DOCTORAL THESIS

Human Rights-Based Legal Conditions for Trustworthy AI

Alexander Antonov

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2024

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
55/2024

Human Rights-Based Legal Conditions for Trustworthy AI

ALEXANDER ANTONOV

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Business and Governance
Department of Law/Ragnar Nurkse Department of Innovation and Governance
This doctoral thesis was accepted for the defence of the degree in Public Administration
11/08/2024

Supervisor:
Professor Dr Tanel Kerikmäe
School of Business and Governance
Department of Law
Tallinn University of Technology
Tallinn, Estonia

Co-supervisor:
Associate Professor Dr Thomas Hoffmann
School of Business and Governance
Department of Law
Tallinn University of Technology
Tallinn, Estonia

Opponents:
Professor Dr Azar Aliyev
Faculty of Law, Economics and Business
Department of Law
Martin-Luther-Universität Halle-Wittenberg
Halle-Wittenberg, Germany

Professor Dr Ronald Leenes
Tilburg Law School
Tilburg Institute for Law, Technology, and Society
Tilburg University
Tilburg, the Netherlands

Defence of the thesis: 16/10/2024, Tallinn

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology has not been submitted for doctoral or equivalent academic degree.

Alexander Antonov

signature

European Union
European Regional
Development Fund

Investing
in your future

ISSN 2585-6898 (publication)
ISBN 978-9916-80-202-1 (publication)
ISSN 2585-6901 (PDF)
ISBN 978-9916-80-203-8 (PDF)
DOI https://doi.org/10.23658/taltech.55/2024
Printed by Koopia Niini & Rauam

Antonov, A. (2024). *Human Rights-Based Legal Conditions for Trustworthy AI* [TalTech Press]. https://doi.org/10.23658/taltech.55/2024

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
55/2024

Inimõigustel põhinevad õiguslikud tingimused usaldusväärse tehisintellekti loomisel

ALEXANDER ANTONOV

TAL TECH KIRJASTUS

Contents

List of publications

The list of author’s publications, on the basis of which the thesis has been prepared:

I. **3.1 Antonov, A**., & Kerikmäe, T. (2020). Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU. In D. R. Troitiño, T. Kerikmäe, R. M. de la Guardia, & G. Á. P. Sánchez (Eds.), *The EU in the 21st Century: Challenges and Opportunities for the European Integration Process* (pp. 135−154). Springer. https://doi.org/10.1007/978-3-030-38399-2_9

II. **3.1 Antonov, A**., Häring, T., Korõtko, T., Rosin, A., Kerikmäe, T., & Biechl, H. (2021). Pitfalls of Machine Learning Methods in Smart Grids: A Legal Perspective. In *Proceedings - 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, 12-14 November 2021, Rome (pp. 248−256). Institute of Electrical and Electronics Engineers (IEEE). Danvers. https://doi.org/10.1109/ISCSIC54682.2021.00053

III. **1.1 Antonov, A**. (2022). Managing Complexity: The EU’s Contribution to Artificial Intelligence Governance. *Revista CIDOB d’Afers Internacionals*, (131), 41−65. https://doi.org/10.24241/rcai.2022.131.2.41/en

Appendix

IV. **2.5** Kerikmäe, T., Nyman Metcalf, K., Hoffmann, T., Minn, M., Liiv, I., Taveter, K., Shumilo, O., Solarte, C. & **Antonov, A**. (2019). 1st Report on Legal Framework and Analysis Related to Autonomous Intelligent Technologies. In Riigikantselei (pp. 1−11).

V. **2.1** Kasper, A., & **Antonov, A**. (2019). Towards Conceptualizing EU Cybersecurity Law. In *ZEI Discussion Paper*, C253. Bonn: Center for European Integration Studies [Zentrum für Europäische Integrationsforschung]. https://hdl.handle.net/20.500.11811/9849

NB: Except for the papers in the appendix, all the publications included in the cumulative doctoral thesis are indexed in the abstract and citation database *Scopus*.

Author's contribution to the publications

The author's contributions to the publications in this thesis are:

I The author of this cumulative thesis is the lead author of the co-authored book chapter. He co-developed the argument of the publication, conducted the literature review and was responsible for writing the introduction, methodology, empirical analysis and conclusion sections.

II The author of this cumulative thesis is the lead and corresponding author of the co-authored book chapter. He co-wrote the introduction and the conclusion sections and conducted the legal analysis.

III The author of this cumulative thesis is the sole author of the article.

IV The author of this cumulative thesis contributed a literature review for the legal analysis.

V The author of this cumulative thesis is the second author of the co-authored article. He co-developed the argument, contributed a literature review, and co-wrote the legal analysis of the case study.

1 Introduction: Focus and aim

"[D]emocratic AI governance necessitates aligning the activities of standard-setting bodies, who influence industrial processes and products, with the societal objectives and the requirements of public accountability enshrined in law."

Transatlantic Reflection Group on Democracy and the Rule of Law in the Age of "Artificial Intelligence", 2023, p. 250

Situated at the intersection of law, technology and human rights studies, this thesis examines legal frameworks for protecting and promoting the fundamental rights of individuals and groups concerning the development and public-sector use of artificial intelligence (AI) systems[1]. The objective of this normative analysis is to determine the legal conditions under which the development and public-sector use of AI can be compatible with fundamental rights.[2] It aims to contribute to the literature on regulating AI, particularly to human rights-based approaches to AI regulation.

