
Brussels, BE 1000

The Wilfried Martens Centre for European Studies is the political foundation and think tank of the European People's Party (EPP), dedicated to the promotion of Christian Democrat, conservative and like-minded political values.

For more information please visit:
www.martenscentre.eu

External editing: Communicative English bvba
Typesetting: Victoria Agency
Printed in Belgium by Puntgaaf, Kortrijk.

This publication receives funding from the European Parliament.

The European Parliament and the Wilfried Martens Centre for European Studies assume no responsibility for facts or opinions expressed in this publication or their subsequent use. Sole responsibility lies with the author of this publication.

Table of contents

Keywords Data accessibility – Human-centric approach – Trustworthy AI – Principle-based approach – Risk-based approach – *Ex post* assessment – *Ex ante* assessment – Data control – Ethics-based approach – Explainability

About the Martens Centre

The Wilfried Martens Centre for European Studies, established in 2007, is the political foundation and think tank of the European People's Party (EPP). The Martens Centre embodies a pan-European mindset, promoting Christian Democrat, conservative and like-minded political values. It serves as a framework for national political foundations linked to member parties of the EPP. It currently has 31 member foundations and two permanent guest foundations in 25 EU and non-EU countries. The Martens Centre takes part in the preparation of EPP programmes and policy documents. It organises seminars and training on EU policies and on the process of European integration.

The Martens Centre also contributes to formulating EU and national public policies. It produces research studies and books, policy briefs and the twice-yearly *European View* journal. Its research activities are divided into six clusters: party structures and EU institutions, economic and social policies, EU foreign policy, environment and energy, values and religion, and new societal challenges. Through its papers, conferences, authors' dinners and website, the Martens Centre offers a platform for discussion among experts, politicians, policymakers and the European public.

About the author

Michał Boni, Ph.D., was involved in many activities of the Polish Transformation from 1989 as both a minister and adviser, and is also author of the long-term strategy Poland 2030. The first minister of digital affairs in Central Eastern Europe (2011–13), and one of the leaders of the consultations and work on the General Data Protection Regulation from 2012, supporting strong cooperation from the various partners, he is also the author of the unique document *Digital Poland* which outlined ways to use EU structural funds (2014–20) to develop the digital economy and society. As a Member of the European Parliament (2014–19), especially active in the Committee on Civil Liberties, Justice and Home Affairs and Committee on Industry Research and Energy, he was responsible for reports on subjects including the 5G road map and gigabit society development, interoperability issues, the European Open Science Cloud, the Re-Use of Public Sector Information Directive, ePrivacy solutions, artificial intelligence and the General Data Protection Regulation (and its implementation). He has also authored reports on eHealth development. He is now working in cooperation with the Martens Centre, the University for Social Sciences and Humanities in Warsaw, SMEEurope and SMEConnect, the Advisory Board of the Digital Enlightenment Forum and Blockchain for Europe.

Executive summary

Artificial Intelligence (AI) is changing our world. This new phenomenon carries many threats, but also offers many opportunities. We need to find a suitable framework to support trustworthy AI. A key challenge remains: can we, as humans, retain control over the technology or will the technology take control of humanity? In responding to this challenge, the following question needs to be considered: What kinds of tools are needed, not only to keep control of AI development, but foremost to multiply the possible opportunities it offers?

The current pandemic has shown how useful and important AI can be in helping us to fight COVID-19. Moreover, it has clearly demonstrated that we cannot afford *not* to utilise it, nor do we have time to lose with regard to its development.

Hence, it is our responsibility to urgently establish an adequate framework for the development of AI systems based on a revision of the existing law and followed by possible new legislative proposals with a clear focus on future-proof tools. We have to generate a suitable governance model that not only has its foundation in law, but that also ensures democratic oversight through the open and collaborative participation of all partners and the validation of AI solutions by science and society. We should build trustworthy AI based on a human-centric and principled approach. The practical implementation of ethical rules in the design of AI (through the existing *ex post* model of analysing the consequences, including unintended ones, as well as a new *ex ante* model that provides an impact assessment in the early stages of development) and the evaluation of the everyday functioning of AI systems are essential.

It will not be possible to develop AI and claim all its economic and social benefits without a clear model for data use (including flows, collection and processing) that fully respects fundamental rights and the principles of cybersecurity. It will not be possible to build trustworthy AI without transparent rules for the relationships between its users (workers, citizens and consumers) and AI designers, developers and deployers (with the symmetry of information required, e.g. practical schemes for 'explainability'). It will not be possible to accurately implement various AI functionalities without undertaking risk assessments and introducing mechanisms to manage those risks.

To achieve all of the above, we need compromises at various levels: between European institutions and stakeholders (businesses, citizens and consumers, taking into account rights), between European institutions and member states (based on common and harmonised solutions), and between political groups, which are currently more focused on their differences than similarities. How can these compromises be achieved swiftly?

The answer is multidimensional and complex; however, we should be brave enough to pursue it. Paradoxically, the unfortunate experience of COVID-19 has brought a lot of positive momentum to our search for answers, proving to be a real AI development game-changer.

Introduction

It is essential to understand how important it is for the EU to be a driving force behind the development of Artificial Intelligence (AI) and to ensure the proper conditions are created for the success of the Data Strategy, which will be a key factor in future growth in many areas. Therefore, the European Commission's decision to start the debate on AI[1] and the Data Strategy[2] in parallel is welcome. What is more, it shows a new paradigm in understanding the Digital Single Market concept as a holistic solution. In this paradigm the conditions needed to complete the Digital Single Market are a conglomeration of AI, data strategy, elements of the new data-based industrial policy, new forms of functioning for e-commerce (e.g. as in the Digital Services Act) and schemes for the development of cybersecurity. All of these contribute to the strategic autonomy of the European economy, which is crucial and necessary for future European competitive advantage. Digital scaling-up is possible only if we can overcome the separation of elements into silos and use every opportunity, whether stemming from digital technologies or from new schemes for services and industrial development. All of these aspects must be included in a consistent regulatory framework.

Moreover, it is indispensable to consider the impact of the COVID-19 pandemic on AI development and the possible impact of AI on solving problems created by the outbreak. AI has supported analytical efforts to simulate the trajectories of the spread of the virus. The use of data in algorithmic processes has helped many countries to trace infected individuals and warn those who may have come into contact with them. However, the use of all such personal data should comply with rules on the protection of privacy and security, and apply the principles of anonymisation and sunset clauses. AI has already enabled a quicker and deeper analysis of the genetic sequence of SARS-CoV-2. As a result of the application of AI during the pandemic, its potential has become much clearer and more understandable to the public, showing its paramount advantage for society.

When I describe the challenges and opportunities of AI, it should be clear that I am referring to the AI systems of algorithmic processes, machine learning (ML) and the development of AI per se; in other words, I mean AI in the broadest sense, in line with the definition used by the European Commission

[1] European Commission, *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*, White Paper, COM (2020) 65 final, 19 February 2020.

[2] European Commission, *A European Strategy for Data*, Communication, COM (2020) 66 final, 19 February 2020.

following the High Level Expert Group on AI (HLEG on AI).[3] However, it should be noted that a broad definition of AI may lead to many practical problems. For example, it can create difficulties in building the proper conformity assessment model—some algorithms are changed very frequently to improve solutions and make them more secure and in these situations there is no time to provide updated assessments.

Political debates on the many different aspects of AI development have long been ongoing. Some of the conflicts address essential aspects, for instance

- whether to start with a stringent regulatory framework or to be more open to ‘soft law’ solutions, allowing for a step-by-step approach to achieve legal conditions;
- whether to be strict with respect to ensuring fundamental rights (privacy etc.) or to allow a looser interpretation of the ‘legitimate use’ of the General Data Protection Regulation (GDPR), allowing for the collection and processing of data under reasonable conditions; and
- whether to perceive the use of AI models only as threats or risks or to be open-minded in also understanding their advantages.

Political groups articulate these various issues very clearly; still, there is widespread general support for the development of AI across the political spectrum. Broadly speaking, the European People’s Party—despite having doubts regarding the regulatory framework—wants to support the development of AI, especially the accessibility of data for AI growth and training (which is key to many industries, the data economy and the changing health care system). The European Conservatives and Reformists group follows a similar logic. Renew Europe and the Socialists have both raised concerns regarding the lack of proper fundamental rights protection and would like to see limitations put in place with regard to data collection.

[3] ‘Artificial Intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions’; HLEG on AI, *Ethics Guidelines for Trustworthy AI*, (April 2019), 36.

My objective is that this paper will offer substantive arguments to make the political discourse more compromise and solutions orientated. Such an approach to AI development also requires that a balanced solution is found, in which data availability can be combined with data security and transparency regarding its collection, use, reuse and processing, with full respect for all aspects of European law.

The element of transparency means that there must be more reasonable and responsible controls in place regarding the data used by AI for training and the creation of new products and services. On the one hand, the key question is how to make control (both human control and through the use of a data-control-based approach) over all kinds of data more tangible, implementable and transparent for data originating both within and outside the EU. On the other hand, there is the question of how to categorise this data, whether as personal or non-personal, and industrial structured or non-structured data. At the same time, the availability of data should not be limited (by borders or by the specificity of some fields), because this would risk halting AI development. The more efficiently (quickly and responsibly) data can be accessed, the more qualified the results of the work of AI will be, with all the accompanying benefits to the economy and people. Finally, it is important to ensure that the use of data remains compliant with the 'FAIR' principles: findable, accessible, interoperable and reusable.[4]

One of the key challenges in the development of a data-based economy and AI opportunities relates to trust. On the one hand, there is a common understanding that AI offers many opportunities, but on the other there is fear about the many threats it poses. Some of these are based in reality, for example, concerns about how the labour market will look—including fear that human jobs will be replaced by robots and AI and the dangers associated with this. However, other fears are often based on ideas from popular culture (e.g. *Frankenstein* and the television series of *Black Mirror* and *Westworld*) which has established many stereotypes and prejudices.

AI should be treated as a General Purpose Technology (i.e. a technology that will have a powerful impact on further technology development) that will have a visible impact on many areas of our everyday life (e.g. health care systems) and offer new economic possibilities. These possibilities depend on our openness to and acceptance of the new technologies and, ultimately, depend on how trusting society is

[4] Commission Expert Group on FAIR Data, *Final Report and Action Plan: Turning Fair Into Reality* (2018).

of AI. But how should the concept of trust be understood in terms of common psychological attitudes? In this regard we need to understand the meaning of the terms ‘undertrust’ (i.e. a deficit of trust and consequent lack of confidence regarding the new world of AI) and ‘overtrust’ (i.e. overly optimistic attitudes focused only on the possibilities),[5] and avoid their dangers by focusing on how to build trust. As emphasised in the European Commission *White Paper on AI*, we need to build an ecosystem of trust. Practically, this would achieve the objective of establishing trustworthy AI development. What is additionally crucial is that it would give AI credibility, which is fundamental for economic and societal development. The essential premise of the human-centric approach I advocate here should be to achieve trustworthy AI models.

