Revista CIDOB d'Afers Internacionals
ISSN: 1133-6595
ISSN: 2013-035X
publicaciones@cidob.org
Barcelona Centre for International Affairs
España

Antonov, Alexander
Managing complexity: the EU's contribution to artificial intelligence governance
Revista CIDOB d'Afers Internacionals, núm. 131, 2022, Mayo-Septiembre, pp. 41-68
Barcelona Centre for International Affairs
España

DOI: https://doi.org/doi.org/10.24241/rcai.2022.131.2.41/en

Disponible en: https://www.redalyc.org/articulo.oa?id=695774517021

Revista CIDOB d'Afers Internacionals
n.º 131, p. 41-65
ISSN:1133-6595 | E-ISSN:2013-035X
DOI: doi.org/10.24241/rcai.2022.131.2.41/en

Reception date: 25.11.21
Acceptance date: 01.04.22

Managing complexity: the EU's contribution to artificial intelligence governance

Gestionar la complejidad: la contribución de la UE a la gobernanza de la inteligencia artificial

Alexander Antonov
Doctoral Candidate, Department of Law, School of Business and Governance, TalTech-Tallinn University of Technology. *Alexander.Antonov@taltech.ee*.
ORCID: https://orcid.org/0000-0001-6692-647X

How to cite this article: Antonov, Alexander. "Managing complexity: the EU's contribution to artificial intelligence governance". Revista CIDOB d'Afers Internacionals, issue 131 (September 2022), p. 41-65. DOI: doi.org/10.24241/rcai.2022.131.2.41/en

Abstract: With digital ecosystems being questioned around the world, this paper examines the EU's role in and contribution to the emerging concept of artificial intelligence (AI) governance. Seen by the EU as the key ingredient for innovation, the adoption of AI systems has altered our understanding of governance. Framing AI as an autonomous digital technology embedded in social structures, this paper argues that EU citizens' trust in AI can be increased if the innovation it entails is grounded in a fundamental rights-based approach. This is assessed based on the work of the High-Level Expert Group on AI (which has developed a framework for trustworthy AI) and the European Commission's recently approved proposal for an Artificial Intelligence Act (taking a risk-based approach).

Key words: European Union (EU), artificial intelligence (AI), governance, fundamental rights, AI Act, trustworthy AI, digital single market

Resumen*: En un contexto de ecosistemas digitales mundialmente cuestionados, este artículo examina el papel y la contribución de la UE al concepto emergente de la gobernanza de la inteligencia artificial (IA). Entendida esta por la UE como el ingrediente fundamental para la innovación, la adopción de sistemas de IA ha alterado nuestra comprensión de la gobernanza. Enmarcando la IA como una tecnología digital autónoma integrada en las estructuras sociales, este artículo argumenta que se puede aumentar la confianza de la ciudadanía de la UE hacia la IA si la innovación que esta comporta se fundamenta en un enfoque basado en los derechos fundamentales. Ello se evalúa a partir del trabajo del Grupo de Expertos de Alto Nivel en IA (que ha desarrollado el marco para una IA fiable) y la propuesta recién aprobada de la Comisión Europea para una Ley de inteligencia artificial (con un enfoque basado en el riesgo).*

Palabras clave*: Unión Europea (UE), inteligencia artificial (IA), gobernanza, derechos fundamentales, Ley de IA, IA fiable, mercado único digital*

In the context of globally contested, dynamically evolving digital ecosystems, with trade and production of goods, services and information incrementally shifting into the digital realm, the objective of this conceptual, exploratory paper is to examine the EU's role in and contribution to development of a novel type of governance, the phenomenon of an emerging Artificial Intelligence systems (AI) governance framework. With AI's key function of amplifying if not automating social processes traditionally carried out by human beings, its wide-scale introduction into society would conceivably have a revolutionary impact on human autonomy. While questions abound as to the exact nature of AI and thus its scope, and the ability to regulate its application in diverse societal contexts, the EU devised a regulatory framework tailored to leverage for the potential of AI, as one of the leaders of AI development, along with the US and China.

The EU's globally unique standards on AI are of particular interest: a) the Trustworthy AI framework, developed by the High-Level Expert Group on AI (AI HLEG, 2019a) and based on a fundamental rights-based understanding, and b) the subsequent European Commission's proposal for a four-dimensional risk-based approach in the Artificial Intelligence Act (AIA)[1] (EU Commission, 2018a; 2018b; 2021d). The EU's AI policies are derived from the intention to create both an «ecosystem of trust»[2] and «ecosystem of excellence»[3] (EU Commission, 2020). More broadly, they are underpinned by the two-fold rationale to both accelerate development of the Digital Single Market and empower citizens and consumers alongside the transformative goal of achieving *Europe's Digital Decade* (EU Commission, 2021a). But how are these two approaches, the fundamental rights-based and innovation-inspired risk-based method, aligned with a view to achieving «Trustworthy AI» in this context?

As such, this study seeks to examine the EU's modes of governance applied in this nascent policy domain through the prism of the AI HLEG's framework of Trustworthy AI, and in turn elaborate on whether they are fully grounded in fundamental rights-based understanding, since this approach arguably presents the most viable mode of increasing citizens' trust in an increasingly autonomous

1. See: COM (2021) 206 final. «Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts».
2. An approach based in applied ethics which is aimed at developing AI under a reflective, citizen-centric, fundamental rights-based rationale.
3. It primarily gravitates around three pillars: responsible investment, innovation, and implementation of AI; See also: COM (2021) 205 final. «Fostering a European approach to Artificial Intelligence».

data-driven technology (AI HLEG, 2019a). Consequently, the key objective is providing initial insights into the extent to which the seven requirements of Trustworthy AI – human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability (AI HLEG, 2019a) – are expressed in the risk-based approach of the AIA, by the following research questions:

– What is AI and how can the technology be conceptualized?
– What is the EU's AI governance framework and its contribution to the emerging field of AI governance in general?
– To what extent is the proposed four-dimensional risk-based approach in the AIA aligned with the Trustworthy AI concept?

While AI is dual use in character, the scope of application of the Trustworthy AI framework is confined to development and deployment of AI in the public sector (EU Commission, 2021b). Additionally, the study is mainly centred on external governance factors, thereby not providing an analysis of variable internal governance factors such as the impact of budgetary matters, e.g. regarding public procurement of AI, on the AI governance framework.

Problematizing AI

Why AI presents a challenge to governance

Characterized as a general-purpose technology (Brynjolfsson and McAfee, 2017) and conceived as a «black-box», AI's double-edged sword character provides compelling reasons for a regulatory regime, comprehensive, holistic and multi-layered in nature. Touted as one of the most strategic technologies of the 21st century, a wide-scale uptake of AI encapsulates the promise to increase the quality of products and services, raise efficiency and create economic growth amounting to €176.6 billion if not trillions annually, provided its adoption in e-commerce as part of the European Digital Single Market Strategy proves successful within the next few years (Scott *et al.*, 2019). On a socio-political level, it promises to help address societal challenges in domains such as *health care* - by detecting cancer cells, in *agriculture* - by decreasing depletion of soil, or in *transportation* - by increasing safety and potentially reducing the carbon footprint (Taeihagh, 2021).