The thesis considers an *AI system* an advanced networked digital information and communications technology (ICT) and draws on the definition of AI in Article 3(1) of the EU Artificial Intelligence Act (AI Act) (Regulation 2024/1689): "AI system means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." Following the High-Level Expert Group on Artificial Intelligence[3] (AI HLEG) (High-Level Expert Group on Artificial Intelligence, 2019a, p. 7), the author refers to *fundamental rights*[4] as the obligations of States to respect, protect and fulfil the freedoms and civil-political and socio-economic rights of individuals and groups under the principles of democracy and the rule of law. By focusing exclusively on *natural persons* as rights-holders, fundamental rights are defined as codified under the Charter of Fundamental Rights of the European Union (the Charter) and moral entitlements (Charter, 2016; High-Level Expert Group on Artificial Intelligence, 2019a). In this thesis, the term *natural persons* refers to individuals or groups who are *citizens* and *data subjects*. By *citizens* or *data subjects*, the thesis refers to citizens, residents, and refugees in the EU or other jurisdictions in their role as *end-users* of AI technology. Additionally, the author refers to individuals or groups adversely affected by an AI system as *affected persons*. Everyone, regardless of legal status, can be an affected person. The terms

[1] This thesis uses the notions of *AI*, *AI system* and *AI technology* interchangeably.

[2] In this thesis, *condition* refers to "a requirement involved in a law, or other legally recognized document that changes the rights and duties of those involved" (online legal dictionary Wex, Legal Information Institute, Cornell Law School: https://www.law.cornell.edu/wex/condition).

[3] To inform on the implementation of the EU AI strategy, the European Commission mandated an expert group to create ethics guidelines for AI, including policy and investment recommendations. On behalf of the Commission, the AIHLEG developed the concept of Trustworthy AI and proposed a three-part framework (*lawful*, *ethical*, *robust AI*) to achieve it. The AI HLEG was comprised of representatives from the European industry and academia. While it established ethical and robust AI, the Commission, Council of the EU, and European Parliament were responsible for addressing lawful AI as part of the ordinary legislative procedure by drafting the AI Act.

[4] This thesis applies the terms *fundamental rights* (used in the European context) and *human rights* (primarily referred to on the international levels) interchangeably throughout.

developers and *providers* denote individuals involved in the development or design of AI. The term *deployers* refers to users of AI systems in the public sector. These are *legal persons or entities* representing a public authority, agency, or body. Individuals who use or deploy AI systems under the authority of the deployer are *civil servants*, for example, public administrators, doctors, or governmental lawyers.

Promising to contribute to economic growth and increase efficiency across various areas, governments aim to create conditions for innovation in big data and AI and increase the use of AI in the market (Marcus *et al.*, 2019; Misuraca & van Noordt, 2020). Another key factor driving the development of AI is its potential for mitigating the impact of climate change in line with the United Nations (UN) Sustainable Development Goals (European Commission, 2019; Vinuesa *et al.*, 2020). For instance, AI is considered crucial for improving energy efficiency by automating the measurement and monitoring of energy consumption in real-time (Kaack *et al.*, 2022; Rolnick *et al.*, 2023).

However, as a general-purpose and dual-use socio-technical system, AI is a double-edged sword (Cohen, 2019; Murray, 2021; Taylor, 2023b; Yeung, 2022). Prior research and recent case law have shown that features of the complex phenomenon of AI, particularly partial autonomy and opacity, can give rise to risks and create material and immaterial harm to individuals and groups, ranging from bias and discrimination to violations of applicable data protection and privacy standards (Alston, 2019; Buolamwini & Gebru, 2018; Eubanks, 2018; Case 11519/20, 2023; Hoffmann-Riem, 2021; Case C-634/21, 2023; Mantelero, 2019; Case C-09-550982, 2020; Pasquale, 2015; Redden *et al.*, 2022; Wirtz *et al.*, 2020; Yeung, 2019, 2022). Furthermore, prior research has revealed that the implementation of AI systems could create significant information and power asymmetries among developers, deployers, and citizens in society (Busuioc, 2021; Cohen, 2019; Gasser & Almeida, 2017; High-Level Expert Group on Artificial Intelligence, 2019a; Taylor, 2023b; Yeung, 2022). Indeed, it has been argued that the information advantages and infrastructural power of the AI industry via-à-vis governments could reduce governments' ability as primary duty-bearers to respect, protect, and fulfil the human rights of rights-holders in their jurisdictions (Cohen, 2019; Floridi, 2020b; Hasselbalch, 2021; Morozov & Bria, 2018; Taylor, 2021; Yeung, 2019; Zuboff, 2019). This, in turn, could weaken governments' abilities to fulfil their monitoring obligations, enforce fundamental rights and provide effective redress for affected persons (Leslie *et al.*, 2021; Mantelero, 2019; A/HRC/38/35, 2018; Wagner, 2019; Yeung, 2019).

Regulating AI, therefore, has recently become a key priority for governments as part of broader global and regional multistakeholder initiatives on AI governance[5], involving industry and academia (Dafoe, 2018; Dignum, 2019; Ebers, 2020; European Commission, 2018; Evas, 2024; Executive Office of the President, 2016a; Mantelero, 2022; OECD, 2021; Rotenberg, 2024; A/78/L.49, 2024; Wendehorst, 2020). In promoting the uptake of the technology while addressing its risks to fundamental rights, governments aim to develop and deploy "human-centric and trustworthy AI" (High-Level Expert Group on Artificial Intelligence, 2019a; Regulation 2024/1689) or "safe, secure, and trustworthy AI systems" (Executive Order No. 13,960, 2020; Executive Order No. 14,110, 2023). This twin objective follows the tradition of regulating emerging technologies (Brownsword, 2019).

[5] *AI governance* is here understood as a "field [which] studies how humanity can best navigate the transition to advanced AI systems, focusing on the political, economic, military, governance, and ethical dimensions" (Dafoe, 2018, p. 5).