To implement this approach, some conditions are needed:

- There needs to be full respect of privacy and fundamental rights, as well as of the ethical values linked to some dimensions of the functioning of new technologies (as seen in the recommendations of the Commission’s HLEG on AI[6]). This will create a principle-based approach, with proper clarity of the differentiation between principles and the tools used to ensure respect of principles.
- The capacity must exist to analyse all threats and frame them as risks, which is key to many dimensions of the functioning of AI. This would support the use of a risk-based approach, with the consequence of assessing the levels and weights of risks and proposals in order to manage risk. It must be considered that probable high-risk solutions (e.g. those in health care) can very often bring large advantages and benefits for all. Such an approach will be crucial for shaping the best regulatory framework and model of governance.
- Technical and business solutions are needed which can strengthen trust and crucially put pressure on both the implementation of standards and the model of design for AI services and products. Here the fundamental question is whether to use an *ex post* model to analyse possible negative impacts (as exists now) or an *ex ante* model to foresee the impact of AI on humans, societies, economies,

[5] This subject was raised during a video conference between representatives of the technology company Philips and the author; the participants from Phillips were Hans Aloys Wischmann (Programme Manager AI, Precision Diagnosis) and Jan Willem Scheijgrond (Head of Global Government and Public Affairs).

[6] HLEG on AI, *Ethics Guidelines for Trustworthy AI*.

values and so on. We may not be able to predict all the effects of automated decision-making processes (i.e. where AI plays the role of the decision-maker), but it would be better for public trust to check the processes of design and to monitor the compliance of pilot projects with principles and values (i.e. ensuring privacy-by-design, ethics-by-design solutions). This is the chance to make the shift from a ‘black box society’ to an ‘auditing society’. At the same time, it would also be useful to open up the idea of combining *ex post* and *ex ante* models, which could positively influence some areas. This can already be observed in health innovations.

- Trust needs to be built in this complicated ecosystem that is being created by AI, and this requires proper oversight. In assessments of the challenges facing AI development, building trust is important both for transparent management and for the much greater participation of users, citizens, workers, patients and clinicians—in terms of the development of e-health solutions—and consumers.[7] One of the key expectations is the development of a clear liability model, in which people who have been harmed by AI systems should receive the same level of protection as those who have had negative experiences in other systems. Which model of oversight would be best? And what kinds of rules, involving which partners, stakeholders and institutions, should this be based on? It is essential that this model is democratic. We need to discuss all aspects of and possibilities for the implementation of democratic oversight.

[7] This issue was raised during my conversation with European consumer organisation expert David Martin; see also D. Martin, *AI Rights for Consumers*, BEUC, ref. BEUC-X-2019-063-23/10/2019 (October 2019).

A two-step approach

The dominant view in many papers on the subject is that AI will not move towards developing human-like perception. It is crucial not only to understand the role of AI better, and the significance of the development of deep learning and algorithms, but also to imagine and implement a new research and development framework, competitiveness rules for this area, limits (or a lack of limits) for experimentation and education.

Meeting the educational challenges is critical both for the current development of AI and for dealing with unpredictable future scenarios, such as the development of AI with human-like perception. This will be especially important when AI takes on a humanoid form. Without new digital AI literacy, the understanding of people and societies will lag behind the development of technology.

As a result, we need to promote a two-step approach for AI development. The first step should focus on temporary challenges, take a mid-term perspective (approximately 5–10 years) and address the following questions:

- How can European competitive advantage be built vis-à-vis Chinese and US development (if we compare investments[8] in this area it is clear that we are not currently the leaders)?
- How can we prepare industries for the entrance of AI?
- What kind of human control over technology is possible?
- What kind of regulatory framework is needed? How can innovation and regulation be combined? How can we ensure the constant development of technologies while fully respecting the principles of ethics and fundamental rights, adequately translated into regulations in various forms (legal acts vs. soft laws)? How can we make the rules for AI development future-proof?[9]
- What kind of institutional framework is needed to support the development of AI technologies

[8] In 2017, private investment in AI in the EU was in the range of €2.4–€3.2 billion, compared with €6.5–€9.7 billion in Asia and €12.1–€18.6 billion in North America. Similar ratios remain when taking into account public expenditures too. See European Commission, *Artificial Intelligence for Europe*, Communication, COM (2018) 237 final, 25 April 2018, 5. For China and US see C. Cornillie, 'Finding Artificial Intelligence Money in the Fiscal 2020 Budget', *BGov.com*, 28 March 2018; K. Hao, 'Yes, China Is Probably Outspending the US in AI – But Not on Defense', *Technology Review*, 5 December 2019; and J. Bughin et al., *Tackling Europe's Gap in Digital and AI*, McKinsey, February 2019.

[9] This issue was already partially visible during discussions on the Medical Devices Directive some years ago.

that meet human needs, based on future-proof regulations and allowing for effective stakeholder collaboration?

- How can people of all generations be educated about AI?
- How should the advantages of AI solutions in everyday life be presented?
- How can the various threats related to the appearance of AI in public discourse be overcome?

Taking into account the above-mentioned questions, we need to work on a proper regulatory framework (i.e. the rules and regulations, 'hard' and 'soft' laws, the exchange of best practices, networks of cooperation) that will introduce some firm resolutions. This framework should be future proof, which means that as much flexibility is incorporated (to update and adjust institutional solutions as the technology changes) as is needed to guarantee respect for human and consumer rights and, in parallel, conditions for innovative development. In this sense, 'flexibility' means the ability to accommodate technical changes while ensuring legal certainty.

The second step should focus on long-term challenges (taking a perspective of approximately 15–20 years) and present much more future-oriented solutions, accepting all the inherent uncertainties and unpredictability that these involve. It should address future AI (e.g. full neural net developments, with the possibility of AI emotion and the model of singularity) and also some new foreseeable challenges, as described in a Centre for European Policy Studies (CEPS) report by Andrea Renda:[10] 'Recent breakthroughs in AI, mostly due to the use of Deep Learning and Deep Reinforcement Learning techniques, are first steps towards distancing the acts of the AI system from the will of the programmer'.

These two perspectives need to be viewed in distinct ways. However, the work on introducing and developing the above-mentioned steps should take place in parallel. Our initial orientation needs to happen very fast—we have to implement the proper framework for our actions as soon as possible.

[10] A. Renda, *Artificial Intelligence. Ethics, Governance and Policy Challenges*, CEPS (Brussels, February 2019), 14.

Values and ethics: challenges

Some years ago, when the debate on the relationship between humans and societies and new technologies started, it was less clear how important values and ethics were vis-à-vis the development of technology. Attention was focused on convenience for users, full accessibility to the new technological opportunities for all and making the new solutions fully understandable for everybody. Now, however, we are discussing the ethical dimensions of interactions between humans and machines, and there are two crucial requirements for the proper development of those interactions: awareness and control.

The human-centric and principle-based approach

If we want to build trustworthy AI systems, we have to be aware not only of threats, but first of all of our rights as they relate to values and ethical principles. This is the essential foundation of the human-centric and principle-based approach.

Today, there is considerable support for adequate debates on these issues. This is evidenced by the publication of documents such as the *Ethics Guidelines for Trustworthy AI* by the Commission's HLEG on AI,[11] the work done by the German Data Ethics Commission,[12] and the analyses prepared by institutions such as Digital Europe[13] and the European Consumer Organisation.[14] Work on this subject has also been presented by companies such as Philips, Siemens, Google and Microsoft.[15] This clearly

[11] HLEG on AI, *Ethics Guidelines for Trustworthy AI*.

[12] Germany, Data Ethics Commission, *Opinion of the Data Ethics Commission* (Berlin, October 2019).

[13] Digital Europe, *Digital Europe's Recommendations on Artificial Intelligence Policy* (13 November 2019).

[14] J. Malinina, *AI Must Be Smart About Our Health*, BEUC, ref. BEUC-X-2019-078-02/12/2019 (December 2019); Martin, *AI Rights for Consumers*; C. Schomon, *Automated Decision-Making and Artificial Intelligence – A Consumer Perspective*, BEUC, ref. BEUC-X-2018-058-20/06/2018 (June 2018).

[15] From conversations with representatives of (1) Philips: Hans Aloys Wischmann (Programme Manager AI, Precision Diagnosis) and Jan Willem Scheijgrond (Head of Global Government and Public Affairs); and (2) Siemens Healthineers: Hanna Marie Hoehn and Lars Rohwer. Google, *Responsible Development of AI* (2018); Google, *Perspectives on Issues in AI Governance* (2018); S. Pichai, 'AI and Google: Our Principles', Google blog, 7 June 2018; J. Gennai, 'Google's AI Principles' (2019, unpublished presentation seen by the author); P. Marczuk, 'Artificial Intelligence. Ethics and Regulations', Microsoft (unpublished presentation seen by the author).

shows how important the topic is. In addition, in many member states work to prepare national strategies for AI development, including the ethical dimension, is being carried out.

Over the course of many years, we have been able to observe the shift from human control over technology to new models in which use of the term ‘autonomous’ has a specific meaning. Some years ago, autonomous cars were the subject of science-fiction novels or films. In real life, they were a visionary concept. Now, we are not only talking about them, but they have already been designed: autonomous cars; the autonomous decisions made by algorithms and AI in health care; the administrative decisions supported by ML and algorithms; AI-enabled mass-scale scoring; the e-commerce calculations on transactions and prices based on analytical efforts (albeit with the potential for discrimination—there are many examples that suggest that profiling algorithms could be discriminatory); models for manipulating behaviours and minds during electoral processes, undermining democracy; autonomous weapons (dangerous because of their non-transparent decision-making processes); and security solutions managed by machines, such as mass surveillance or facial-recognition models. Obviously, these ‘autonomous’ solutions raise concerns.

Of course, these ‘autonomous models’ are very useful and convenient, improving management skills, procedures and possibilities, and working towards smart solutions. They are important and needed for drivers, patients, workers, users, consumers and citizens. They bring efficiency and effectiveness to many areas. But, because the decision-making processes used are not known to us (due to the information asymmetry of the algorithmic sequences between users and AI or the algorithmic developers and deployers), we—as human beings in our various social roles—have many concerns about the transparency and accountability of those processes.