Tensions between social and economic dimensions of AI arise when societal trust towards the use of the technology both by private actors and civil servants in the public sector weakens. This is primarily due to ill-considered deployment or even deliberate abuse of AI, as already witnessed in domains such as in law enforcement in the US context of predicting the likelihood of recidivism, in the Chinese context of using remote biometric identification and social credit systems or, more pertinently for this paper, in the allocation of welfare benefits in the EU (Chiusi *et al.*, 2020). These and other cases present challenges for wide-scale societal adoption of AI and call for regulatory measures to restrain what has been sometimes conceived in scholarly accounts as the emerging «digital leviathan» (Langford, 2020).

To harness AI, the social and legal discourse in the still fragmented study of AI governance therefore centred primarily on questions of how to achieve «responsible AI», an AI «ethical by design» and «Trustworthy AI» (Van den Hoven, 2017; Theodorou and Dignum, 2020; Hamulák, 2018). With engineers being one of the key stakeholder groups in the development of AI, research initiatives such as *Z-inspection* have been established with the aim of involving AI developers in iterative co-design frameworks and engaging them in discussions with a diverse group of domain specific experts (Zicari *et al.*, 2021). This holistic, interdisciplinary, co-design methodology adds an important element to concretization of the requirements for Trustworthy AI of the AI HLEG, increases awareness of the socio-technicity of AI and thus reaffirms the importance of a fundamental rights-based approach to AI governance within the EU. Additionally, debates exist, centred around creating liability and accountability frameworks in cases of potentially discriminating or flawed AI (Ebers, 2021). In an ever more networked, datafied society, structured by Information and Communication Technologies[4], the human agency and hence human dignity, democracy and the rule of law, pillars and values of the European integration project as such, are fundamentally put to the test by uptake of AI (AI HLEG, 2019a; Murray, 2020).

The EU faces the challenge of establishing a regulatory framework for the design, development and application of AI which does not «unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans» but instead «augments, complements and empowers human cognitive, social and cultural skills».

4. Understood to be a «broad and unconsolidated domain (…) of (i) products, (ii) infrastructure and (iii) processes (…) that includes telecommunications and information technologies, from (a) radios and (b) telephone lines to (c) satellites, (d) computers and (e) the Internet» (ITU, 2015).

The two-fold challenge as such, resonating with the previous discourse on emerging technologies (Larsson, 2021), is to devise a tailored, proportionate regulatory mechanism which on the one hand provides a degree of latitude for innovation in AI and on the other hand addresses pitfalls already detected. In other words, the EU faces the challenge of establishing a regulatory framework for the design, development and application of AI which does not «unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans» but instead «augments, complements and empowers human cognitive, social and cultural skills» (AI HLEG, 2019a).

As contended, a fundamental rights-based approach helps to situate and contextualize wide-scale application of AI-based socio-technical systems, aligned with principles of proportionality and necessity. It provides access to redress and accountability mechanisms in adverse cases involving AI, in particular to vulnerable groups in society. Additionally, from the viewpoint of the AI developer, the fundamental rights framework helps anticipate and thus address potential risks arising from deployment of AI (Smuha, 2021) at an early stage. The question concerns how we may achieve Trustworthy AI in the context of variable endogenous and exogenous factors, not to mention the global competition for AI development, epitomized in notions such as «AI race» between the US, China and the EU, and growing calls for «digital sovereignty» (Pohle and Thiel, 2020).

Conceptual framework of AI systems

AI as an autonomous digital technology embedded in societal structures

AI can be divided into different methods and sub-disciplines (Gasser and Almeida, 2017), the most promising of which is comprised of machine learning (ML) based applications e.g. in the areas of natural language processing, image recognition or robotics. After inception of AI as a research discipline amid the Dartmouth Workshop in the summer 1956[5], the current political and economic

5. The Dartmouth Summer Research Project took place in the summer of 1956 at the private university Dartmouth College (Hanover, New Hampshire). It ran for roughly eight weeks and was organised by John McCarthy (Computer expert), Marvin Minsky (scientist), Nathaniel Rochester (chief architect of the IBM) and Claude Shannon (mathematician, electrical engineer and cryptographist). This event is widely considered to be the founding event of artificial intelligence as a field.

interest in AI was preceded by various ups and downs, or so-called «AI winters» and «AI summers» (Russel and Norvig, 2010). Catalysed by an exponential increase in the amount of machine-readable data, coupled with acceleration of computational power afforded by improved statistical ML methods, there seems to be a new momentum for widescale uptake of AI both within public administration and the private sector.

The intangible nature of continuously self-learning AI, which runs on software and code, or digitized information encoded into bit strings designed to translate electronic impulses to achieve a certain goal by either amplifying or even replacing a human being's mental or physical capacity, complicates our understanding of how to devise a human-centric regulatory framework. Defined in the AIA as «software that is developed with (…) (a) Machine learning approaches (…), (b) Logic- and knowledge-based approaches (…), (c) Statistical approaches (…) and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with» (EU Commission, 2021c), AI can be best conceptualized «as a medium that is materialized into particular code-based devices» (Lawson, 2017) implicating a change in the mental state of a human being and/or a physical state of objects.

It follows, that the notion and understanding of the digital technology of AI hinges on the context of its application, its impact on the material world and the underlying means or methods by which certain pre-programmed goals are expected to be achieved. Its materialization always translates into an effect on the real world and is thus bound to actualization of its designer's pre-programmed processes. As such, elements of human intelligence and knowledge are replicated and represented in the technological ability to *perceive* the digital and/or physical environment, *interpret and process* structured or unstructured data, *decide* to take the most rational action in the context of attaining a pre-defined goal, *learn* from this process and *inductively* establish new rules to attain the same goal(s) more efficiently (AI HLEG, 2019a). Consequently, AI can be conceptualized as a technological and socio-technical system as soon as the threshold of its effects materializing into actualization through material devices or artefacts, computer-based and/or robotic devices has been reached. In this vein, it is worth recalling the relevance of data, since the process of harnessing data translates into a reflection of the values, norms and societal structures in which AI is deployed, if not embedded in (Rahwan, 2018; Larsson, 2019; Larsson 2021).

While this framework could be contested on account of being too broadly applicable, rendering previous software-based ICTs «AI systems», the novelty of AI is best reflected in its inherent feature of autonomy, afforded by ML-based

self-learning algorithms and its as yet still narrow, but gradually developing intelligence. More broadly, AI is unique in comparison to traditional socio-technical systems owing to its 'autonomous, adaptive, and interactive' features (Van de Poel, 2020; Troitiño, 2021), with the design changing continuously though interaction and engagement with the environment across time and space. The nexus between AI's nature and deployment in real-world contexts thus necessitates additional reflection on the impact of this novel type of socio-technical system on the human environment, and its interactions with human beings in particular. Framing AI as an autonomous socio-technical system embedded in a socio-legal, economic environment thus helps establish the link between AI design and the impact of design choices for AI development on human-computer interaction. It increases the role of citizens empowering design approaches, which can extend to human needs from a human-centric perspective in the context of AI innovation. This is even more relevant if one considers, that innovation in AI has been primarily taking place in private, commercially driven research clusters, and the rationale of these was tilted rather in favour of consumers than citizen empowerment, with a key impact on AI design and value choices therein (Umbrello, 2022).

The novelty of AI is best reflected in its inherent feature of autonomy, afforded by ML-based self-learning algorithms and its as yet still narrow, but gradually developing intelligence.