Initially, governments reviewed the adequacy of extant legal frameworks and developed non-binding ethics guidelines, recommendations, declarations, and codes of conduct for AI (Ad Hoc Committee on Artificial Intelligence (CAHAI), 2020; Black & Murray, 2019; Brownsword, 2019; Executive Office of the President, 2016a, 2016b; High-Level Expert Group on Artificial Intelligence, 2019a; Hoffmann-Riem, 2021; **IV**).[6] In this review process, policymakers and legislators aim to strike a balance between a precautionary approach related to safeguarding fundamental rights and a permissive approach related to minimising regulatory burdens not to forestall the economic and societal potential the introduction of AI is projected with (Brownsword, 2019; Brownsword & Goodwin, 2012; Crootof & Ard, 2021; Larsson, 2020). In the European jurisdictions, the world's first regulatory instrument for AI is emerging: the EU AI Act (Regulation 2024/1689). The risk-based approach of the AI Act follows the tradition of regulating emerging technologies and is grounded in EU product safety legislation (Evas, 2024; Martini *et al.*, 2024; Regulation 2024/1689; Wendehorst, 2020).

To examine and determine legal conditions for protecting and promoting the fundamental rights of individuals and groups concerning the development and public-sector use of AI systems, this thesis primarily focuses on the AI HLEG fundamental rights-based approach and its three-part framework of *lawful*, *ethical,* and *robust AI* for **Trustworthy AI** under the **EU AI strategy** (European Commission, 2018; High-Level Expert Group on Artificial Intelligence, 2019a, pp. 6-8). The framework has influenced the global AI regulatory debate, particularly the AI strategies of democracies worldwide, as reflected in the OECD AI Principles and the White House AI Principles (Fjeld *et al.*, 2020; Thiebes *et al.*, 2021). The author conceives the realisation of Trustworthy AI as *adequate protection and promotion of the fundamental rights of individuals and groups* concerning the development and public-sector use of AI. The thesis particularly focuses on the *lawful AI* dimension of the framework. For this purpose, several EU regulatory instruments have been analysed: First, legislation for cybersecurity, particularly the Cybersecurity Act, the Network Information Security (NIS) 1 Directive and, later, NIS 2 Directive (Directive 2016/1148; Directive 2022/2555; Regulation 2019/881); second, the General Data Protection Regulation (GDPR), a global standard for data protection law (Regulation 2016/679); and third, the AI Act (including the Commission's 2021 Proposal for an Artificial Intelligence Act and the amendments by the European Parliament and of the Council on the Proposal), which is expected to establish global standards for AI (COM/2021/206 final; C/2024/506; Regulation 2024/1689).

[6] For human rights-based approaches to the regulation of AI, see in particular The AIHLEG Ethics Guidelines for Trustworthy AI: High-Level Expert Group on Artificial Intelligence (2019), the Council of Europe principles, ethical guidelines and requirements of human rights, democracy and the rule of law as part of the Feasibility Study on AI: Ad Hoc Committee on Artificial Intelligence (CAHAI) (2020), The UNESCO Recommendation on the Ethics of Artificial Intelligence (2021), The guiding principles of the UN Secretary-General's High-Level Advisory Body on AI: United Nations, Advisory Body on Artificial Intelligence (2023); for others, see G20 AI Principles: G20 (2019), the Governance Principles for a New Generation of Artificial Intelligence: the Government of the People's Republic of China (2019), the GPAI principles: The Global Partnership on Artificial Intelligence (GPAI) (2020), The Blueprint for an AI Bill of Rights: The White House (2022), The Bletchley Declaration: UK Government (2023), G7 Hiroshima Process: G7 (2023), the (updated) OECD AI Principles: OECD (2024), The Seoul Declaration: The Republic of Korea & UK Government (2024).

This thesis assumes that the EU legal frameworks do not yet guarantee adequate protection and promotion of the fundamental rights of individuals and groups concerning the development and public-sector use of AI. In other words, the EU aims to create a digital single market in line with the values and principles of the Charter and the competencies provided under the EU Treaties (Newman, 2020; Savin, 2020; Troitiño, 2022). However, the attempt to combine a permissive and precautionary approach to AI through these regulatory instruments following a risk-based approach has some significant limitations (Laux *et al.*, 2024; Mantelero, 2022; Smuha *et al.*, 2021; Taylor, 2023a; **III**). As argued in **Publication III**, Trustworthy AI can be realised if the four ethical principles of the AI HLEG, *respect for human autonomy, prevention of harm, fairness*, and *explicability*, are translated into *human rights-based* legal requirements.

In the literature on **human rights-based approaches to AI**, several scholars identified the existing human rights frameworks as a starting point for regulating AI, despite their limitation that they were drafted before the uptake of AI (Beduschi, 2020; Kriebitz & Lütge, 2020; Latonero, 2018; Leslie *et al.*, 2021; McGregor *et al.*, 2019; Niklas, 2020; Prabhakaran *et al.*, 2022; Quintavalla & Temperman, 2023; Raso *et al.*, 2018; Sartor, 2020; Smuha, 2020; Yeung *et al.*, 2020). The universal applicability and the institutional system that monitors and enforces human rights law when infringements occur provide an essential foundation for regulating AI (Leslie *et al.*, 2021; McGregor *et al.*, 2019; Yeung, 2019). In the **algorithmic accountability** strand of the literature, some legal scholars focused on assessment mechanisms beyond transparency requirements, elucidating how international human rights law can be applied to evaluate potential human rights harms in the technology's life cycle (McGregor *et al.*, 2019). Others examined how the UN human rights-based approach to development can inform the creation of public policies and laws on AI, particularly concerning the use of AI in welfare administration (Niklas, 2020). Some specifically assessed existing international human rights standards from a corporate responsibility perspective to evaluate how they apply to companies developing AI (Kriebitz & Lütge, 2020; Lane, 2023). However, these discourses tend to give limited attention to the role of *capability approaches* to protecting and empowering individuals and groups concerning the development and public-sector use of AI. Considering that AI is primarily developed by AI developers of tech companies pursuing commercial rather than public interests, capability approaches play a key role in empowering underrepresented stakeholders in AI governance, particularly voices from civil society and individuals from the Global South (Cohen, 2017; Taylor, 2023a, 2023b).