As a result, we need to return to a model in which we—as humans—can take control of technological solutions. This means that those solutions (the concepts behind new products and services, and the everyday functioning of many devices and apps) should be based on certain principles and should require us to better understand (and be aware of) the rules which govern the work of AI-based processes, services and products. As always, as in many cultures and in many marketplaces, these principles relate to our fundamental rights.

Box 1 Ethical principles as presented by the HLEG on AI

In their *Ethics Guidelines*, the HLEG on AI identified five groups of fundamental rights:[16]

- respect for human dignity (humans should be treated as moral subjects, not objects for manipulation, scoring or sorting);
- freedom of the individual (sovereignty of decisions and choices, equal accessibility to all opportunities, privacy and security protection);
- respect for democracy, justice and the rule of law (AI as a guardian of democracy, not the instrument of its destruction);
- equality, non-discrimination and solidarity (crucial to avoiding unfairly biased outputs; key to the quality and transparency of the data used for training AI and the fair processing of data, especially those relating to vulnerable groups); and
- citizens' rights (maintaining the rights of citizens vis-à-vis governments, authorities and all other powers; based on international law, which means the same rights for all people, from all parts of the globe).

Translating those rights into ethical principles, the HLEG on AI[17] mentioned four principles:

- respect for human autonomy,
- prevention of harm,
- fairness, and
- explicability.

[16] HLEG on AI, *Ethics Guidelines for Trustworthy AI*, 10–11.
[17] Ibid., 12.

The European Commission has produced a Communication describing the seven key requirements indicated in the *Ethical Guidelines* of the HLEG on AI.[18] They are

- human agency and oversight;
- technical robustness and safety;
- privacy and data governance;
- transparency;
- diversity, non-discrimination and fairness;
- societal and environmental wellbeing; and
- accountability.

The essential challenge is how to implement those rights and principles not only in reality (with due consideration for individuals' and businesses' patterns of behaviours), but when building completely new technological solutions. How can we limit the risk of unpredictable solutions resulting from the unsupervised deep learning of AI (if and how it is decided that deep learning should be unsupervised), and what instruments are needed for this? How can we combine the need to get the best results in fighting crime using AI tools (e.g. facial recognition) with the clear need to avoid unregulated mass surveillance? How should we avoid the malicious use of AI solutions and control the asymmetry of power and information while using new procedures to manage people with the support of algorithms, ML systems and AI? And finally: how can we make the human-centric and principle-based approach fully operational? Going beyond this, we need to create tools that can ensure adherence to these principles.

It is very interesting to observe how established companies are seriously and responsibly involved in finding the best solutions. Among others, Google[19] has started to prepare some responses to societal expectations of the development of AI systems. The company has proposed several instruments to

[18] Ibid., 26–31.

[19] J. Dean and K. Walker, 'Responsible AI: Putting Our Principles into Action', Google blog, 28 June 2019.

provide for explainability, auditability, interpretability, international standards, testing and validation, contestability and user feedback. This list is very useful and broad, and offers many ways in which to work on making AI systems much more transparent. However, one may have doubts about some of Google's motives. It is therefore in the interest of Google's shareholders' (and those of all similar companies) to ensure the credibility and trustworthiness of these motives. This is necessary to bring together the perspectives of shareholders and stakeholders (clients, citizens and business contractors). A transparent framework can help to achieve this. This would offer a new way of understanding relations between the company and its environment. As usual in the market economy, shareholders are focused on profits. However, with the digital transformation, they are increasingly acknowledging their social responsibilities. At the same time, stakeholders are not simply passive consumers. They are active in shaping the relationship between business and users. The more they are aware, the more they can support the development of personalised services and products, to the benefit of the shareholders. Hence transparency should be linked to the efficiency and convenience outcomes that benefit both providers and users. Transparency is also necessary for trustworthy AI development.

Box 2 Responsible AI practices

Like many companies, Google has started to work on its Responsible AI strategy.[20] Since announcing its AI principles in 2018, the company has initiated:

- The education and empowerment of employees for a deeper understanding of AI functions and, critically, consideration of how to use AI responsibly.
- Trainings on ML fairness, aimed at employees.
- Invitations to AI ethics speakers to share their thoughts and ideas with both Google's global workforce and other audiences.
- The development of a framework for responsible and ethical AI that benefits everyone.

[20] Ibid.; Gennai, 'Google's AI Principles' (unpublished presentation seen by the author).

- The establishment of a People and AI Research Group to share people's and scientists' views on issues.
- Research papers (and their publication) on topical AI issues and the dimensions of AI, crucial for better understanding this new phenomenon and avoiding mistakes.
- The development, preparation and making open-sourced of 12 new tools, for instance, the What-If Tool, which allows users to analyse an ML model without writing code. 'It enables users to visualise biases and the effects of various fairness constraints as well as to compare performance across multiple models'.[21] The Google Translate tool has also been developed, focusing on reducing gender bias by delivering feminine and masculine translations for users—without the data ever leaving their device, further enhancing user privacy.
- The engagement of product teams to contribute to building human-centric AI products.
- The possibility of businesses and organisations using the Cloud AI Hub to access already-trained ML models.
- An assessment of the potential of the AI Hub (as used outside the company) for harmful dual use, abuse or presentation of misleading information.
- The establishment of a much more efficient text-to-speech network that not only means that the system only has to be trained once, but that requires much less data and time to adapt to new speakers. This network 'could help individuals with voice disabilities, ALS, or tracheotomies',[22] and also recognises the potential of such technologies to be used for harmful applications (such as synthesising an individual's voice for deceptive purposes).

[21] Dean and Walker, *Responsible AI.*

[22] Ibid.

Trustworthiness is fundamental to the sustainable development of AI. It should be human centric, providing humans with control over the technology, thus overcoming the information asymmetry between users and AI developers and deployers. This would fulfil the need to make the human-centric and principle-based approach fully operational and ensure the transparency of AI solutions.

I will now focus more closely on four issues: explainability, data control, risk assessment and cybersecurity.

Explainability

For transparency and trust, the concept of explainability is fundamental. There is a very strong link between the ability to explain and the ability to make accurate predictions in AI, as was shown in the CEPS report.[23] For example, deep learning models based on neural nets offer the highest prediction accuracy, but the lowest level of explainability. Bayesian belief nets can bring a higher level of explainability and mid-level prediction accuracy. Decision trees offer many more possibilities for explainability, but with a lower level of prediction accuracy.[24] It is necessary to know more about the complicated models algorithms use: 'For example, frequently-used algorithms are based on so-called "Neural Networks", which work with hidden layers of relationships and combinations of all different characteristics in the data. This makes it difficult to assess whether or not a person is being discriminated against on grounds of their gender, ethnic origin, religious belief or other grounds'.[25]

What does this mean?

Thus, the framework for making some solutions explainable not only depends on the will of the regulators, but much more heavily on the substance and nature of the learning techniques used. And this leads to the advancement of AI. Simple models require fewer sources of data and less complicated

[23] Renda, *Artificial Intelligence*, 60.

[24] Some new possibilities for using modelling techniques were presented in Accenture Consulting, *Model Behavior. Nothing Artificial. Emerging Trends in the Validation of Machine Learning and Artificial Intelligence Models* (2017), 7.

[25] EU Agency for Fundamental Rights, *#Big Data: Discrimination in Data – Supported Decision-Making* (2018), 6.

analytical procedures, while more advanced ones need a wide variety and huge scale of data and more advanced analytical schemes.

When working on practical models of explainability we need to know what can be explained, and what cannot and why not. In addition, it should be made clear that explainability does not mean full access to everything (i.e. all code, as has been suggested by some groups in the public discourse on the subject), because many technical solutions and software programmes are subject to intellectual property protections. By avoiding non-transparent schemes, possible harm to users and hidden functions (the 'black-box effect') we should be able to find a balanced model for public decisions and decisions that have a significant impact on humans.

One possible way to solve the problem is to create adequate levels of explainability by:

- *Addressing the users.* This would give them all the key information needed to ensure proper understanding. In reference to consumer rights, it would offer a reasonable and as easy way as possible to contest AI decisions and to ensure effective redress against decisions made by AI or the humans operating it. It should also focus on the collective redress rules and be orientated to solve the liability problem as was analysed and noted in the Commission document, *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics.*[26] It is crucial to take into account various types of redress rules and mechanisms for different types of device and technological solution. In supporting consumer rights' development, the collation of clear and understandable redress rules is key to making the system more transparent.
- *Addressing the auditors.* First of all, an auditing system needs to be established that centres on standardised rules, norms and certification schemes, with the documentation from the companies and authors of solutions made fully available (this is crucial for the full evaluation of potential high-risk applications, and is the background for the risk-management model proposed by the European Commission[27]). It should not only reference the technical aspects of the AI, but also the evaluation and testing models used to complete the conformity assessment (e.g. to ensure full respect of

[26] European Commission, *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics*, Report, COM (2020) 64 final, 19 February 2020.

[27] European Commission, *White Paper on Artificial Intelligence*, parts C, D, E and F.

EU law). In addition, taking into account cybersecurity principles, it is important to find a proper balance between opening all information up for auditing (including legally binding trade secret provisions) and limiting access for security reasons to the designers and administrators of critical infrastructure, for example, the energy network.

- *Addressing the investigators.* The investigators should know as much as possible (i.e. have access to key data and information, as well as technical premises and projects) about all aspects of the functioning of the AI systems, because they will have to assess the effects of incidents and harm cases—both physical and mental.
- *Addressing the researchers.* One could consider giving investigators access to the data and algorithms. Such access would enable an independent assessment of the effectiveness and side-effects of the AI (content control and profiling), and the potential for harmful dual use, abuse or presentation of misleading information. This should certainly be correlated with some aspects of the conformity assessment processes for AI systems.

Proper implementation of a solution for explainability still requires many analyses and evaluations—and is a field for common work among all partners: business, academia, institutions and regulators. The final proposal may include references to both legislation (i.e. the explainability obligations related to products and services based on AI, and the model of voluntary labelling for non-high-risk AI applications proposed by the European Commission[28]) and 'soft law' (e.g. codes of conduct, co-regulations and self-regulation in the different sectors).

Data control and data development

The model of data control is one of the most important factors supporting trustworthy AI. On the one hand, it requires the full and harmonised implementation of the GDPR. This is necessary to complement AI

[28] Ibid., part G.

development in many areas: among others, health care using AI models; smart cities that combine the use of all kinds of data, both personal and non-personal; and public services improvement using transparent data processing. We also need to avoid the misinterpretations which currently exist in this field.