Hence, to concretize the AI HLEG Trustworthy AI framework and thus leverage the potential of AI's autonomous features, engineers in particular are called on to familiarize themselves with and apply systems thinking approaches to AI design, which are based on fundamental rights-based rationales. To this end, e.g. Umbrello (2022) suggests utilizing the Value Sensitive Design (VSD) approach, a framework providing a rich toolkit on the study of computer-human interaction, and when adapted to the specificities of AI, to AI governance in particular (Umbrello and van de Poel, 2021). As a reflective, interdisciplinary cross-cultural specific method, it can elicit and foster awareness of the importance of design choices in AI innovation, and of the long-term impact of embedding context-specific values in AI on citizens, end-users and other stakeholders in emerging AI digital ecosystems. This is of relevance, since VSD can complement the work of the AI HLEG by means of translating the Trustworthy AI criteria through AI design into specific norms, and may therefore help promote a fundamental rights-based system thinking for AI governance.

The second difficulty is best described by the so-called «AI effect», a paradox underlying deployment of all AI based technologies: As soon as AI has been adopted by the broader public, it loses the character of AI and becomes a conventional

technology (Troitiño, 2021). However, defining AI based on the severity and scale of effects of its deployment on human-beings and the environment in general, as stipulated in the AIA in the form of the four-dimensional risk-based approach, helps us address this challenge not only on the semantic level but conceivably on the level of the rule of law and fundamental rights (EU Commission, 2021c). For example, an AI system «which deploys subliminal techniques beyond a person's consciousness to materially distort a person's behaviour» (EU Commission, 2021c) could never be treated as a conventional technology in and by a democratic society. Tensions could arise during application of «real-time remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement» (ibid.), with exceptions to its application facing substantial criticism by the European Parliament, including various European NGOs.

Nevertheless, provided that the same applications are restricted by the fundamental rights law principles of necessity and proportionality, implies that even these will not be approved without any critical reflections by a democratic society informed by an independent press and thus aware of the potential dangers AI poses to its own dignity. Problems might arise in EU countries where the rule of law, and independence of the press and judiciary is gradually being subverted. This might complicate judicial review procedures to contest individual applications of AI in administrative courts e.g. by law enforcement which might in turn have a negative effect on trustworthiness of the digital technology.

Additionally, further awareness regarding use of biometrics is required in its deployment against citizens from «third countries», e.g. in the context of border protection. The field of biometrics will therefore remain contested and provide fertile ground for debates as to whether applications from thence should be transferred wholly or partially to the «Prohibited AI Practices» criteria (EU Commission, 2021c). Additional grey areas set the basis for tensions between human rights principles and safeguarding public security temporarily infringing those rights could be mentioned, but what it boils down to is that as long civil society or legal entities, are informed about AI's pitfalls and empowered to challenge them through judicial review procedures, adverse impacts of AI would continuously be questioned by society rendering the «AI effect» a pertinent criteria upon which AI deployment could be assessed. Consequently, remaining under public scrutiny for the entire life cycle of AI, the paradox presents a threshold criterion which the democratic character of society can be assessed against.

In essence, since AI can be construed as an autonomous digital technological artefact embedded in a socio-legal, economic environment, regulating AI resolves around the central questions of «by *whom* and *for which purpose* [AI systems] will be designed and related to that *by whom* they are owned and deployed and *in which contexts* they will be applied» (Antonov and Kerikmäe, 2020).

Steering AI: from governance to AI governance

The notion of «governance» applies to various domains and fields, broadly denoting (a) complex type(s) of steering and co-management by public and private actors over social, political and/or economic processes, either on an international, national or sub-national level. Generally speaking, the framework of governance may therefore be deciphered as a social ordering exercise, with the following elements factored into the definition: (i) multiplicity of actors - institutions, states, international and non-governmental organizations, (ii) variety of mechanisms and (iii) structures, (iv) degrees of institutionalization and (v) distribution of authority (Katzenbach and Ulbricht, 2019).

While some political scientists have questioned the analytical depth of the framework, owing to its wide scope and applicability (Kohler-Koch and Rittberger, 2006), in the uptake of disruptive technologies and the call for a holistic understanding of AI's societal ramifications (Murray, 2020), the governance framework first of all provides guidance as a toolbox for exploring and potentially addressing the complexity around AI, the diversity of actors and processes around the technological infrastructure, and in turn modalities and configurations in the inter-relationship between our existing understanding of the framework of governance and AI, thus the necessary measures, whether they are policies, legislation, regulations, including alternative modalities of regulation (Lessig, 1999) - such as ethics guidelines (Larsson, 2021), standards, codes of conduct -, or adjudication, through which social, political and economic frictions caused by the uptake of AI may be attenuated and hence public and private interests proportionately balanced out.

In essence, the concept originated at the end of the 1970s, when private actors entered the field of public governance driven by the motives of cost-reduction and efficiency, areas and domains over which the State traditionally assumed political authority, eventually leading to what Rhodes (1996: 661) described as a 'fragmentation of political authority'.[6] On the international and external dimension of the governance framework, amid the end of the Cold War and spurred on by the process of globalization, state-centric thinking had gradually been replaced by new types of governance culminating in the consensus definition by the Commission on Global Governance (1995: 2): «Governance is the sum of the many ways individuals and institutions, public and private, manage their common affairs».

6. See also: Calcara *et al*, 2020: 8.

This definition aptly captures the changing character of the State, where after gradually being shaped by endogenous and exogenous factors, «governing» shifted to «governance». The most crucial aspect to be named for both dimensions is the uptake of ICTs and the subsequent process of digitalization[7]. The phenomenon of fragmentation therefore applies equally to, is represented in and by the current context of technological disruption, to which large digital platforms, in particular GAFAM (Google, Amazon, Facebook, Apple and Microsoft) contributed.

As such, impacting all dimensions of society, digitalization has further eroded traditional understanding of functioning of the State. The continuing process of fragmentation characterized by increasing technological complexity bears varying impacts on policymaking: on its design, on selection and participation of new actors, on the policy goals and hence methods by which these are set out for being attained.

New governance frameworks such as *e-governance* (Garson, 2006), *Internet governance* (Kettemann, 2020), *cybersecurity governance* (Von Solms and von Solms, 2018) and *algorithmic governance* (Katzenbach and Ulbricht, 2019) have emerged. Lessig's (1999) prescient observation at the beginning of the uptake of ICT during the late 90's that «code [was] law» and that the architecture of cyberspace displayed unique features demanding a revision of traditional governance frameworks provides a well-grounded precursor to the understanding of current societally disruptive phenomena such as commercially driven social profiling and nudging or the proliferation of hate speech and the dissemination of disinformation, propelled by and ensuing from innovation in ICTs via cyberspace.

Essentially, the vast economic power digital platforms exert over society (Zuboff, 2019; Nemitz and Pfeffer, 2020) and the datafication resulting therefrom (Murray, 2020) have called for a revision of traditional governance frameworks. In particular, the growing knowledge asymmetry between programmers and policymakers present reasons for rethinking governance (Lessig, 1999; Van den Hoven, 2017; Buiten, 2018). Derived from the Internet governance discourse (Kettemann, 2020: 30), the notion of «multistakeholderism» has gained traction, calling for a broad-based participation of experts with technical, legal, social and economic expertise in policy and law-making processes.

7. Digitalization is understood here to be both a technical and social process of translating analog data and information, traditionally stored in the form of texts into machine-readable format by means of a binary code (Altwicker, 2019).