Other strands in the literature, particularly from **data justice and critical data studies**, therefore called on for incorporating an ecosystemic approach to regulating AI based on the capability approaches of Nussbaum and Sen, both globally and locally (Dalton *et al.*, 2016; Taylor, 2017, 2023a; Taylor & Mukiri-Smith, 2021). Reassessing fundamental rights in the context of ICTs and big data, they argued for including the concept of group privacy in the regulatory discussion on AI instead of primarily focusing on protecting individual privacy rights (Taylor *et al.*, 2017a; Taylor *et al.*, 2017b). These approaches have also highlighted the need for incorporating more inclusive, participatory, and context-sensitive requirements into legal frameworks for AI (Dalton *et al.*, 2016; Taylor, 2016, 2017; Taylor *et al.*, 2017a; Taylor & Mukiri-Smith, 2021; Taylor *et al.*, 2017b).

Yet, while these previous discourses attempt to regulate AI indirectly from a socio-technical perspective, aiming to mitigate human rights harms and empower individuals by ensuring transparency, fairness, and accountability through procedural requirements, they pay limited attention to the developers' role and the impact of design and architecture on fundamental rights (Cohen, 2017). Owing to the growing role of AI developers in influencing public values and public infrastructures, developers increasingly impact the fundamental rights of individuals and groups, including their capabilities (Aizenberg & van den Hoven, 2020; Cohen, 2012, 2017, 2019; Friedman & Hendry, 2019; Umbrello, 2022; Umbrello & van de Poel, 2021; Yeung *et al.*, 2020). Regulating AI has also been increasingly identified as an inherently interdisciplinary challenge in legal and computer science discourses (Bex, 2023; Brownsword, 2020, 2022; Graber, 2021; Hoffmann-Riem, 2021; Hydén, 2020; Larsson, 2019; Lindgren & Dignum, 2023; Nyman-Metcalf & Kerikmäe, 2020; Prabhakaran *et al.*, 2022; Rahwan, 2018; Selbst *et al.*, 2019). However, legal scholars have yet to respond to AI developers' calls to bridge disciplines and elaborate common legal requirements for protecting and promoting the fundamental rights of individuals and groups concerning AI (Cohen, 2017, 2019; Prabhakaran *et al.*, 2022; Shahriari & Shahriari, 2017). While calls for adopting socio-technical perspectives on regulating AI are emerging in human rights discourses, they remain limited (Cohen, 2017, 2019; Lindgren & Dignum, 2023; Prabhakaran *et al.*, 2022; Taddeo *et al.*, 2023; Taylor, 2023a; Yeung *et al.*, 2020).

Since the AI HLEG's framework for **Trustworthy AI** is addressed equally to developers, deployers, citizens, and society at large, it laid a foundation for examining how existing human rights principles and standards could be translated into practically applicable legal requirements for AI (High-Level Expert Group on Artificial Intelligence, 2019a, pp. 6-8; Noorman *et al.*, 2019; Smuha, 2019, 2020). Prior research involving legal scholars showed how the AI HLEG's ethical principles relate to existing principles, standards, and mechanisms under international human rights law and the Charter and how they can be translated into new law (Chatila *et al.*, 2021; Mantelero, 2022; Smuha, 2020; Smuha & Morandini, 2022; Smuha *et al.*, 2021; Yeung *et al.*, 2020). Yet, the few concrete human rights-based proposals have remained either underinclusive or overinclusive (Leslie *et al.*, 2021; Mantelero, 2022; Smuha *et al.*, 2021; Yeung *et al.*, 2020). They either focused solely on one specific mechanism, impact assessments (see Mantelero, 2022), or excluded (see Leslie *et al.*, 2021; Smuha *et al.*, 2021; Yeung *et al.*, 2020) an under-researched component, namely participatory design approaches, involving developers, deployers and citizens in the development and assessment of AI systems in the public sector. Additionally, the existing understanding of cybersecurity is primarily grounded in technological solutions (**V**) and preventive and responsive measures to addressing vulnerabilities in AI systems have so far been taken in isolation (**II**). Only a few scholars have conceptualised cybersecurity from a human rights perspective (Deibert, 2018; Papakonstantinou, 2022; Shackelford, 2019; Taddeo, 2019; Taddeo *et al.*, 2023; Taddeo *et al.*, 2019; Von Solms & Van Niekerk, 2013), and they have not yet proposed how Trustworthy AI could be realised.

To address this gap, this **interdisciplinary thesis** attempts to make an initial contribution to the growing body of knowledge on AI regulation by **combining human rights** and **socio-technical perspectives**.

Research questions and outline of the thesis

Consequently, the overarching research question this thesis seeks to address is:

Under what legal conditions can the fundamental rights of individuals and groups be protected and promoted regarding the development and public-sector use of AI systems?