On the other hand, data control requires raising awareness among all users. We, as individuals, need to be more sensitive in relation to our data, as our control has a significant impact on privacy protection and cybersecurity—for which some kind of cyber-hygiene is needed. The best illustration of data control is when I am responsible for my data, for all my choices and behaviours without redundant burdens, and I know that my data in some situations will be further processed. I should also understand that this processing must follow clear rules for data anonymisation and pseudonymisation (with adequate safeguards).[29]

In light of AI development, it is also vital to have a proper framework for data flows in Europe and all over the world. Without the right conditions for data flows and data sharing models, it is not possible to train algorithms, ML or AI, or to make them more efficient. We have some rules in the EU and agreements with third countries such as Japan and the US, and soon with South Korea. However, this is not sufficient and further work is needed on forging clear and transparent contractual arrangements rather than blocking access to global data. If we want to use AI properly, we need to be able to gather sufficient data and sometimes the European market may not provide enough. But above all, the EU must fully use the diversity and the richness of its data and properly understand that AI's strength is built on this.

In addition, there is another aspect of data that is crucial for AI development. It is not only the availability of data that is important, but also their quality—and while there is a clear obligation to protect personal data, in some cases, anonymised data are not entirely useful for processing and producing highly qualified outcomes. If we want to change the European economy and use the potential of data for economic and social development on a major scale (including to grow small and medium-sized

[29] The significance of anonymisation and pseudonymisation stems from GDPR interpretations, facilitating the avoidance of breaches related to personal data and indicating the applicability of data processing for many purposes. There is also a meaningful debate on the legitimacy of the use of personal data. We should also note the mandatory obligation to carry out data protection impact assessments as outlined in art. 35 of the GDPR. However, anonymised or pseudonymised data cannot be used for every purpose—there are some limits to the efficiency of AI.

enterprises (SMEs) and to ensure internationalisation), it is obvious that we have to create a 'hierarchy' among the different types of data, in order to structure them and prepare them for common use.

In Box 3, we see how Metro AG, a German food company has used algorithmic, AI and ML techniques for data processing to make the business much more efficient.

Box 3 Metro AG: using intelligence to simplify customers' lives

German wholesale and food specialist company Metro AG (a business-to-business company) decided to utilise all the digital possibilities of the Google Cloud Platform to build a new model of relations with its customers.

The company's technology unit, Metronom (employing approximately 2,000 members of staff) prepared and is now leading implementation of the following solutions:

- Making the company's and customers' data analysis more accessible across the whole business.
- Creating a new data lake (built with the Google Cloud Platform) to store information in an unlimited way.
- Making it possible to integrate advanced analytics solutions and ML for super-precise observation of all items and actions to better support services for customers.
- Strengthening analytical power (real-time reporting of data streamed direct from stores and applications) for scaling up services, as emphasised by Marko Schwob, Domain Owner Analytical Platform Engineering at Metronom:[30] 'scaling is about more than storage space, it is about having analytical power available on demand'.

[30] *Google*, 'Metro. Stocking Technology and Business Intelligence to Simplify Customers' Life'.

- Supporting internal product teams. There are more than one hundred data analytics workbenches, each one tailored to a certain function, with easy access to all computing possibilities. As a result the company provides clear advice, for instance on hygiene laws for certain foods including how to store foods such as ice cream, which involves integrating information from local weather forecasts.
- Creating a new business model for all partners based on more accurate buying decisions, thus saving money (by reducing costs).
- Developing a permanent process for optimising analytical tools, achieving omnichannel analytics (the ability to trace a customer's total experience of all interactions with Metro).

The above-mentioned solutions have led METRO to reduce the instability of its e-commerce platform by up to 80%, scaling capacity to match a 45-times increase in daily events in just three months from starting to use the data lake and reducing infrastructure costs by between 30% and 50%. In other words, the company has been able to successfully and radically improve its customers' lives.

In terms of upcoming developments, a solution is also required for the problems of data portability (art. 20 of the GDPR) and openness to industry-specific solutions. Additionally, the problem of data ownership should be solved, taking into account the sectoral basis for specific data-access principles. In the context of autonomous car development and Internet of Things (IoT) opportunities, there are significant issues that need to be addressed, such as who owns the data, what conditions are needed to share the data, what kind of functioning open data sources can be treated as a common good and in what direction development of open data collection should move.

In parallel to the debate on the *White Paper on AI*, the EU is discussing key aspects of the Data Strategy.[31] The main challenge is making all actors in the marketplace (public institutions and the private sector) understand the importance of data for future economic development and the improvement of the social life and public services. In this context, managing the following key aspects is essential:

[31] European Commission, *A European Strategy for Data.*

imbalances in market power, data interoperability and quality, data governance, data infrastructures and technology, and data literacy. Ensuring the existence of proper rules, infrastructure, interoperability and a cross-sectoral governance framework would go some way to solving these shortcomings and make data processing a European competitive advantage.

There is a need to develop the proposed concept of a 'High Impact Project on European data spaces and federated cloud infrastructures'.[32] It is clear that there are many interconnections and interdependencies between data infrastructure development and stronger incentives for AI systems development. As a first step, it is crucial for European industries to develop opportunities related to the optimisation and prediction models used by businesses and smart public institutions. This can be achieved by ensuring data accessibility and AI support. Developers of AI need to meet the collective (industrial and institutional) deployers of AI. This process requires energy-efficient and trustworthy edge and cloud infrastructures or, as the European Commission describes them, 'Infrastructure-as-a-Service, Platform-as-a-Service, Software-as-a-Service services'.[33]

In 2018, the EU established the European Open Science Cloud as a space for the exchange of data in the scientific area. Formally it was a success, but the technological requirements have not been so easy to implement. The real problem is a deficit in the culture of sharing. No patterns and behaviours of sharing have been developed—we are far from the spirit and model of a sharing economy.

Thinking about future advantages, the European Commission has proposed the establishment of additional Common European Data Spaces.[34] This is a much needed and ambitious goal. But first, we should analyse all possible barriers including member states' competences, difficulties in exchanging data because of the various interests of partners (among them competitiveness expectations), the lack of harmonised solutions and fragmentation of existing models, the lack of interoperable schemes, and cultural and mental habits. The barriers in these areas may prove very difficult to overcome.

[32] Ibid., 16.

[33] Ibid.

[34] These include an industrial (manufacturing) data space, a Green Deal data space, a mobility data space, a health data space, a finance data space, an energy data space, an agriculture data space, a skills data space and data spaces for public administration. Ibid., 22.

For all the proposed data spaces, we have to ensure an adequate level of data control. Hence, The ability to implement a data control model is closely related to the explainability solution, and it does not matter whether this is for personal or non-personal data, individual or collective subjects. The most important elements are the creation of a joint model of data flows rules (globally and in the EU), ensuring data accessibility, the development of a culture of data sharing and the establishment of clear instruments for users to build the capacity to control data.

Risk assessment

The tangible integration of the principles of trustworthy AI is closely linked to the assessment of risk. The German Data Ethics Commission gave a very inspirational presentation on some practical suggestions relating to the 'criticality of an algorithmic system'.[35] The full integration of ethical principles into AI systems requires proper risk analysis and a proper response to those risks via risk-management options. One possible way to ensure adequate risk analysis is to assess threats based on the likelihood and severity of the potential harm. We need to know how decision-making processes or components of the algorithmic processes could affect fundamental and consumer rights (and to what extent) and lead to harm. The German Data Ethics Commission has produced a 'Criticality Pyramid and Risk-Adapted Regulatory System'[36] which establishes five levels of threats:

- applications with zero or negligible potential for harm,
- applications with some potential for harm,
- applications with a regular or significant potential for harm,
- applications with a serious potential for harm, and
- applications with an untenable potential for harm.

[35] Germany, Data Ethics Commission, *Opinion of the Data Ethics Commission*, part 3: 'Algorithmic systems'.
[36] Ibid., part 3, Figure 2.

The German Data Ethics Commission also makes some suggestions on the range of measures which could be applied. These include:

- the lack of a need for special measures;
- meeting formal and substantive requirements (e.g. the publication of a risk assessment, monitoring of procedures, *ex post* controls);
- additional measures, such as *ex ante* approval procedures;
- a live interface for permanent oversight led by supervisory authorities (which leads to the question of which EU institutions have the best capacity to run this?); and
- the total or partial ban of an algorithmic system.

What is critical for the future of the risk-assessment model is to find common agreement on this or similar proposals (e.g. the European Commission's recommendations on risk assessment). How can the adequacy of these solutions (in technical, legal and practical terms) be verified? Which European institutions/agencies should take responsibility for the preparation of this model, and its implementation and management (i.e. the role of a supervisory institution or network of institutions, and under what framework)? How can the risk-assessment model as a process be created in such a way that it is open to technological changes, sensitive to new phenomena in relations between humans and machines, and adaptable to innovations?

The proposed system could act as the foundation for further analysis and development. But it seems to me that the use of risk-analysis and risk-management solutions could be very helpful in establishing concrete instruments to maintain and develop the principle-based approach to AI development.

In the European Commission proposal, the risk-based approach, crucial to establishing the proper legislative and institutional framework, is linked to the two important dimensions of threats and harms. These relate first to the risks to fundamental rights, and second to the risks to safety and the effective functioning of the liability regime in a broader sense, including cybersecurity issues.[37] All these risks

[37] European Commission, *White Paper on Artificial Intelligence*, part A: 'Risks for safety and the effective functioning of the liability regime'.

can affect individuals (legal breaches and physical incidents), businesses (legal uncertainty) and public institutions with their goal of making public services more effective through the use of AI. As a result, the Commission proposes that: ‘AI application should generally be considered high-risk in the light of what is at stake, considering whether both the sector and the intended use involve significant risks’.[38] The relevant sectors include health care, transport, energy and various parts of the public sector, such as asylum and migration services, border control, the judiciary, and social security and employment services. As the Commission proposes, any list of sectors in the regulations should be exhaustive, with the obligation to be periodically reviewed.

In my opinion, this can only be the starting point (the early stage of establishing the real set of rules) and we need to take into consideration the factors and rules on which a real assessment of additional sectoral indications should be based. Could we imagine that autonomous decision-making processes might lead us to the destruction of a building, because the data used for decision-making had come from fake information? There is a significant problem with data quality and credibility. The more we need data, the more we need to ensure they are proper. In the time of disinformation, fake news and fake science, we can also imagine fake construction data. Hence we need completely new procedures for data verification, especially for public data.