In light of this, the development of AI can be treated as a continuation of innovation in ICTs or digital technologies in general. In simplified terms, relying on present infrastructure capabilities, the technical, logical and social layer of cyberspace (Schmitt, 2017), with the uptake of AI, digitalization culminates in cognification of social processes (Kelly, 2016), the presentation of still narrowly defined human intelligence. The problem of *control* over these processes coupled with the speed of light of electronic signals travelling along fibre optic cables is compounded by the inherent and most characteristic feature of AI, its autonomy, grounded in ML-based inductive learning mechanisms, and thus its «black-box» character, which in turn disrupts traditional understanding of transparency, fairness, legality and accountability. An additional factor complicating the governance of AI is the multiplicity of actors involved in its design, uptake, maintenance and auditing. Finally, the speed of innovation in AI even leads some researchers to contend that AI misaligned with our value system might potentially pose an existential threat to humanity itself (Bostrom, 2014; Dafoe, 2018).

A new governance domain can be discerned from the current discourse in interdisciplinary technology governance literature: AI governance. Scholars and policymakers alike have proposed diverse approaches for steering AI.

Consequently, in addition to the non-exhaustive list of governance frameworks above, a new governance domain can be discerned from the current discourse in interdisciplinary technology governance literature: *AI governance*. While dynamic, yet fragmented and unconsolidated (Butcher and Beridze, 2019; Taeihagh, 2021) in character, scholars and policymakers alike have proposed diverse approaches for steering AI. Being in most respects holistic in scope, the discourse on AI governance primarily gravitates around the question of anticipating and decreasing future risks in the short-term and long-term through ethics guidelines, institutional building processes and codification of ethical guidelines into hard-coded norms (Larsson, 2021).

As one of the first researchers to set out a framework on AI governance, Gasser and Almeida (2017) and Dafoe (2018) sought to concretize reflections concerning AI's societal ramifications and opportunities around the dimensions of the (i) *Social and legal layer*, (ii) *Ethical layer* and (iii) *Technical layer*, and, respectively, (i) *the Technical Landscape*, (ii) *AI Politics* and (iii) *Ideal Governance* into policies, new institutions and norms through iterative multistakeholder consultations on an international level. Sub-communities such as the AI4Good initiative by the UN, tied to the UN Sustainable Development Goals (ITU, 2021; Cowls *et al.*, 2021), and the Ethically Aligned Movement of the Institute of Electrical and Electronics Engineers community (IEEE, 2019) have grown

out of motivation to create explainable, ethical AI which in the broader sense caters to the call to «open[ing] the black box of AI» (Buiten, 2019) to society.

Assisting AI engineers in particular to translate ethical guidelines and AI specific values through design, Umbrello and van de Poel (2021) suggested bottom-up and hybrid approaches to AI governance, drawing on the self-reflective VSD framework (Umbrello, 2021). By proving how AI HLEG principles can be concretized as higher order values through AI design while accounting for ethical tensions and potential dilemmas therein, their work lends not only credence to the operationalizability of the AI HLEG framework but also reaffirms how instrumental fundamental rights-based system thinking is to AI governance and innovation. To achieve Trustworthy AI, they call on AI engineers to primarily design *for* human values to counter the current rationale of market-driven innovation in AI and its influence on AI design, and provide complimentary guidelines on how to do so more effectively for the long term (Umbrello and van de Poel, 2021).

In the same light, arguments for treating AI not merely as a computer-science based technology, but one that is embedded and situated in societal structures, hence raising ethical, value-based and normative questions alike, are reflected in socio-legal discourses on AI (Rahwan, 2018; Floridi *et al*., 2018; Dignum, 2019; Larsson, 2021). In particular Rahwan's (2018) proposal of a socially inspired algorithmic contract sheds light on the fundamental societal challenge of retaining human agency in the uptake of AI, exemplified in the author's model of «society in the loop» (ibid.). Scrutinizing and explicating the interrelationship between AI development and human values on a societal level provides avenues for revised understanding of how society as a whole can prosper through the uptake of AI. This reflection is epitomized in his calling for a renewed social contract in an ever more quantifiable environment where «humans and governance algorithms» interact with each other (ibid.). In turn, the entire life cycle of AI, from its design to auditing must be understood within the societal context in which the digital technology is deployed.

The EU's contribution to AI governance

Elements of proposals for governance frameworks on AI have found expression in policy-documents such as in the multiple *national AI strategies* (OECD, 2021; Van Roy *et al*, 2021), and in particular in the EU's *Ethics Guidelines* and *Policy and Investment Recommendations for Trustworthy AI* (AI HLEG, 2019a and 2019b), the EU Commission's *White Paper on AI* (2020) and the EU's legislative proposal on AI, the AIA (European Commission, 2021c).

For example, the VSD framework is aligned with the AI HLEG's seven key requirements and reflected in the EU's Trustworthy AI framework, which treats «not only the trustworthiness of the AI system itself but also (…) the trustworthiness of all *processes* and *actors* that are part of the system's life cycle» being iterative in nature (AI HLEG, 2019a). On the meta level, the EU's holistic proposal for the AIA presents not only the first-ever concretization of frameworks discussed in scholarly discourses on AI governance, in particular on AI4Good and Responsible AI discourses (Rahwan, 2018; Floridi *et al.*, 2018; Dignum, 2019; Theodorou and Dignum, 2020; Cowls *et al.*,, 2021; Larsson, 2021), but also the first global legislative instrument to address the growing recommendation in AI governance literature for institution building and a governance structure on AI, thus consolidating international debates around AI governance. This is also partially due to the participation of some of the leading scholars in the AI HLEG, such as Floridi or Dignum, whose research spans topics from the interdisciplinary fields of computer science, philosophy and law. And vice versa, the reports by the AI HLEG have shaped and shifted the discourse on AI governance from previous debates, focussing on existential *risk-centred* AI governance (Dafoe, 2018) towards *Trustworthy* AI governance, reflected in the adoption of OECD principles on AI and the US AI principles (Thiebes *et al.*, 2021).

Trust, excellence or both?

Tailored legislative measures and standardization efforts are linchpins for the EU's global competitiveness in AI uptake (Data Ethics Commission, 2019). The head of the executive of the European Union, Commissioner Ursula von der Leyen, committed to the key objective in the *Commission's Political Guidelines for 2024* to devise a regulatory framework for AI, based on European values and norms (EU Commission, 2021a).

From 2018, with the establishment of the 52-member strong multistakeholder AI HLEG[8] up until the AIA, the EU has drawn on variable policy instruments to create a governance framework for AI, three of which stand out: (i) *Policy and*

8. The independence of the High-Level Expert Group on AI is worthy of special note. As such, «the views expressed in [their documents] reflect the opinion of the AI HLEG and may not in any circumstances be regarded as reflecting an official position of the European Commission» (AI HLEG, 2019a).

Ethics guidelines on AI, laying the groundwork for the concept of human-centric, Trustworthy AI, and setting a global standard; (ii) *White Paper on AI*, setting out the vision for an AI ecosystem based on trust and excellence; and (iii) *Legislative Proposals*, in particular the Digital Services Act, Digital Markets Act and the latest proposal for an Artificial Intelligence Act (AI HLEG, 2019a and 2019b; EU Commission, 2021a).