By analysing relevant EU legal frameworks, the main aim is to determine legal conditions necessary and sufficient for protecting and promoting the fundamental rights of individuals and groups concerning the development and public-sector use of AI. To this end, the thesis looks at the following sub-questions:

- What policy and legal measures has the European Union implemented as part of its AI strategy to protect and promote fundamental rights?
- To what extent can the main legal principles and mechanisms under the GDPR address risks to the fundamental right of the protection of privacy concerning the development and use of AI?
- In which aspects does the AI Act fall short in guaranteeing adequate protection and promotion of fundamental rights regarding the development and public-sector use of AI?

Two book chapters (**I**; **II**), two articles (**III**; **V**), and a contribution to a report of Estonia's AI Task Force for developing legal frameworks for AI (**IV**) attempt to provide answers to the research questions. The individual publications address specific aspects of the main research problem.

I and **IV** mapped the EU policy and early legislative actions on AI, responding to the first sub-question. After the Commission released the White Paper on AI in early 2020, it called for a review of applicable EU legal frameworks to assess if they adequately safeguard EU values and principles concerning AI (European Commission, 2020b). Hence, **II** examined the GDPR and pertinent cybersecurity legislation**.** In response to the second sub-question, **II** conducts a technical and legal analysis regarding the implementation of smart metering systems (SM) in residential households. The book chapter identifies end-user concerns in the technical literature and, thereafter, analyses these concerns from a legal perspective. It particularly focuses on machine learning (ML) approaches in SM, since their developers seldom consider privacy or cybersecurity aspects at the development stage of the technology. Yet, as argued in **II**, this can cause human rights harm to end-users of SM at the use stage of the technology. The findings of **V** informed the conceptualisation of cybersecurity in **II**.

Article **V** identifies harms caused by a significant ransomware attack, *WannaCry*, and analyses its implications for emerging EU cybersecurity legislation. The analysis of the identified harms shows a link between data protection and privacy, cybersecurity, and *trust* in ICTs. Going beyond existing approaches to cybersecurity, which primarily focus on technical solutions and exclude the perspectives of the human user of ICTs, **V** suggests securing not only interconnected information systems and networks, including data but also the *aggregate interactions* among human users and information systems and networks in society. Applying this understanding of cybersecurity to the development and public-sector deployment of AI systems, **II** suggests complementing existing requirements under the GDPR by additional mechanisms, particularly fundamental rights impact assessments, before the implementation of SM in residential households.

After the Commission presented the AI Act Proposal in April 2021, the author examined in **III** how the proposal translated the four ethical principles of the AI HLEG into law and identified some significant gaps for realising Trustworthy AI. In response to the third sub-question, these gaps pertain to a lack of obligations for developers and deployers, and substantive and procedural rights for citizens as underrepresented stakeholders in AI governance. **III**, therefore, suggests greater citizen participation in shaping AI regulation. This relates to both *ex-ante* and *ex-post* regulatory mechanisms for AI. **III** particularly recommends introducing effective redress mechanisms for ensuring greater transparency and explicability in AI-informed individual decision-making by deployers. Additionally, complementary to legal requirements, the author proposes introducing institutional measures, such as securing adequate funding for national competent authorities to enable effective monitoring and enforcement of existing legal obligations under the GDPR and upcoming obligations under the AI Act. Moreover, another gap identified is the need to increase civil servants' digital literacy skills. This is crucial for increasing understanding about (yet unknown) AI-specific human rights harms and enabling better assessment of potential harms related to the public-sector use of AI in the long term.

In addressing the research questions, the author follows Cohen's understanding of fundamental rights as *civil liberties*, *capabilities,* and *affordances* (Cohen, 2017, p. 85). Additionally, the thesis treats law as *Law 3.0*, considering law as a normative tool alongside technological instruments (Brownsword, 2016, 2020, 2022) and emphasising that developers of AI significantly influence the ordering function of the law and its legal effects (Cohen, 2017, 2019; Nemitz, 2021; Nemitz & Pfeffer, 2020). Arguing that developers can either constrain or afford the fundamental rights of individuals and groups, the findings in **II-III** show how interdisciplinary, fundamental rights-based *systems thinking* approaches can be applied to protect and promote fundamental rights concerning AI. Article **III** particularly emphasises the introduction of *participatory* systems thinking approaches, involving legislators, policymakers, developers, deployers, and citizens in the development of AI systems and the assessment of the use of the technology in the public sector. Following the value sensitive design (VSD) approach (Friedman & Hendry, 2019, pp. 38-44), this thesis further treats natural persons as either *direct* or *indirect stakeholders*. Whereas the thesis refers to AI developers and deployers as *direct stakeholders* because they either directly develop or use AI technology in their daily work, citizens are *indirect stakeholders*. In comparison to AI developers and deployers, citizens are those individuals and groups that are still excluded from the development of the technology and the assessment of its use in the public sector.

While regulating AI is increasingly viewed as a societal challenge, **I-III** show that existing legal frameworks exclude requirements enabling deployers and citizens to participate in the development or assessment processes of the technology. The decision on how to mitigate potential human rights harms remains at the developers' discretion through self-conformity assessments. Furthermore, existing legal frameworks for AI treat AI primarily as a neutral tool or product. In this regard, the mitigation of potential human rights harms is addressed through legal or technical solutions only. However, as shown in **II** and **III**, this might undermine the fundamental rights of individuals and groups, including concerning their rights to *respect for private and family life*, *the protection of personal data, equal treatment and non-discrimination*, and citizens' right to *good administration*. Considering the information asymmetry among developers, deployers, and citizens, **I-V** show the need for adopting interdisciplinary approaches to regulating AI, including context-sensitive and inclusive legal requirements that allow direct and indirect stakeholders to shape the technological normativity of AI before deploying AI systems in different application contexts in public sectors. Therefore, as stated in **I** and **III**, the debate on AI regulation needs to be guided by the question 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” (**I**, p. 144). By conceptualising “AI as an autonomous digital technology embedded into societal structures and contexts, mediated through digital devices” (**III**, p. 59), the thesis argues that regulating AI requires revisiting existing understandings of AI and adjusting obligations and rights iteratively with the involvement of citizens. In turn, to mitigate potential human rights harms to individuals and groups and promote human rights concerning the development and public-sector use of AI, this thesis proposes five legal conditions and requirements to realise them: *Fundamental rights impact assessments for robust AI*, *effective redress mechanisms*, *socio-technical digital literacy*, *monitoring and enforcement capacity of national supervisory authorities*, *fora for participatory design and the inspection process for Ethical AI*. By examining existing legal frameworks for AI, the proposed conditions and specific requirements aim to promote AI regulation as a “collective responsibility” (Cohen, 2017; Yeung, 2019, p. 70) shared among legislators, policymakers, developers, deployers, and citizens.