According to the Commission, the second criterion focuses on the possibility of high-level risks resulting from AI solutions having an enormous impact on affected parties: ‘material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities’.[39] Thus when establishing types of high-risk applications, we should leave space for exceptional instances and phenomena, for example, those related to discriminatory threats during recruitment processes or to the use of remote biometric identification.

Even if I agree with some of these concrete concepts, we are still in the early phase of reflection on the classification of the various types of risks related to the functioning of AI systems. Moreover, more clarification is needed as to whether the first and second criteria should function separately or be linked.

[38] Ibid., part C: ‘Scope of the future EU regulatory framework’.

[39] Ibid., part C.

To better adjust the risk-assessment system to the different types of threats and to create a framework for turning threats into risks, the Commission has produced a precise[40] taxonomy of the requirements needed for the adequate evaluation of AI systems and to create a background for conformity assessments. Analysing potential high-risk AI applications, the following key features should be taken into consideration:

- training data (sufficient for the qualified design of applications, avoiding dangerous situations);
- data and record-keeping (key for storing data and if necessary tracing data back for the verification of algorithm sequences);
- the information to be provided (all information applying to the users and consumers; this is also about explainability);
- robustness and accuracy (analysis of the accuracy of prepared solutions, the reproducibility of outcomes, and the proper reaction of AI systems to errors and inconsistencies during the lifecycle of products/services);
- human oversight (full model of the validation of solutions from the human perspective—the foundation for oversight); and
- specific requirements for certain AI applications, such as those used for the purposes of remote biometric identification (returning to the assessments needed in the conflict between security expectations and individual freedom).

I consider this taxonomy a well-designed basis for work on future solutions.

Finding an adequate model of risk assessment (adjusted for the various types of threats), which enables AI development and does not block innovative solutions, is essential. It requires clear rules for conformity assessment and the creation of an institutional network to ensure the effectiveness of all the necessary administrative processes and decisions.

[40] Ibid., part D: ‘Types of requirements’.

Cybersecurity

As a complement to the principle-based approach, I want to emphasise the significance of cybersecurity issues. These also relate to the safety-first approach presented by the European Commission as the backbone for risk assessment, described above.

A feeling of certainty is fundamental and critical for users. To avoid threats which might arise due to a lack of proper care being taken regarding the security of the infrastructure needed for AI systems development (i.e. the network of High Performance Computing Centres, clouds and networks, such as the 5G network), this certainty should be delivered and institutionally guaranteed. Safeguards are needed to establish a resilient AI environment, safe from all kinds of attacks. The problem of components coming from various parts of the world (i.e. third countries) could be a particular cause for concern (as has been seen regarding the risks of developing the 5G network). The proper response should be ensured through existing EU legislation (i.e. the NIS Directive, the Cybersecurity Act and further implementation of the EU model of certifications), but also by building up a general understanding of the need for strategic autonomy, as is often noted by Paul Timmers.[41] Eventually, it would be useful to generate an additional security-based approach as a key factor in the safe development of AI.

Sustainable AI development and a resilient AI environment ought to be guaranteed by the proper implementation of all EU regulations and—in addition—by a growing awareness of the significance of cybersecurity issues for the strategic autonomy of AI development in the EU. Considering all the challenges described in this section, an impact assessment is of supreme importance. Therefore, we should provide one by analysing all aspects of the cybersecurity of AI (algorithmic systems and ML) and the possibility of implementing the human-centric and principle-based approach. We should also consider the various risk-assessment models and the data-control-based approach, alongside the implementation of the relevant architecture. Finally, an analysis of the future-proof approach is

[41] P. Timmers, 'Ethics of AI and Cybersecurity When Sovereignty Is at Stake', *Minds and Machines*, 29 (2019), 635–45.

necessary. The public consultations opened by the European Commission on the White Paper on AI resulted in a strong requirement to deliver an impact assessment of the AI and Data Strategies.

Governance by law; governance by oversight (institutional options)

There is an ongoing debate in the EU on the need for a regulatory framework for AI development and the tools that will be useful for the better governance of this new challenge.

Assessment of the existing legal framework

First, we need to assess the existing legal framework in the EU and establish whether or not it is sufficient for the future safe development of AI. Currently, we can base future aspects of the regulatory framework on:

- GDPR, where there is a need to ensure full harmonisation and the common interpretation of multiple articles and existing provisions (this should be complemented by finalising the work on the ePrivacy Directive). It is essential to remember that Article 35 of the GDPR has made data protection impact assessments mandatory for all processing related to new technologies which could put rights and freedoms at a high level of risk. We also need to remember the mission to disseminate the impact of GDPR—making this regulation the global reference point for modelling privacy protection.[42]
- The Law Enforcement Directive.[43]

[42] European Parliament and Council Regulation (EU) no. 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L119 (27 April 2016), 1.

[43] European Parliament and Council Directive (EU) 2016/680 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ L119 (27 April 2016), 89.

- The Regulation on the free flow of non-personal data.[44]

- The Open Data Directive.[45]

- The Cybersecurity Act and the NIS Directive, which require full implementation, including the introduction of certification schemes, the promotion of cybersecurity risk-assessment and management patterns (we need to review the cybersecurity aspects for the future development of AI), and the promotion of cyber-hygiene educational efforts.[46]

- Guidelines for the implementation of 5G networks, ensuring compliance with cybersecurity requirements (as detailed by the Commission in several documents).[47]

- The HLEG on AI's *Ethics Guidelines for Trustworthy AI* and the incoming results of further analysis and recommendations currently being discussed by the group.

- All solutions that address the requirements for various types of devices, from medical to those used in farming development or in the IoT, with consideration of software requirements. Some such solutions exist, while some require review.

- An adequate review of the Product Liability Directive,[48] which should be linked with the specific challenges generated by AI (algorithms, ML models, automated decision-making processes, and the AI tasks used in some concrete services such as medical or financial). It will be important to find proper understandings of and the division between mental and physical harms, and it is crucial for business-to-customer operations that a proper assessment is made of the necessary

[44] European Parliament and Council Regulation (EU) no. 2018/1807 on a framework for the free flow of non-personal data in the European Union, OJ L303 (14 November 2018), 59.

[45] European Parliament and Council Directive (EU) 2019/1024 on open data and the re-use of public sector information, OJ L172 (20 June 2019), 56.

[46] European Parliament and Council Regulation (EU) no. 2019/881 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) no. 526/2013 (Cybersecurity Act), OJ L151 (17 April 2019), 15; European Parliament and Council Directive (EU) 2016/1148 concerning measures for a high common level of security of network and information systems across the Union, OJ L194 (6 July 2016), 1.

[47] NIS Cooperation Group, *Cybersecurity of 5G Networks, EU Toolbox of Risk Mitigating Measures* (January 2020); European Commission, *Secure 5G Deployment in the EU – Implementing the EU Toolbox*, Communication, COM (2020) 50 final, 29 January 2020.

[48] European Council Directive 85/374/EEC on the approximation of the laws, regulations and administrative provisions of the member states concerning liability for defective products, OJ L210 (25 July 1985), 29.

compensations (i.e. what kind and how). It is necessary to consider suggestions of co-liability schemes (e.g. it is not clear who would take financial responsibility for autonomous car accidents when the reasons may not relate simply to the AI or computer systems, but partly also to the motor or to tyre issues).

- An analysis of whether current EU consumer law is fit for the practical implementation of the rules of AI development (Directives on Unfair Commercial Practices, Unfair Contract Terms, Consumer Rights, Sales of Consumer Goods, Price Indication and the above-mentioned Product Liability).[49]
- Results of the ongoing debates on digital platforms (their responsibilities, rules and the problem of their accountability), which may also affect discussions on the Digital Services Act and the modernisation of the eCommerce directive.

The above-mentioned package will be very useful for further policymaking and shows that we have something to build on and are not starting from zero. Progress requires, however, a thorough analysis of the sufficiency and insufficiency of the existing legal framework before we can take the next decide on a way forward. From the perspective of the existing legal framework there may be shortcomings which could undermine the adequate functioning of AI systems in the future. If the requirements of transparency are not established, it will be difficult to identify legal breaches and to find clear solutions for liability.

Since stand-alone software cannot be defined as either a product or service (especially a service based on AI systems), it is difficult to apply EU safety legislation. This also concerns respect for the rules regarding the lifecycle of products and services, for instance, in situations where software should be updated after it has been on the market for a certain period of time.

[49] European Parliament and Council Directive 2005/29/EC concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) no. 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive'), OJ L149 (11 May 2005), 22; European Council Directive 93/13/EEC on unfair terms in consumer contracts, OJ L95 (5 April 1993), 29; European Parliament and Council Directive 2011/83/EU on consumer rights, amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and repealing Council Directive 85/577/EEC and Directive 97/7/EC of the European Parliament and of the Council Text with EEA relevance, OJ L304 (25 October 2011), 64; European Parliament and Council Directive 1999/44/EC on certain aspects of the sale of consumer goods and associated guarantees, OJ L171 (25 May 1999), 12; and European Parliament and Council Directive 98/6/EC on consumer protection in the indication of the prices of products offered to consumers, OJ L80 (16 February 1998), 27.

Some risks are not currently being explicitly addressed (in particular regarding some applications related to the development of the IoT supported by AI) because there is a problem with when and how to assess them: when the product/service enters the market, during future updates of the software or, in the case of the results of self-learning, during the use of the AI itself. If the value chain of a solution supported by AI has many components (e.g. products placed on the market by a party that is not a producer), the proper allocation of responsibilities will prove difficult because very often EU and national product liability rules are incompatible.

A review and assessment of the existing laws related to the various dimensions of the functioning of AI is necessary. Once this has been delivered, we will be able to establish the most desirable components of a future-proof regulatory framework.

New regulatory framework

We should consider how to make the regulatory framework more adaptable, flexible and open to the disruption of AI development. For this we can use ‘some experimental approaches to regulation’, such as ‘regulatory sandboxes’ (ongoing monitoring of the market and the social impacts of innovative projects), the ‘incorporation of technology roadmaps’ (through the use of multistakeholder platforms as inputs into the policymaking processes), and ‘ongoing monitoring’ of all policy impacts, which could lead to the use of sunset clauses in the regulations.

What is particularly important from my point of view is the new way of understanding the need for rules and regulations. Before even considering a new regulatory framework, we must tackle some pertinent issues:

- Provide rules that are supportive of and not limiting to AI development in Europe.
- Establish relevant regulations that solve problems instead of imposing bans and restrictions, and are fully implementable.