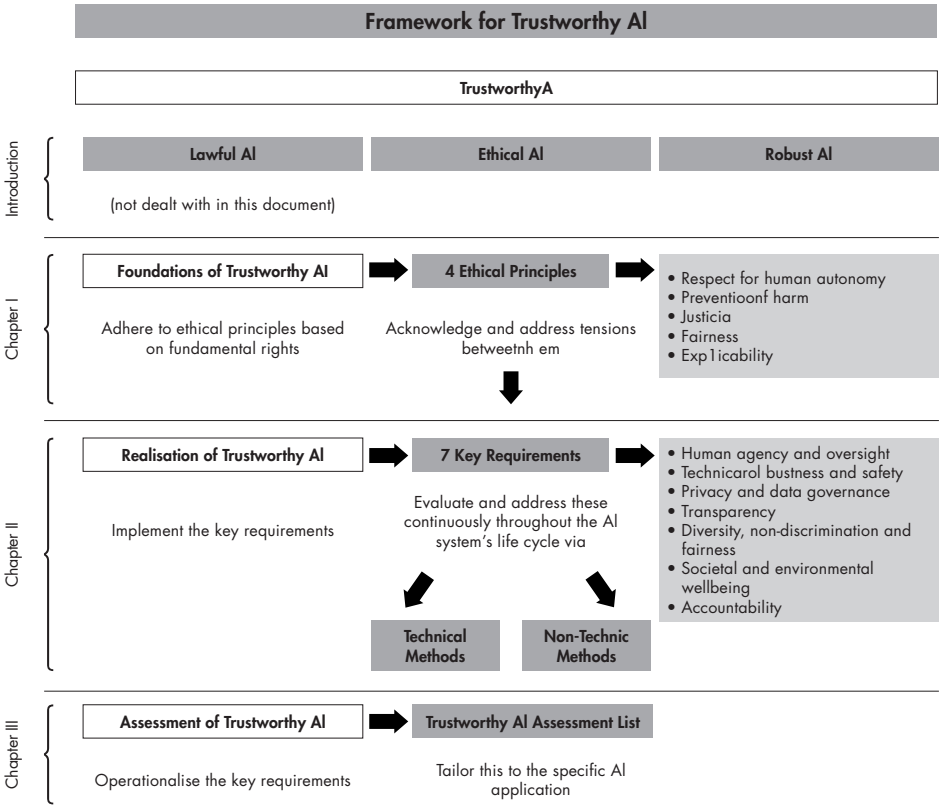
These joint efforts have in turn incrementally fed into a comprehensive yet not uncontested regulatory framework on AI. Along with the earlier introduction of GDPR, Brussels latest policies on AI governance could in turn be plausibly conceived as reaffirming the EU's traditional assumed role as a principle-based «regulatory superpower» (Bradford, 2020a; Bakardjieva Engelbrekt *et al.*, 2021). What Bakardjieva Engelbrekt *et al.* (2021) construe as a «technological shift», has been arguably addressed within the context of the «Brussels effect», referring to the EU's innate ability as the world's largest Single Market to promulgate global legal standards which later find inspiration and adoption beyond the remit of its jurisdiction (Bradford, 2020b). One must assume Bradford's empirically tested concept is reflected in the EU's efforts on global regulation of tech, with the GDPR being exemplary for this. The same pattern continues to apply for its regulatory efforts to tame the monopolistic power of GAFAM and other global tech players (ibid.). While this paper does not intend to adopt Bradford's methodology, it draws on its leitmotif, to wit the EU's inherent resolve to act as a beacon for democracy, the rule of law and promoting fundamental rights, all holistic concepts and principles which are more relevant than ever in the «technological shift». These same globally challenged fundamental rights-based principles find expression in the Trustworthy AI framework, a concept which the EU aims to export and attain leadership with, through AI.

Tailored legislative measures and standardization efforts are linchpins for the EU's global competitiveness in AI uptake. The head of the executive of the European Union, Commissioner Ursula von der Leyen, committed to the key objective to devise a regulatory framework for AI, based on European values and norms.

The EU Commission set out to steer AI development, marketing and application primarily based on a four-dimensional, risk-based approach (EU Commission, 2021c). While more comprehensive and thus nuanced in scope compared to the previously envisaged binary proposal in the *White Paper on AI*, which solely distinguished between low and high-risk AI applications (EU Commission, 2020), the risk-based classification method in the AIA still appears to give rise to criticism. For «AI made in the EU» to be deemed trustworthy (figure 1), developers and designers of these systems must adhere to three key requirements: (i) An AI must comply with all legal norms; (ii) adhere to ethical

standards and values in democratic societies; and (iii) present a high degree of both technical, e.g. in the area of cybersecurity, and social robustness when it comes to AI safety and principles such as explicability, fairness, prevention of harm and respect for human autonomy, in particular (AI HLEG, 2019a).

Figure 1. Framework for trustworthy AI of the High-Level Expert Group on AI (AI HLEG)

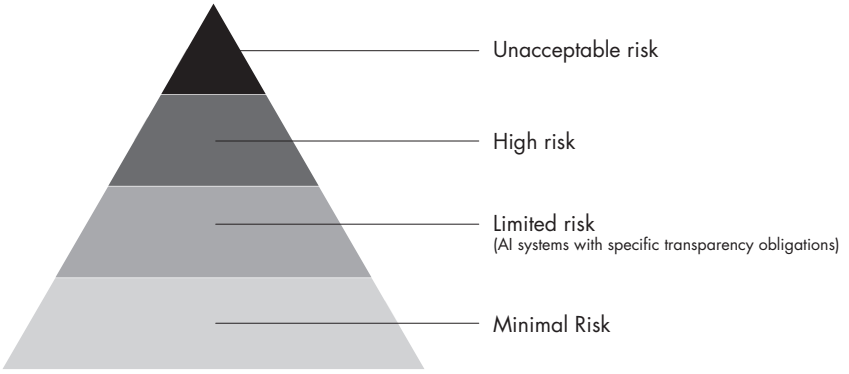


Surce: AI HLEG (2019a).

Whereas the AI HLEG expounded on the second and third requirements, the EU Commission as the executive body of the EU, devised a legislative proposal to realize «lawful AI», complementing the Trustworthy AI framework with the four-

dimensional risk-based approach (figure 2; EU Commission, 2021c). This in turn begs the question, of how the fundamental rights-based Trustworthy AI framework is represented in the AIA, specifically against the backdrop of the EU Commission's intention to create «a light governance structure» on AI (EU Commission, 2021b). In other words, can the EU attain Trustworthy AI with a light governance structure, encapsulated in the risk-based approach? Drawing on the five elements of the governance framework outlined above, combined with the seven key requirements of Trustworthy AI, the paper revisits *feedback* received on the AIA proposal, in particular commentary provided by human rights scholars (see e.g. Smuha *et al.*, 2021) and contends, that the AIA in its current form lacks fundamental elements to be deemed trustworthy in the «Lawful AI» dimension.

Figure 2. Four-dimensional risk-based approach of the Artificial Intelligence Act (AIA)



Source: European Commission (2021d).