The author does not assign any agency or legal personality to AI systems. In line with the AI HLEG, only individuals or organisations developing and deploying AI systems should be held accountable for human rights violations (High-Level Expert Group on Artificial Intelligence, 2019a). The thesis focuses on regulating narrow AI systems, which can be rule-based or knowledge-based (Barocas *et al.*, 2023, pp. 27-28). It does not deal with AI that poses only minimal risk, such as spam filters, nor does the normative analysis extend to AI applications for military purposes. Neither does the author focus on discussions about “existential risk” (Bostrom, 2014) posed by AI, which is important but often overshadow the debate on mitigating potential human rights harms emanating from existing and mundane AI use cases. To this end, the thesis examines two use cases involving the application of two different AI models in the public sector.

The first case in **II** focuses on SM, which facilitates monitoring energy consumption in residential households. **II** particularly analyses the use of non-intrusive load monitoring techniques (NILM) enabled by SM. With AI models in smart grids applied for recommendations and predictions for more efficient use of energy, **II** shows a low awareness among developers of NILM regarding potential human rights harms to citizens. By mapping end-users’ privacy and cybersecurity concerns and conducting a normative analysis of these concerns combining technical and legal perspectives, **II** reveals how developers, deployers, and citizens can pre-emptively address potential human rights harms. **II** also examines the end-users’ concerns from a cybersecurity perspective, as the analysis of the ransomware attack *WannaCry* in **V** demonstrates the importance of understanding the harms of cyberattacks for prioritising the goals, limitations, and scope of legal frameworks for ICTs. The analysis of end-users’ concerns regarding cybersecurity in **II**, such as tampering with the training and input data, reveals the need to introduce legal requirements for cybersecurity in AI to address vulnerabilities of the technology and realise *robust AI*. Additionally, **II** shows that applicable legal mechanisms in the GDPR, such as data protection impact assessments, must be complemented by fundamental rights impact assessments before deploying AI in the public sector.

The second case of this thesis pertains to the use of AI in the administration of welfare related to access to essential public services. The author primarily focuses on the judgment of the Hague District Court in the *Systeem Risicoindicatie* (SyRI) case

(Case C-09-550982, 2020). This case concerned the so-called SyRI system. SyRI was a risk model implemented by government authorities to determine the likelihood of individuals committing welfare fraud and helped inform governmental decision-making on the allocation of welfare benefits in the Netherlands. This was the first time a Court rendered the use of an advanced ICT-based system by the State incompatible with human rights obligations (Alston, 2019, 2020; Case C-09-550982, 2020).

While the Court could not determine the exact functioning of the risk model due to a lack of information by the State, the thesis treats SyRI within the meaning of the AI Act (Regulation 2024/1689), namely as an AI system with varying levels of autonomy whose output generates either a prediction, a recommendation, or a decision. Compared to **II**, the analysis of the *SyRI* case illustrates that the sources for potential human rights harms can lie not only in the AI model but might also emanate from deployers, who purchase an AI system and select the purpose of its deployment. It shows that human cognitive biases can reinforce AI's opacity and bias despite applicable legal frameworks, such as the SyRI legislation. Therefore, to address potential human rights harms, for example, emanating from deploying AI for administrative purposes, the analysis of the *SyRI* case shows the need to include not only legal but also *institutional* requirements for the development and public-sector use of AI. Consequently, the proposed legal conditions in this thesis include both legal and institutional requirements to ensure *external coherence* (Taylor, 2023a, p. 30). External coherence refers here to situations where developers or deployers comply with existing legal obligations, yet where these obligations are insufficient to adequately protect and promote fundamental rights of individuals and groups. This is important, considering that the AI Act primarily relies on *self-conformity assessments* by developers of stand-alone high-risk AI systems and where the prime expertise to monitor and comply with the obligations remains with the developers and owners of the technology.

Both cases reveal the need for implementing mechanisms that enable citizens to participate in joint assessment and development processes in the public sector. As shown in **II** and **III** and further discussed in Chapter 4, these mechanisms are fundamental rights impact assessments, the participatory design approach of VSD, and the co-assessment inspection process for Ethical AI. The recommendations are, therefore, addressed beyond legislators and policymakers to citizens since realising Trustworthy AI requires the participation of citizens in the development and assessment of the technology.

The introductory chapter of this cumulative thesis is organised as follows: Following this chapter, Chapter 2 outlines the methodology and explains the research strategy and methods. Chapter 3 provides the theoretical background of the thesis. To this end, Section 3.1 illustrates the potential impact of public-sector AI applications on human rights, focusing on the core human rights-related normative concern of discrimination, and data protection and privacy and cybersecurity issues. Thereafter, Section 3.2 presents the analytical framework of the thesis. Section 3.3 describes earlier proposals for legal frameworks for AI, including the two emerging schools of thought in AI regulation. Chapter 4 presents and discusses the main contributions and limitations of this thesis. Based on the individual findings in **I-III** and in the appendix papers **IV-V**, the author determines the legal conditions necessary and sufficient for realising Trustworthy AI. The thesis proposes *human rights-based* legal conditions as part of a human rights-based approach to AI and provides a systematic exposition of this approach. In conclusion, Chapter 5 summarises the main results and provides avenues for future research.