- Consider how to guarantee flexibility and openness to changes and adaptations resulting from innovation.
- Differentiate between strict general legislative rules ('hard law'), rules addressed to specific sectors and rules implementable by 'soft law' (codes of conduct, co-regulation and self-regulation models) in terms of responsibility for the new AI ecosystem.
- Involve all partners (without creating artificial borders between strictly European and external businesses) to find the best model of regulatory framework.
- Build a new 'architecture for data control' using existing laws and by creating practical solutions to avoid redundant burdens with reference to the explainability model.
- Develop skills, competences and attitudes for risk assessment and management, and promote not only professional, but also public awareness of the principled and human-centric issues crucial for AI development.
- Make the HLEG on AI recommendations operational, ensuring respect for ethical principles and fulfilment of joint data accessibility, transparency and security expectations.
- Consider how to translate rules and regulations that address AI systems in a way which supports business and economic development, especially the growth of SMEs (with a new model for business-to-business interactions and transactions).

According to Renda there are three areas that will be crucial to a regulatory framework for AI development:

- *Fundamental principles*. That AI development is lawful. This considers the importance of proper legislative solutions, taking care to meet all the principles mentioned by the HLEG on AI in Ethics *Guidelines for Trustworthy AI*.
- *Good practices*. That AI development is responsible, which means (a) that companies or public institutions that use the AI models take transparent responsibility with clear references to the key principles; (b) that development is open to allow the exchange of experiences in a model of best practice as per codes of conduct; (c) that there is confirmation by all partners of the importance of

some procedures (e.g. in health care) as an expression of cooperation between patients, clinicians and managers of AI systems; and (d) that experiences of governance patterns, monitoring, control and feedback—all non-mandatory opportunities for making AI human-centric—are shared.

- *Sustainable AI*. That development of AI is sustainable, with reference to the broad context in which AI use is growing, especially to the UN's Sustainable Development Goals as a collection of objectives important for the UN and the EU's 2030 Agenda. These include inclusion, full and productive employment, innovation, zero carbon growth, quality of education and women's empowerment, among others.

It is obvious that the final shape of the new regulatory framework should be the result of a merit-based compromise between AI development needs, political considerations regarding the future advantages of trustworthy AI and a proper understanding of the purpose of legislative efforts. All this should build European competitive advantage and the right conditions for strategic autonomy.

Governance challenges

I agree with these considerations. However, from the perspective of creating an effective policymaking model, we also need to consider the second dimension of governance (in addition to governance by law): governance by oversight.

A certain set of solutions is needed to make the AI development principles fully implementable and workable. These include the explainability model, risk assessments, data control schemes and architecture, as well as solutions ensuring the protection of fundamental and consumer rights. These require legislative solutions and a genuine legal framework which fully encompasses the participation of various partners. Undoubtedly, all these elements of building a trustworthy AI system should be founded in a robust environment that ensures the best forms of management and oversight. How should this necessary and expected institutional oversight of AI development be established and ensure a trustworthy model of growth?

If the key challenges stem from autonomous decision-making processes and data governance issues (such as those impacting data collection from various sources and of different types, and the mechanisms of data use, reuse, processing and sharing, all of which have the potential to harm people), it is clear that we need different models and levels of oversight and, probably, the involvement of many kinds of institutions. All of them should be focused on effectively implemented human-centric AI.

From an institutional point of view, there are several existing European agencies and bodies which should be involved in the work on AI development and ready to support the European Commission, European Parliament and the Council:

- the European Data Protection Board, which is responsible for data protection, and explanation and interpretation of the GDPR;
- the EU Agency for Fundamental Rights, which is responsible for ensuring fundamental rights in all areas of European activity, monitoring various phenomena and producing reports;
- the HLEG on AI, which is focused on providing an understanding of ethical principles and transferring them into practical requirements and procedures (through suggestions for legislative solutions and the proposal of 'soft law' models);
- the European Network Information Security Agency, which responds to cybersecurity threats, harmonising the efforts of all member states, and is working on certification schemes for products, services and all processes requiring cybersecurity resilience frameworks;
- the European Consumer Consultative Group, which references many analyses and recommendations crucial to broad consumer protection in the EU, in cooperation with national bodies responsible for consumer issues; and
- those institutions and organisations responsible for the certification, standardisation and establishment of norms important to the proper step-by-step development of AI.

And there are clearly many more that contribute. However, if we wanted to point to an institution that could lead and prepare a risk assessment for the different types of AI measures, algorithmic systems and the effects of ML activities, it would be challenging to find one ready to take on this responsibility. There are also many social organisations (whether business, science or citizen orientated) equally involved in AI-related decision-making policy processes and solutions. Therefore, the focus on institutional support for the AI ecosystem (governance by law and by oversight) should be complemented by the full involvement of all partners in building a system of excellence.

What is critical is to build a new network model among all these institutions, agencies and bodies and to establish the necessary rules to enable them to cooperate and indicate their priorities. This is the only way to lead the ongoing work on rules and regulations that ensures the participation of all key stakeholders at all levels. There are many good examples of dialogues between partners/stakeholders which aim to find a proper model for data use and functioning AI that benefits us all.

Box 4 Philips: healthcare opportunities for all through cooperation with all partners

While working on the use of AI as a key support for better diagnosis and therapies in health care services, Philips recognised the problems of ‘overtrust’ (a too-optimistic attitude among some people implementing AI functions) and ‘undertrust’ (a too-pessimistic attitude that ignored the potential benefits and exaggerated the threats and risks of AI). Both problems are very much present at the moment, across societies. In health care, the main fear is of the replacement of doctors by machines, which could lead to a model without human factors and values in the relationship between patient and medical service.

For Philips it is clear that trust should be balanced and commensurate with the evidence provided in the validation of each AI tool. The best way to respond to these societal challenges is to involve all stakeholders at an appropriate level and at the most appropriate time to exchange views on future products, devices and services, that is, before they are fully implemented into the health care system. This means the *ex ante* assessment of the ethical context for future offerings, including the patients’ and doctors’ perspectives, and the *ex ante* validation of the functioning of AI-enabled medical products and their efficiency and effectiveness—within the established regulatory frameworks for all medical products.

Why is Philips' approach exemplary?

Philips is open to involving all key partners in concrete discussions on AI in health care. This means that while engineers and AI operators are invited, it is primarily patients (representative organisations), clinicians, consultants and scientists, including ethics specialists, who are consulted. Philips has actually started such dialogues through blog posts on their published 'Data and AI Principles' website.[50]

The process of monitoring and assessing pilot projects and experiments works in parallel with education and skills development, enabling a better understanding of the new solutions by all participating partners. At the same time, the platform encourages not only debate and education, but the sharing of views on the risks and scientific knowledge of what some products and services are for—if it is safe to do so and with full respect for fundamental and patients' rights (especially when the topic relates to data).

During this dialogue and interaction, which also highlights the need for decentralised oversight of new uses of AI in medical services, a model of practical transparency aimed at all users (patients and doctors) is implemented and has become the most persuasive tool for disseminating ideas about new health care possibilities. This is the best way to build trust and credibility. It is vital for Philips to acknowledge publicly that even autonomous AI processing models require human oversight in order to avoid biases and discrimination, and to ensure fairness. This innovative and state-of-the-art approach by Philips enables the proposed solutions to be fully validated by both science and society.

The case study of Philips in Box 4 offers an encouraging example of a decentralised model of cooperation, allowing validation by science and society. The alternative—the establishment of a new European institution responsible for all dimensions of AI—would not, in my view, be feasible.

[50] https://www.philips.com/a-w/research/blog.html

Since the European Commission's publication of the *White Paper on AI*, it has been clear that we are not on the way to creating a dedicated agency responsible for AI governance. This means that a framework for cooperation between competent bodies and authorities and the creation of a network are the expected solutions to the matter of oversight.

These would require, however, rules of cooperation for all partners and a clear message to the member states that they should be involved by using (or establishing, in many cases) national authorities focused on all aspects of AI development. It is crucial to avoid fragmentation; it is also essential to look at capacity building at all levels, from sectoral, via national to the European. How should this be done? How should the EU testing centres, which will play a fundamental role in supporting AI systems development with clear requirements and no redundant burdens, be allocated and spread out across the Union?

The proper institutions should be equipped (financially and technically, as well as with skilled people) with all the competences needed to prepare and provide conformity assessments in order to test new solutions and support the *ex ante* model of assessment, leading to certification schemes. But we have to ask: to what extent should conformity assessment focus on the near-daily updating of algorithms (required either for security reasons or for the convenience of users)? National authorities, as well as sectoral ones, should focus on these issues, and make adequate use of the capacity of existing structures (such as those in finance, pharmaceuticals, aviation, medical devices, consumer protection and data protection) rather than duplicating current functions.

In my view, there is no precise description of the relations between all the institutional partners in this collaborative framework, especially as it will require proper and proportional relations between the national and European levels, as well as the sectoral level (there are several new areas which face the potential development of high-risk solutions for which no institution has the capacity to take responsibility, e.g. social security, workforce development). A European governance structure could be established as a forum for the exchange of opinions and best practices, offering advice on standardisation activities and certification models, and also supporting the voluntary labelling of non-high-risk applications.

The Commission's proposals on governance by law and (from my perspective) by oversight should be developed and discussed institutionally, as well as with social partners and all stakeholders in AI development. This is the starting point. We also have to consider the concerns related to this model of governance. What kind of criteria should be used to ensure good governance, avoiding bureaucratic burdens, slow decision-making processes and the ill-preparedness of governing bodies? The need for such a networking model of governance for some aspects of AI development is due to the assessment that currently no institution working alone could provide an expected or adequate response to the growing challenges.

Democratic oversight

In order to build a networking model of governance and trustworthy AI, it is essential and critical to ensure proper democratic oversight. AI development could interfere with fundamental and consumer rights and undermine social order (e.g. through biases caused by a decision-making process that lacks proper supervision of the data sources and that takes an easy-going attitude towards unintended consequences). It is our responsibility to avoid such a situation arising. Hence, it is important to create an oversight network which cares about democratic principles and should be based on the following visible pillars/aspects:

- awareness of the network partners, with educational efforts aimed at all generations;
- evidence-based knowledge of multidimensional AI development that avoids stereotypes and naive prejudices;
- tangible and just representation of all groups that have roles in society and the economy: citizens, workers, consumers, business leaders (also SMEs), AI developers, AI deployers, scientists, regulators and policymakers;
- an openness towards future-proof solutions tailored to sector-oriented models; and
- a readiness to cooperate, with an understanding of the significance of the added value of possible results.