Multiplicity of actors

While the EU allowed for public participation in the process towards the AIA, e.g. by means of surveying small and medium-sized companies on the Trustworthy AI framework, establishing a Europe-wide forum on AI policy discussions, the «AI Alliance», and the AI HLEG, questions still linger as to the inclusiveness of the process. This criticism is reflected in the AIA, which neither provides for *procedural* nor *substantive rights* for EU citizens affected by adverse AI (Smuha *et al.*, 2021), either in areas of low or high-risk applications, and is at odds with the Trustworthy AI principles of *human agency*, *fairness*, *societal wellbeing* and *accountability*. As such,

the AIA in its current form does not empower citizens to the degree of allowing for informed and transparent redress mechanisms (Smuha *et al.*, 2021), in cases of potentially flawed or illegal uses of AI, as previously envisaged by the AI HLEG.

Hence the aspect of EU citizen participation must be improved in the EU's AI governance structure. Based on the rule of law principles of legality, fairness and accountability, the essential ingredient of inclusiveness ought to gain additional weight in the debate on AI governance at an EU level. Inclusive governance frameworks entail elements that allow for citizen participation, rendering them legitimate. They are geared to increasing transparency of the policymaking process itself. Finally, they may be adjusted over time, and are thus iterative and context-dependent in scope. While not exhaustive, the combination of these key factors permits citizens' trust in the policymaking process and in state institutions to develop and grow (Pierre and Peters, 2021), independent of the time and context in which technologies, digital, autonomous or complex in nature, are introduced.

Consequently, voices of ordinary EU citizens need to be reflected in the AIA, empowering them in particular, but not only by means of complaints and accountability mechanisms. To this end, the European Parliament, as the representative for 450 million EU citizens, must advocate broader citizen participation in future formulation and definition of the AIA, introducing e.g. the right to direct participation in potential revisions of AI-risk application. EU citizens could e.g. deliver cases and complains directly to the *European Artificial Intelligence Board* or to their respective national supervisory authorities.

Mechanisms, structures, institutionalization and authority

Whereas the AIA envisages establishing a *European Artificial Intelligence Board*, with experts from 27 EU member states and the EU Commission being represented therein to permit adequate enforcement of the AIA, questions remain as to good administrative practices of the Act, since the required levels of expertise around AI infrastructure must be built up, calling for broad-based training of experts and the level of investment required for infrastructure capabilities of national supervisory authorities, as experience around the GDPR context indicated these institutions were generally underfunded in various EU countries, thus hamstringing and crippling effective administration of GDPR compliance and enforcement of data protection laws.

For the long-term perspective on implementation and monitoring of the AIA, providing training to civil servants on advanced digital literacy, in particular on how to identify AI-specific threats to human rights is of particular importance, considering that designers of AI systems are provided great leeway in their decision

to conduct *ex ante* conformity assessments, which, questionably, pertains only to high-risk AI applications[9] listed under Art. 6(2) and Annex III (EU Commission, 2021c) but not to low-risk ones. Owing to the complex nature and context-dependency of AI, if left unaddressed, this might bear negative impacts on EU citizens' agency and privacy rights. While creating a database on high-risk AI applications at the European level respects the aim of effective supervision and monitoring of critical AI systems, from a fundamental rights law perspective, one may question whether the database should not extend to all four risk dimensions.

Additionally, transposing a spirit of fundamental rights protection into the AIA, it would be worthwhile to establish a complementary external auditing body, which iteratively revisits the criteria of *necessity* and *proportionality*, interrogating whether temporal infringements of fundamental rights are *necessary in a democratic society*, legitimate and hence *proportionate* (Smuha *et al.*, 2021). This is of particular relevance in public sector uses of AI, taking into account that the AIA allows for use of biometric identification systems in the field of law enforcement. Adverse applications and misconceived interpretations of the necessity and proportionality requirements could bear serious repercussions on the Trustworthy AI requirements of *human agency and oversight*, *privacy and data governance*, *diversity*, *non-discrimination and fairness*, *societal wellbeing* and *accountability*. While necessary for monitoring product safety and health requirements in the area of AI, in general, market surveillance authorities can neither replace nor assume the roles of institutions, whose core competence lies in safeguarding EU citizens' fundamental rights, extending to sensitive areas such as data protection and privacy.

Consequently, referring back to the final research question, while displaying and respecting elements of the Trustworthy AI requirements, in particular in terms of *unacceptable risks* of AI usages that run counter to democratic values, the rule of law and fundamental rights as such, the first global horizontal legislative proposal on AI is tilted rather in favour of creating an ecosystem of excellence. When compared to GDPR, the AIA must thus be understood as being an innovation-inspired legislative proposal, which shifts the debate on AI governance

9. Examples of high-risk AI applications: «(i) safety component of products subject to third party ex-ante conformity assessment; and (ii) stand-alone AI systems with mainly fundamental rights implications' in areas of: Biometric identification and categorisation of natural persons; Management and operation of critical infrastructure; Education and vocational training; Employment, workers management and access to self-employment; Access to and enjoyment of essential private services and public services and benefits: Law enforcement; Migration, asylum and border control management; Administration of justice and democratic processes».

from a language of *Trustworthy AI*, a rights-based approach espoused by the AI HLEG, to a language of *risk-based AI*, deemed to be rather innovation-friendly, with *economic* and *human values* remaining in tension with one another.

Concluding remarks

The rationale of this paper has been to examine whether the EU's governance framework in its current form is aligned with the goal of creating an inclusive, fundamental rights-based and thus Trustworthy AI ecosystem. After taking account of the double-edged sword character of AI systems and the challenges these presented in developing a human-centric AI governance framework, this paper revisited the latest discourse on AI governance literature from an EU perspective. Thereafter, it assessed EU policy measures on AI and discussed the EU's AI governance structure from a fundamental rights-based perspective, contrasting the EU Commission's innovation-inspired proposal of the AIA, construed as representing an *ecosystem of excellence*, with the Trustworthy AI framework of the AI HLEG, primarily conceived to create an *ecosystem of trust*. Situating EU policy measures into scholarly discourses on AI governance and vice versa, this paper provides initial insights into the EU's role in and contribution to an emerging AI governance framework.

This paper calls for additional mechanisms empowering and allowing EU citizens to participate directly in the future shaping and direction of implementation of the AIA, aligned with furthering the Trustworthy AI requirements of human agency and oversight, and accountability.

By framing AI as an autonomous digital technology embedded into societal structures and contexts, mediated through digital devices, the paper has contended that potential tensions between both approaches arise and find expression in the AIA's four-dimensional risk-based approach, partially compromising Trustworthy AI principles, in particular but not limited to *human agency*, *fairness* and *accountability*. However, given the nascent nature of the field of AI governance, the first-ever legislative proposal on AI opens avenues for broad-based discussions on how democratic societies, based on the rule of law and fundamental rights-inspired values, intend to live in an ever more AI conditioned, technology driven environment.

This paper thus calls for additional mechanisms empowering and allowing EU citizens to participate directly in the future shaping and direction of implementation of the AIA, aligned with furthering the Trustworthy AI

requirements of *human agency* and *oversight*, and *accountability*. Iteratively questioning the risk-based approach, future research would need to call for case-based, context-dependent empirical studies on the effectiveness of self-conformity assessments of AI designers weighed against the Trustworthy AI concept. Extensive public opinion surveys would need to be conducted in all EU member states, to collect data on EU citizens' perception on both the rights-based and risk-based concepts. These findings must underpin and inform discussions on both approaches to achieve inclusive Trustworthy AI. No less important are measures to collect additional data on the environmental impact of uptake of AI technologies, in close conjunction with studies on role of AI in potentially reducing the carbon footprint.