2 Methodology

This cumulative thesis in the interdisciplinary fields of law and technology and human rights studies applies qualitative research methods using both deductive (**I-II**; **IV-V**) and inductive methods (**II-III**) (Creswell, 2012; Hervey *et al.*, 2011). To address the research problem, ideas have been developed from particular issues to general concepts (Creswell, 2012). All publications are conceptual and theoretical. They rely on written sources, both primary and secondary, and draw on desk research, literature reviews, and documentary analysis to gain a comprehensive understanding of the broad phenomenon of *protecting and promoting the fundamental rights of individuals and groups regarding the development and public-sector use of AI systems*. For the legal analysis, both primary and secondary law, including EU legal acts, preparatory documents, and case law, have been consulted. Due to the thesis's interdisciplinary nature, the methodological approaches vary.

The thesis is divided into expository (**I-II**; **IV-V**) and evaluative research (**III**) (Hervey *et al.*, 2011). While expository research facilitates exploring how the phenomenon under study relates to the facts of the world and “what” legal options are available to address the legal uncertainties it raises, evaluative research helps to find answers to questions “to what extent” or “why” by critically examining the facts of the world and assessing the adequacy of existing laws in addressing legal uncertainties (Hervey *et al.*, 2011, pp. 9-10). In this regard, the thesis takes an external viewpoint on the law and includes critical perspectives beyond the traditional legal approach. This “law in action” approach considers not only the rules but also the facts, including the impact of power relations on drafting laws and their interpretation in different institutional settings (Ballin, 2020, pp. 51-52, 86 *ff*; Hervey *et al.*, 2011). The research problem has therefore been addressed from socio-legal (**I**; **III**; **V**), technical and legal (**II**), and legal perspectives (**IV**). Table 1 provides an overview of the methodological approaches used in the thesis.

***Table 1**. Overview of methodological approaches (source: author)*

Publication	Level of analysis	Main research question	Research strategy	Methods
I.	AI policy (international/EU) + Legal framework (EU)	What are the European Union and the AI HLEG's objectives in the AI strategy to address ethical and legal challenges posed by AI?	Socio-legal analysis: expository	Desk research, literature review, documentary analysis
II.	Individual + Legal framework (EU)	How does current EU legislation protect the prosumer's data and privacy rights concerning ML-based SM?	Technical and legal analysis: expository Case: ML-based SM	Desk research, literature review, documentary analysis
III.	Concept + Legal framework (EU)	To what extent is the proposed four-dimensional risk-based approach in the AI Act aligned with the Trustworthy AI concept?	Socio-legal analysis: evaluative	Desk research, literature review, documentary analysis
IV.	Project + Legal framework (EU/Country)	Mida välismaal on tehtud? [What has been done abroad?]	Legal analysis: expository	Desk research, literature review, documentary analysis
V.	EU + Legal framework (EU)	What harms do EU cybersecurity-related laws seek to prevent?	Socio-legal analysis: expository Case: ransomware attack *WannaCry*	Desk research, literature review, documentary analysis

Research strategy and method

The initial step involved identifying legal uncertainties and (material and immaterial) harms that potentially result from AI's widespread development and adoption and identifying gaps in existing laws to address these issues. This is a commonly applied approach in law and technology scholarship (Crootof & Ard, 2021). The concept of legal liability, established as a central problem in this field of study, informed the analysis of legal requirements for AI (Pałka & Brożek, 2023; Schrepel, 2023). Legal liability was examined from a human rights perspective to understand how to mitigate potential harms related to the development and public-sector use of AI (Shelton, 2015; Wagner, 2019). This allowed for analysing the phenomenon from not only a legal but also an institutional perspective.

Concerning the expository research sections of the thesis, as part of a **report on legal frameworks for AI** (**IV**), the author analysed various national AI policies, recent legislative changes in selected EU Member States' liability regimes, and the EU AI strategy. For this purpose, the author consulted the EU database EUR-Lex, the Publications Office of the European Union, the Legislative Observatory of the European Parliament, the Public Register of Council documents, and the Google Search database. The initial findings on the objectives of the EU AI strategy informed the empirical section of **Publication I**. The book chapter establishes a general understanding of AI system types and provides an overview of the domains where the technology is applied. The author conducted a literature review on regulatory approaches to ICTs using the Scopus and Google Scholar databases to identify research gaps in the regulatory discussion on AI. This facilitated the examination of theoretical frameworks for regulating ICTs that could be applied to conceptualise normative issues concerning AI and identified Lessig's four modalities of regulation, the VSD approach, and the fundamental rights-based approach of the AI HLEG as part of the tripartite framework of Trustworthy AI (lawful, ethical, robust AI). These frameworks guided the subsequent steps in the thesis. To ensure the protection of fundamental rights concerning AI, the findings in **I** showed the importance of including AI developers in developing legal guidelines for AI. Additionally, it confirmed that one of the main normative uncertainties raised by AI for the protection of fundamental rights concerns the rights to *respect for private and family life* and the *protection of personal data*, as highlighted in several EU documents, including in the AI HLEG's ethics guidelines and policy and investment recommendations for Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019a, 2019b).