At the same time, we should consider how to involve member states in this networking model of governance and how to make it effective. With such an approach, governance by law can be complemented by governance by oversight.

A regulatory framework is important and crucial for the governance of AI development, yet alone it is not sufficient. AI will develop fast, the use of AI opportunities will increase, new horizons for AI functionalities will open up and we cannot predict the new possibilities arising from AI-related matters. Hence we also need to govern AI development through democratic oversight by a well-designed network of cooperative partners: from governments and public institutions through to science and businesses, and citizens, workers and consumers.

Policy recommendations

We are starting a new chapter in the debate on AI development in light of the discussions on the Data Strategy, the new model for the Digital Single Market (perhaps even a new paradigm), the need for an industrial policy 4.0 and the passing of the Digital Services Act. All of the above mentioned are necessary and key to the future of European competitive advantage.

Moreover, there are two additional issues which should be taken into account. The first is an assessment of whether a common approach can be achieved and under what conditions. This is a matter of convincing both sides of the political spectrum that an agreement and balanced solutions are necessary. The political mindset should be shifted: political battles do not matter if they do not lead us to compromise, which is important for AI development based on a human-centric approach, and in parallel allow us to build new opportunities for the European economy, crucial for strategic autonomy.

Second, the experience of COVID-19 should make us more aware of new, previously inconceivable possibilities for AI. COVID-19 has also caused a rapid growth in the use and uptake of digital solutions; it has been a genuine digital game-changer.

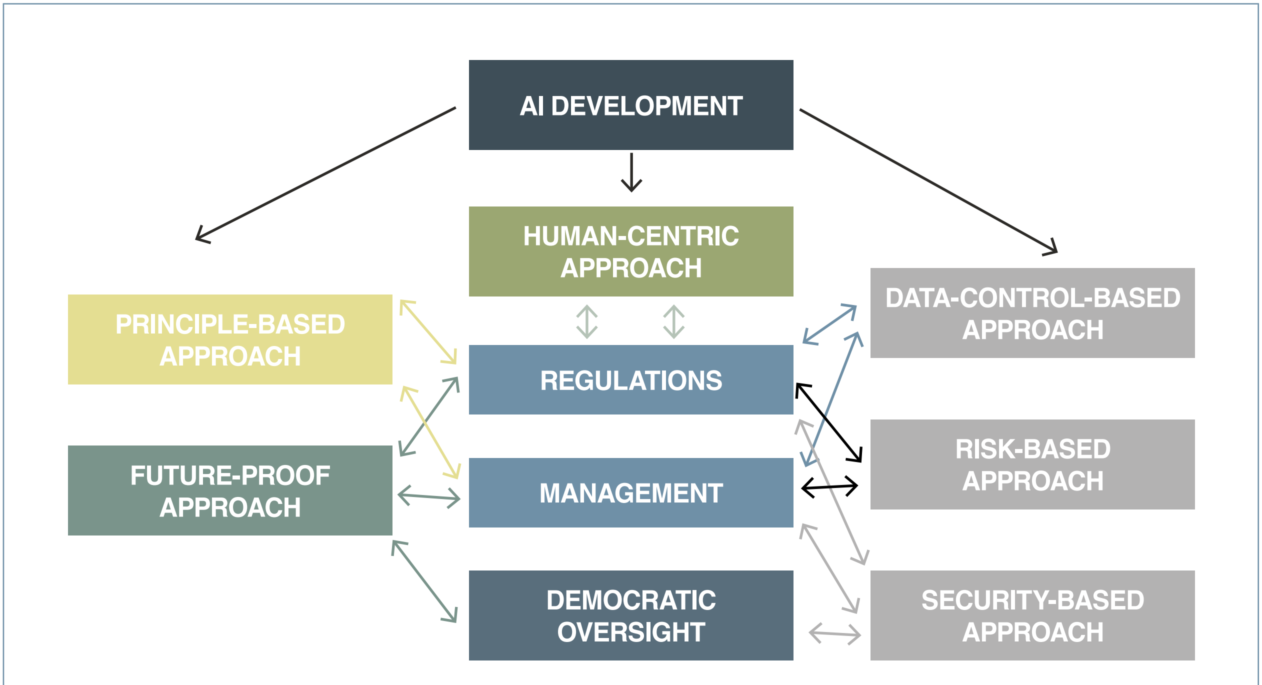
It is obvious that at this stage of the debate on the future of AI and digital issues in the EU, we need to take a holistic view and develop a broad understanding of many aspects of digital development. We also need to embrace a two-step approach to AI development. The economic dimension and aspirations of the European digital advantage must be linked with societal, cultural, political and even psychological views of the future. Our development, as humans and societies, relies on this combined approach and we must take this holistic view to find the proper direction for our opportunities.

There are several recommendations, which in my mind, should be reflected in this debate:

1. It is important to visualise and practically implement the multidimensional model of basic objectives and conditions necessary to create the best future regulatory framework, management patterns and oversight of AI systems development, as is presented below.

   Figure 1 offers a very good example of the complementary model of governance by law and governance by oversight. In my opinion, a holistic view of AI systems development and the multidimensional model presented would be useful in helping to overcome tensions in political debates on the subject and would establish an adequate balance between 'hard' and 'soft' law, strict regulations and the flexibility to be future-proof, security and privacy, legislation and democratic oversight, and AI development for economic growth and the human-centric approach. Once we know what kinds of dilemmas exist, we will be able to solve them.

Figure 1 AI development



Source: Author's own formulation.

2. When looking for a basis for upcoming regulations, we should first assess existing law. This means evaluating its sufficiency or lack of it. This should lead to the necessary corrections and a greater focus on the proper implementation of multiple solutions (such as GDPR, the Cybersecurity Act and broadening the scope of consumer rights). Only after full reviews of the existing legislation

should we start to deliberate on what new legislation is needed and why, taking into account that all solutions should be future-proof and, with due consideration for flexibility, allow for the adjustment of legal frameworks to recognise and allow innovation. All proposed legislative solutions should be monitored and impact assessed. From a political point of view, the issue of how to define and reflect flexibility in the rules is and will be meaningful. The only way forward is to assess the usefulness of the concept of flexibility during the impact assessment. I am confident that the current experience of COVID-19 and the significant role of AI in the fight against it will create more arguments in favour of flexibility, albeit ones which will not undermine key ethical principles.

3. There are two conditions requiring different policy decisions that crucially need to be met for AI development to occur in line with the regulatory framework. These are the implementation of a standardisation model (schemes harmonised all over the EU, open to global requirements) and the development of the 5G infrastructure (a key tool, not least for the opportunities offered by the IoT), with all the European safeguards that would make the networks and AI systems much more resilient. Common work on standardisation patterns and a safe 5G infrastructure is indispensable at the European level and should involve all member states. From a political perspective, it is clear that the experience of COVID-19 will speed up work on establishing an adequate environment for AI development and induce stronger, faster, more effective and much more harmonised cooperation among all member states.

4. If AI systems development is multidimensional and involves many reference points (from fundamental rights, via expectations concerning technical understandability and the common interpretability of some mechanisms, to consumer rights) it is clear that the regulatory framework should be built on risk assessment. There is no one-size-fits-all solution. The model of risk analysis (with levels of risk ranging from highly risky to non-risky) and options for risk management need to be described precisely and take into account an ongoing assessment of threats. This is a real challenge because of the lack of adequate and competent institutions ready to conduct this assessment. It is also—to some extent—problematic because of the unpredictability of certain behaviours and thus the results of autonomous decision-making processes. This unpredictability may result in unintended consequences, an issue which also needs addressing.

A correlation between levels of risks and concrete options to remedy them is crucial to the design of a new regulatory framework which has strict requirements and voluntary solutions. What is more, it requires a clear decision as to whether we base risk analysis on sectoral challenges or on the potential for harm and the negative impact of concrete applications, as the Commission has suggested. The risk-management approach needs proper adequacy rules, a clear model for responsibilities, institutions with the capacity to make conformity assessments and a focus on adaptability to innovation—so far, this approach is in the very early stage of preparation in the EU.

5. In order to avoid hazardous and unpredictable cases, we should consider a new model of assessment of possible harmful impacts on humans (such as biases or discriminatory solutions). We should move from the current model of *ex post* assessment to *ex ante* assessment, or establish a combination of *ex ante* models (crucial for preventing harm) and *ex post* schemes (supportive of innovation). This will require pilot projects, tests and a change in business models. As a result, there is a high probability that some inventions may experience delays in commercialisation. Nevertheless, these changes are necessary as they are the way to build trust between users and technology. The debate about AI as a crucial tool in supporting the fight against COVID-19 could pave the way to testing the capability and efficiency of the *ex ante* model (used to assess possible harmful impacts in the early stages of designing AI solutions). Additionally, there is the new idea of establishing an 'ethical and Technology Assessment' prior to the deployment of AI systems.

6. A key issue in AI development is ensuring efficient data accessibility, transparency, security and—foremost—governance under the clear rules of a human-centric approach, utilising a new architecture of data control. Such rules should be based on a proper, redefined implementation of the GDPR and on increased user awareness and readiness to make individual choices, giving consent after being informed. All of this needs to be combined with openness in data sharing, providing the rules of data sharing are transparent and the culture of data sharing supported by public policies (currently, there is a lack of a data-sharing culture). All of these aspects require political agreement. The redefined implementation of the GDPR (with a harmonised interpretation of certain provisions in all countries of the EU), rational cooperation to finalise the

work on the ePrivacy Directive, promotion of the new data-control architecture and dissemination of the importance of awareness of individual decision-making (through consent or cyber-hygiene behaviours) are essential for AI development. At the same time, these issues need to be tackled by various political groups as they have been circulating in political discourse for many years, evoking tensions rather than leading to compromise solutions.

7. We should discuss and decide how to prepare data for AI processing/computing. This requires work (1) with industries to provide sector-orientated solutions and to make industrial data better adjusted for AI processes; (2) on all kinds of medical data, including on how to use them in a proper way and with greater efficiency for better diagnosis and therapies; and (3) on public data, to make them much more open to common use as a common good. Finally, a serious debate needs to take place in the EU about limits on the data used for AI development, or rather on taking an open view of the various kinds of data sources (i.e. the availability of data without borders) under the rules of GDPR and other laws affecting data flows. Moreover, the challenge of COVID-19 has led to greater recognition of the need for data. Collecting, sharing, maintaining, processing and reusing data requires a fast and effective response during a pandemic. In addition, there is a much greater understanding of the importance of open data as a common good in the current circumstances. In this context, the experience of COVID-19 should accelerate our response to building an appropriate data-use framework and should reconcile political perceptions.