In essence, placing citizens' rights centre stage and empowering them through digital transformation is key for development and uptake of Trustworthy AI. The EU Commission's proposal provides a globally unique starting point for these discussions, on international, national and sub-national levels. What is called for is additional political will at the EU Commission level, not only to partially endorse but also fully integrate ideas from hybrid if not bottom-up AI governance approaches, in particular those rooted in the fundamental rights-based system thinking methods of Value Sensitive Design.

Only through understanding, and based on that long-term process, iteratively evaluating the risks of widescale societal uptake of AI, can democratically elected public officials help empower citizens to a degree that allows them to leverage digital technology for societal good in globally contested digital ecosystems.

Bibliographical references

AI HLEG - High-level Expert Group on Artificial Intelligence. «Ethics guidelines for trustworthy AI». *European Commission*, (8 April 2019a) [Accessed on: 15.03.2022] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

AI HLEG - High-level Expert Group on Artificial Intelligence. «Policy and investment recommendations for trustworthy Artificial Intelligence». *European Commission*, (8 April 2019b) [Accessed on: 15.03.2022] https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Altwicker, Tilmann. «International Legal Scholarship and the Challenge of Digitalization». *Chinese Journal of International Law*, vol. 18, n.º 2 (2019), p. 217-246, DOI: https://doi.org/10.1093/chinesejil/jmz012

Antonov, Alexander and Kerikmäe, Tanel. «Trustworthy AI as a Future Driver for Competitiveness and Social Change». In: Troitiño, David; Kerikmäe, Tanel; de la Guardia, Ricardo Martín and Pérez Sánchez, Guillermo Á. (eds.). *The EU in the 21st Century Challenges and Opportunities for the European Integration*, Cham: Springer, p. 135–154, DOI: https://doi.org/10.1007/978-3-030-38399-2_9

Bakardjieva Engelbrekt, Antonina; Leijon, Karin; Michalski, Anna and Oxelheim, Lars (2021). «What Does the Technological Shift Have in Store for the EU? Opportunities and Pitfalls for European Societies». In: Engelbrekt, Antonina Bakardjieva *et al.* (eds.). *The European Union and the Technology Shift*. Cham: Springer Nature, 2021, p.1-25.

Bostrom, Nick. *Superintelligence: Paths, dangers, strategies.* Oxford: Oxford University Press, 2014.

Bradford, Anu. *The Brussels effect: How the European Union rules the world*. Oxford: Oxford University Press, 2020a.

Bradford, Anu. «The Brussels Effect Comes for Big Tech». *Project Syndicate*, (17 December 2020b) [Accessed on: 15.03.2022] https://www.project-syndicate.org/commentary/eu-digital-services-and-markets-regulations-on-big-tech-by-anu-bradford-2020-12

Brynjolfsson, Erik and McAfee, Andrew. «What's Driving the Machine Learning Explosion?». *Harvard Business Review*, 18 July 2017 [Accessed on: 15.03.2022] https://hbr.org/2017/07/whats-driving-the-machine-learning-explosion

Buiten, Miriam. «Towards Intelligent Regulation of Artificial Intelligence». *European Journal of Risk Regulation*, vol. 10, n.º1 (2019), p. 41-59. DOI: https://doi.org/10.1017/err.2019.8

Butcher, James and Beridze, Irakli. «What Is the State of Artificial Intelligence Governance Globally?». *The RUSI Journal*, vol. 164, n.º 5–6 (2019), p. 88–96. DOI: https://doi.org/10.1080/03071847.2019.1694260

Calcara, Antonio; Csernatoni, Raluca and Lavallée, Chantal (eds.). *Emerging Security Technologies and EU Governance: Actors, Practices and Processes.* London: Routledge, 2020.

Chiusi, Fabio; Fischer, Sarah; Kayser-Bril, Nicolas and Spielkamp, Matthias. «Automating Society Report 2020». *Algorithm Watch* (30 September 2020) [Accessed on: 15.03.2022] https://automatingsociety.algorithmwatch.org/wp-content/uploads/2020/12/Automating-Society-Report-2020.pdf

Commission on Global Governance. *Our Global Neighbourhood: The Report of the Commission on Global Governance*. Oxford: Oxford University Press, 1995.

Cowls, Josh; Tsamados, Andreas; Taddeo, Mariarosaria and Floridi, Luciano. «A definition, benchmark and database of AI for social good initiatives». *Nature*

Machine Intelligence, vol. 3, n.º 2 (2021), p. 111-115. DOI: https://doi.org/10.1038/s42256-021-00296-0

Dafoe, Allan. «AI governance: a research agenda». *Governance of AI Program, Future of Humanity Institute*, (27 August 2018) [Accessed on: 15.03.2022] https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

Data Ethics Commission of the [German] Federal Government. «Opinion of the Data Ethics Commission». *DEK,* (December 2019) [Accessed on: 15.03.2022] https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3

Dignum, Virginia. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Cham: Springer Nature, 2019.

Ebers, Martin. «Liability for Artificial Intelligence and EU Consumer Law». *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, vol. 12, n.º 2 (2021), p. 204–220.

EU Commission. «Artificial Intelligence for Europe». *EC*, COM/2018/237 final (25 April 2018a) [Accessed on: 15.03.2022] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

EU Commission. «Coordinated Plan on Artificial Intelligence». *EC*, COM/2018/795 final (7 December 2018b) [Accessed on: 15.03.2022] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0795

EU Commission. «European White Paper on Artificial Intelligence: a European approach to excellence and trust». *EC*, COM(2020) 65 final (19 February 2020) [Accessed on: 15.03.2022] https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

EU Commission. «A Europe fit for the digital age». *CE*, (9 March 2021a) [Accessed on: 15.03.2022] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age_en

EU Commission. «Fostering a European approach to Artificial Intelligence». *EC*, *COM*/2021/205 final (21 April 2021b) [Accessed on: 15.03.2022] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2021:205:FIN&qid=1619355277817

EU Commission. «Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts». *EC*, COM (2021) 206 (21 April 2021c) [Accessed on: 15.03.2022] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

EU Commission. «Regulatory framework proposal on Artificial Intelligence» (31 August 2021d) [Accessed on: 15.03.2022] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Floridi, Luciano *et al*. «AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations». *Minds and Machines*, vol. 28, n.º 4 (2018), p. 689–707, DOI: https://doi.org/10.1007/s11023-018-9482-5

Garson, George D. *Public information technology and e-governance: Managing the virtual state*. Burlington: Jones & Bartlett Learning, 2006.

Gasser, Urs and Almeida, Virgilio AF. «A Layered Model for AI Governance». *IEEE Internet Computing*, vol. 21, n.º6 (2017), p. 58–62, DOI: https://doi.org/10.1109/MIC.2017.4180835

Hamulák, Ondrej. «La carta de los derechos fundamentales de la union europea y los derechos sociales». *Estudios constitucionales*, vol. 16, n.º1 (2018), p. 167-186, DOI: http://dx.doi.org/10.4067/S0718-52002018000100167

Institute of Electrical and Electronics Engineers (IEEE). «Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems». *IEEE*, (31 March 2019) [Accessed on: 15.03.2022] https://ieeexplore.ieee.org/servlet/opac?punumber=9398611

International Telecommunication Union (ITU). «Measuring the Information Society Report 2015». *UN*, (2015) [Accessed on: 15.03.2022] https://www.itu.int/en/ITU-D/Statistics/Pages/publications/mis2015.aspx

International Telecommunication Union (ITU). «AI for Good». *UN*, (2021) [Accessed on: 15.03.2022] https://aiforgood.itu.int/about/

Katzenbach, Christian and Ulbricht, Lena. «Algorithmic governance». *Internet Policy Review*, vol. 4, n.º 8 (2019). DOI: https://doi.org/10.14763/2019.4.1424

Kelly, Kevin. *The inevitable: Understanding the 12 technological forces that will shape our future*. New York: Viking Press, 2016.

Kettemann, Matthias C. *The normative order of the internet: A theory of rule and regulation online*. Oxford: Oxford University Press, 2020.

Kohler-Koch, Beate and Rittberger, Berthold. «Review Article: The Governance Turn in EU Studies». *Journal of Common Market Studies*, vol. 44, (2006), p. 27-50, DOI: https://doi.org/10.1111/j.1468-5965.2006.00642.x

Langford, Malcolm. «Taming the digital leviathan: Automated decision-making and international human rights». *American Journal of International Law*, vol. 114, (2020), p. 141-146, DOI: https://doi.org/10.1017/aju.2020.31

Larsson, Stefan. «The Socio-Legal Relevance of Artificial Intelligence». *Droit et société*, vol. 103, n.º 3 (2019), p. 573-593, DOI: https://doi.org/10.3917/drs1.103.0573

Larsson, Stefan. «AI in the EU: Ethical Guidelines as a Governance Tool. Why Ethics Guidelines?». In: Bakardjieva Engelbrekt, Antonina; Leijon, Karin; Michalski, Anna and Oxelheim, Lars (eds.). *The European Union and the Technology Shift*. Cham: Springer, 2021, p. 85-111.

Lawson, Clive. *Technology and isolati*on. Cambridge: Cambridge University Press, 2017.

Lessig, Lawrence. *Code and Other Laws of Cyberspace*. New York: Basic Books, 1999.

Murray, Andrew. «Talk at Sixth Annual T.M.C. Asser Lecture on Law and Human Agency in the Time of Artificial Intelligence». *Annual T.M.C. Asser Lecture* 2020, (26 November 2020) [Accessed on: 15.03.2022] https://www.asser.nl/annual-lecture/annual-tmc-asser-lecture-2020/

Nemitz, Paul and Pfeffer, Matthias. *Prinzip Mensch: Macht, Freiheit und Demokratie im Zeitalter der Künstlichen Intelligenz*. Bonn: Dietz, 2020.

Organisation for Economic Co-operation and Development (OECD). «National AI policies and strategies». *OECD. AI Policy Observatory* (September 2021) [Accessed: 15.03.2022] https://www.oecd.ai/dashboards

Pierre, Jon and Peters, Guy. *Advanced Introduction to Governance*. Cheltenham: Edward Elgar Publishing, 2021.

Pohle, Julia and Thiel, Thorsten. «Digital sovereignty». *Internet Policy Review*, vol. 9, n.º **4** (2020), p. 1-19. DOI: https://doi.org/10.14763/2020.4.1532

Rahwan, Ilyad. «Society-in-the-loop: programming the algorithmic social contract». *Ethics and Information Technology*, vol. 20 (2018), p. 5–14, DOI: https://doi.org/10.48550/arXiv.1707.07232

Rhodes, Roderick. «The New Governance: Governing Without Government». *Political Studies*, vol. 44, n.º 4 (1996), p. 652-667, DOI: https://doi.org/10.1111/j.1467-9248.1996.tb01747.x

Russell, Stuart and Norvig, Peter (eds.). *Artificial Intelligence: A Modern Approach*. London: Pearson, 2010.

Schmitt, Michael N. *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. Cambridge: Cambridge University Press, 2017.

Scott, Marcus; Petropoulos, Georgios and Yeung, Timothy. «Contribution to growth: The European Digital Single Market. Delivering economic benefits to citizens and businesses». *European Parliament*, PE 631.044 (January 2019) [Accessed on 15.03.2022] https://www.bruegel.org/wp-content/uploads/2019/02/IPOL_STU2019631044_EN.pdf

Smuha, Nathalie. «Beyond a human rights-based approach to AI governance: Promise, pitfalls, plea». *Philosophy and Technology*, vol. 34, (2021), p. 91-104, DOI: https://doi.org/10.1007/s13347-020-00403-w

Smuha, Nathalie; Ahmed-Rengers, Emma; Harkens, Adam; Li, Welong; MacLaren, James; Piseli, Ricardo and Yeung, Karen. «How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act». *SSRN Electronic Journal*, (2021), p. 1-59.

Taeihagh, Araz. «Governance of artificial intelligence». *Policy and Society*, vol. 40, n.º 2 (2021), p. 137-157, DOI: https://doi.org/10.1080/14494035.2021.1928377

Theodorou, Andreas and Dignum, Virginia. «Towards ethical and socio-legal governance in AI». *Nature Machine Intelligence*, vol. 2, n.º 1 (2020), p. 10-12, DOI: https://doi.org/10.1038/s42256-019-0136-y

Troitiño, David Ramiro and Kerikmäe, Tanel. «Europe facing the digital challenge: obstacles and solutions». *IDP. Revista de Internet, Derecho y Política*, n.º 34 (2021), p. 1-3, DOI: https://doi.org/10.7238/idp.v0i34.393310

Umbrello, Steven. «Conceptualizing Policy in Value Sensitive Design: A Machine Ethics Approach». In: Thompson, Steven John (ed.). *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*, Hershey: IGI Global, 2021, p. 108-125.

Umbrello, Steven. «The Role of Engineers in Harmonising Human Values for AI Systems Design». *Journal of Responsible Technology*, vol. 10, (2022), p. 1-10, DOI: https://doi.org/10.1016/j.jrt.2022.100031

Umbrello, Steven and Van de Poel, Ibo. «Mapping value sensitive design onto AI for social good principles». *AI and Ethics*, vol. 1, n.º 3 (2021), p. 283-296, DOI: https://doi.org/10.1007/s43681-021-00038-3

Van de Poel, Ibo. «Embedding values in artificial intelligence (AI) systems». *Minds and Machines*, vol. 30, n.º 3 (2020), p. 385-409, DOI: https://doi.org/10.1007/s11023-020-09537-4

Van den Hoven, Jerome. «Ethics for the Digital Age: Where Are the Moral Specs? – Value Sensitive Design and Responsible Innovation». In: Werthner, Hannes and van Harmelen, Frank (eds.). *Informatics in the Future*. Cham: Springer, 2017, p. 65-76.

Van Roy, Vincent; Rosetti, Fiammetta; Perset, Karine and Galindo-Romero, Laura. «AI Watch - National strategies on Artificial Intelligence: A European perspective». *European Union*, JRC122684 (June 2021) [Accessed on: 15.03.2022] https://publications.jrc.ec.europa.eu/repository/handle/JRC119974

Von Solms, Basie and von Solms, Rossouw. «Cybersecurity and information security–what goes where?». *Information and Computer Security,* vol. 26, n.º 1 (2018), p. 2-9, DOI: https://doi.org/10.1108/ICS-04-2017-0025

Zicari, Roberto V. *et al*. «Co-design of a trustworthy AI system in healthcare: Deep learning based skin lesion classifier». *Frontiers in Human Dynamics*, vol. 3, (2021), p. 1-20, DOI: https://doi.org/10.3389/fhumd.2021.688152

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books, 2019.