8. If we want to build trust, we need to do more than adhere to the principles of fundamental rights—this will not be enough to ensure trustworthy AI development. We must create tools to support these principles (i.e. to uphold them) and ensure that the rules of explainability are clear. The four levels and objectives of explainability aimed at users, auditors, investigators and researchers could help to indicate adequate, realistic and clearly expressed obligations for AI developers. This will also require an adequate auditing system.

It is clear that AI development should be closely linked to ethics and principles. What is more, trustworthy AI must be built on the transparency of all processes to avoid any threats or harms arising from it. Fundamental rights are the backbone of trustworthy AI. However, to appropriately enforce them, they need to be associated with consumer rights, and have clear, easy and understandable redress mechanisms guaranteed in law. Compensation models and liability schemes need to be prepared and clearly adjusted to the AI context, as is now being discussed in the European Parliament.

9. Many aspects of AI development and many phases in the value chain of AI functionalities will need to be assessed or will require ongoing supervision. For this, the appropriate institutional context is missing. At present there is a lack of proper capacity to take responsibility for safe AI development. I suggest establishing a networking model of governance (with rules of cooperation between institutions and partners, including social, business and scientific representatives) functioning at all possible levels (from companies via sectors and national institutions to the European level) to ensure fully democratic oversight (the real participation and influence of all stakeholders). We should consider the experience of COVID-19, both in terms of data use and practical governance, as a pilot study for the full launch of AI. Hence, we ought to conclude our debate on the AI White Paper and Data Strategy urgently, achieving material and political agreement. This agreement should enable democratic oversight tools to be exercised, focused on AI development and the development of its functionalities. Such oversight should likewise be in place for the use of contact-tracing apps.

10. The future of AI development needs a proper environment. An acute element of building this environment relates to the aspect of AI awareness, that is, the ability to understand the changing world and technologies, and people's readiness to adapt. The key to better understanding the phenomenon of AI is digital AI literacy. This will require a massive educational effort aimed at all generations and focused not only on skills, but also on attitudes and competences. This is the only way to overcome threats and fears and to provide people with a feeling of certainty and stability in the fast-moving new world. We not only have to be aware of threats, but also of our rights and expectations. We should be able to live in a world in which we do not lose control over technology but instead make AI development human-centric.

Bibliography

Accenture Consulting, *Model Behavior. Nothing Artificial. Emerging Trends in the Validation of Machine Learning and Artificial Intelligence Models* (2017).

Bartlett, J., *The People vs. Tech: How the Internet Is Killing Democracy (and How We Save It)* (London: Penguin Random House, 2018).

Bertelsmann Stiftung, *Automating Society. Taking Stock of Automated Decion-Making in the EU* (January 2019).

Bruegel, *Europe Needs a DARPA* (February 2020).

Bruegel, *The Dynamics of Data Accumulation* (February 2020).

Bughin, J., et al., *Tackling Europe's Gap in Digital and AI*, Mckinsey (February 2019), accessed at https://www.mckinsey.com/featured-insights/artifical-intelligence/tackling-europan-gap-in-digital-and-ai/.

Centre for European Policy Studies, *Artificial Intelligence and Cyber Security* (January 2020).

Chivot, E., 'Is the EU's AI Policy Headed in the Right Direction?', Center for Data Innovation (15 July 2020), accessed at https://www.datainnovation.org/2020/07/is-the-eus-ai-policy-headed-in-the-right-direction/.

Commission Expert Group on FAIR Data, *Final Report and Action Plan: Turning Fair Into Reality* (2018), accessed at https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf.

Cornillie, C., 'Finding Artificial Intelligence Money in the Fiscal 2020 Budget', *BGov.com*, 28 March 2018, accessed at https://about.bgov.com/news/finding-artificial-intelligence-money-fiscal-2020-budget/.

Dean, J. and Walker, K., 'Responsible AI: Putting Our Principles into Action', Google blog, 28 June 2019, accessed at https://www.blog.google/technology/ai/responsible-ai-principles.

Digital Europe, *Digital Europe's Recommendations on Artificial Intelligence Policy* (13 November 2019), accessed at www.digitaleurope.org.

EU Agency for Fundamental Rights, *#Big Data: Discrimination in Data – Supported Decision-Making* (2018).

European Commission, *A European Strategy for Data*, Communication, COM (2020) 66 final, 19 February 2020.

European Commission, *Artificial Intelligence for Europe*, Communication, COM (2018) 237 final, 25 April 2018.

European Commission, *Report on the Safety and Liability Implications of Artificial Intelligence, Internet of Things and Robotics*, Report, COM (2020) 64 final, 19 February 2020.

European Commission, *Secure 5G Deployment in the EU – Implementing the EU Toolbox*, Communication, COM (2020) 50 final, 29 January 2020.

European Commission, *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*, White Paper, COM (2020) 65 final, 19 February 2020.

European Commission Joint Research Centre, *Artificial Intelligence. A European Perspective* (2018).

European Council Directive 85/374/EEC on the approximation of the laws, regulations and administrative provisions of the member states concerning liability for defective products, OJ L210 (25 July 1985), 29.

European Council Directive 93/13/EEC on unfair terms in consumer contracts, OJ L95 (5 April 1993), 29. European Council on Foreign Relations, *Machine Politics: Europe and the AI Revolution* (July 2019).

European Parliament and Council Directive 98/6/EC on consumer protection in the indication of the prices of products offered to consumers, OJ L80 (16 February 1998), 27.

European Parliament and Council Directive 1999/44/EC on certain aspects of the sale of consumer goods and associated guarantees, OJ L171 (25 May 1999), 12.

European Parliament and Council Directive 2005/29/EC concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/

EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) no. 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive'), OJ L149 (11 May 2005), 22.

European Parliament and Council Directive 2011/83/EU on consumer rights, amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and repealing Council Directive 85/577/EEC and Directive 97/7/EC of the European Parliament and of the Council Text with EEA relevance, OJ L304 (25 October 2011), 64.

European Parliament and Council Directive (EU) 2016/680 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ L119 (27 April 2016), 89.

European Parliament and Council Directive (EU) 2016/1148 concerning measures for a high common level of security of network and information systems across the Union, OJ L194 (6 July 2016), 1.

European Parliament and Council Directive (EU) 2019/1024 on open data and the re-use of public sector information, OJ L172 (20 June 2019), 56.

European Parliament and Council Regulation (EU) no. 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L119 (27 April 2016), 1.

European Parliament and Council Regulation (EU) no. 2018/1807 on a framework for the free flow of non-personal data in the European Union, OJ L303 (14 November 2018), 59.

European Parliament and Council Regulation (EU) no. 2019/881 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) no. 526/2013 (Cybersecurity Act), OJ L151 (17 April 2019), 15. European Parliamentary Research Service, *A Governance Framework for Algorithmic Accountability and Transparency* (October 2018).

European Parliamentary Research Service, *Automated Tackling of Disinformation* (December 2018).

European Parliamentary Research Service, *Cost of Non-Europe in Robotics and Artificial Intelligence* (June 2019).

European Parliamentary Research Service, *The Ethics of Artificial Intelligence: Issues and Initiatives* (March 2020).

European Political Strategy Centre, *The Future of Work? Work of the Future! On How Artificial Intelligence, Robotics and Automation Are Transforming Jobs and the Economy in Europe* (May 2019).

Frey, C.-B., *The Technology Trap. Capital, Labour and Power in the Age of Automation* (Princeton: Princeton University Press, 2019).

Friends of Europe, *The Case for a Global AI Framework* (November 2019).

Germany, Data Ethics Commission, *Opinion of the Data Ethics Commission*, Berlin (October 2019), accessed at www.datenethikkommission.de.

Google, 'Metro: Stocking Technology and Business Intelligence to Simplify Customers' Lives', accessed at https://cloud.google.com/customers/metro.

Google, *Perspectives on Issues in AI Governance* (2018), accessed at https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

Google, *Responsible Development of AI* (2018), accessed at https://ai.google/static/documents/responsible-development-of-ai.pdf.

Hao, K., 'Yes, China Is Probably Outspending the US in AI – But Not on Defense', *Technology Review*, 5 December 2019, accessed at https://www.technologyreview.com/2019/12/05/65019/china-us-ai-military-spending/.

HLEG on AI, *Ethics Guidelines for Trustworthy AI* (8 April 2019), accessed at https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

Jacques Delores Institute, *Establishing Trust in an AI-Powered Future* (November 2019).

Malinina, J., *AI Must Be Smart About Our Health*, BEUC, BEUC-X-2019-078-02/12/2019 (December 2019).

Martin, D., *AI Rights for Consumers*, BEUC, BEUC-X-2019-063-23/10/2019 (October 2019).

NIS Cooperation Group, *Cybersecurity of 5G Networks, EU Toolbox of Risk Mitigating Measures* (January 2020).

Pichai, S., 'AI and Google: Our Principles', Google blog, 7 June 2018, accessed at https://www.blog.google/technology/ai/ai-principles/.

Renda, A., *Artificial Intelligence. Ethics, Governance and Policy Challenges*, CEPS (Brussels, February 2019).

Schneider, S., *Artificial You. AI and the Future of Your Mind* (Princeton: Princeton University Press, 2019).

Schomon, C., *Automated Decision-Making and Artificial Intelligence – A Consumer Perspective*, BEUC, ref. BEUC-X-2018-058-20/06/2018 (June 2018).

Timmers, P., 'Ethics of AI and Cybersecurity When Sovereignty Is at Stake', *Minds and Machines*, 29 (2019), 635–45, doi:10.1007/s11023-019-09508-4.

Artificial Intelligence (AI) is changing our world. This new phenomenon carries many threats, but also offers many opportunities. We need to find a suitable framework to support trustworthy AI. A key challenge remains: can we, as humans, retain control over the technology or will the technology take control of humanity? In responding to this challenge, the following question needs to be considered: What kinds of tools are needed, not only to keep control of AI development, but foremost to multiply the possible opportunities it offers?

It is our responsibility to urgently establish an adequate framework for the development of AI systems based on a revision of the existing law and followed by possible new legislative proposals with a clear focus on future-proof tools. We have to generate a suitable governance model that not only has its foundation in law, but that also ensures democratic oversight through the open and collaborative participation of all partners and the validation of AI solutions by science and society. We should build trustworthy AI based on a human-centric and principled approach. The practical implementation of ethical rules in the design of AI and the evaluation of the everyday functioning of AI systems are essential.

Wilfried
Martens Centre
for European Studies