Publication II, therefore, looked at how extant EU secondary legislation protects the fundamental rights to *private and family life* and the *protection of personal data* concerning the development and use of AI. The book chapter was written in conjunction with the publication of the White Paper on AI by the Commission and the call therein to review the applicability of existing EU legislation about AI (European Commission, 2020b). The analysis involved evaluating how adequately extant legal principles, rules, and mechanisms enshrined in EU secondary legislation, particularly in EU data protection and cybersecurity laws, can address end-users concerns related to the development and use of a ML-based technology. The study considered the importance of both technical and non-technical methods to establish legal frameworks for AI based on fundamental rights, in line with the objectives in the ethics guidelines for Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019a). To achieve this, legal and technical methods from engineering science were combined, using an interdisciplinary systems thinking approach. Drawing on the technical expertise of engineers in AI and their work on the use of NILM in nearly zero energy buildings at the home university, the authors chose to study the highly privacy-invasive technology of SM.

The Commission made the introduction of SM in residential households mandatory as part of the Clean Energy for All Europeans package and the European Green Deal, seeing SM as crucial to climate change mitigation efforts, especially when combined with ML techniques. However, using the technology could severely affect human agency and human dignity. The publication first reviewed the technical literature to identify end-users' concerns about the development and use of SM in residential households. By end-users, we referred to electricity prosumers treated in the thesis as citizens and data subjects. We categorised these concerns based on data protection and privacy and cybersecurity and considered them to be generalisable to other AI applications. We then analysed the end-users' concerns from a legal perspective, taking a positivist, doctrinal approach while adopting a critical perspective on the law (Ballin, 2020; Hervey *et al.*, 2011). This involved identifying the legal provisions about the end-users' concerns and then interpreting how adequately the relevant legal norms and mechanisms can address them. We used a textual and teleological approach to treaty interpretation to evaluate the applicability of the EU's secondary law in this regard, considering the overarching principles and objectives stipulated in the primary law of the EU Treaties (Lenaerts & Gutiérrez-Fons, 2013). Based on the combined technical and legal analysis, we inductively developed a tool as part of a table that aims to enhance the protection of fundamental rights to *private and family life* and *personal data* by providing developers, deployers, and citizens guidance for either developing or using AI.

Publication III is part of the evaluative research section of the thesis. After the Commission presented the proposal for an Artificial Intelligence Act in April 2021, the article examined how adequately the legal requirements in the proposal, including the four-dimensional risk-based approach, protect and promote citizens' fundamental rights. Following the methods used in law and technology scholarship (Crootof & Ard, 2021), the author identified a preferred legal approach to AI that is precautionary. This allowed for the reassessment of the EU legal framework for AI. The author collected information through desk research, a review, and a documentary analysis of AI governance literature in the EU using the Scopus database, EU database EUR-Lex, and Google's site search function. The article looked at the four ethical principles and seven requirements in the AI HLEG framework for Trustworthy AI and evaluated their translation into legal rights and obligations in the proposal (lawful AI). This helped determine the proposal's shortcomings. It showed that the proposal did not fully implement the fundamental rights-based approach of the AI HLEG and its principles of the *prevention of harm, respect for human autonomy, fairness,* and *explicability*.

When protecting civil liberties and empowering citizens, one of the main proposed amendments in the article was to add substantive and procedural rights for citizens, including but not limited to effective redress mechanisms. Additionally, the article identified the need to include provisions that make access to educational programs in AI mandatory by law to increase digital literacy and, thus, citizen's capabilities. Based on the experience with enforcing the GDPR, this would also benefit future public officials' monitoring of AI use cases when implementing AI laws. Thirdly, since *ex-ante* self-conformity assessments for high-risk AI applications give developers significant discretion and decision-making authority regarding AI standards and values, the paper calls for incorporating mechanisms into legal frameworks for AI that make public participation in the future shaping of AI systems mandatory. This includes rights for public participation in developing AI, particularly in revising high-risk AI applications and creating public registries of AI use cases. This would enable the implementation, enforcement, and monitoring of legal frameworks for AI based on fundamental rights, as envisioned in the human rights-based Trustworthy AI framework. **III**, therefore, argued that the Commission opted for a *permissive* approach, which does not adequately protect and promote fundamental rights.

Publication V, listed in the appendix and part of the expository research section, studied a prominent ransomware attack, *WannaCry,* which significantly impacted the EU cybersecurity strategy and legislative action in this area. The discussion paper aimed to determine the types of harm caused and provide legal guidance on the scope and goals of emerging EU cybersecurity laws during the EU's initial legislative actions on cybersecurity. We collected information via the EU database EUR-Lex and the Google Search database using Google's site search function. This included EU legal acts, preparatory and other EU documents such as transcripts of speeches, and commentary by cybersecurity experts and policymakers in the media and selected blogs (Financial Times, The New York Times, The Washington Post; The AI Blog) due to the recent nature of the attack. The findings informed the conceptualisation of cybersecurity in **II**.

Based on the individual findings in **I-V**, the thesis proposes human rights-based legal conditions for protecting and promoting the fundamental rights of individuals and groups concerning the development and public-sector use of AI as part of a human rights-based approach to regulating AI. This approach is precautionary, participatory, bottom-up and complementary to the risk-based approach in the AI Act. As illustrated by Figure 1, the conditions are derived from the AI HLEG framework for Trustworthy AI, particularly the four ethical principles contained in the theoretical model, and extant legal sources, namely the legal principles and mechanisms within the GDPR and pertinent EU cybersecurity legislation. Additionally, the author evaluated the AI Act, including amendments to the Commission's Proposal during the ordinary legislative procedure before the publication of the AI Act in the *Official Journal of the European Union*. The conditions aim to respond to the normative concerns raised by AI as identified in **II** and the harms mapped in **V**.

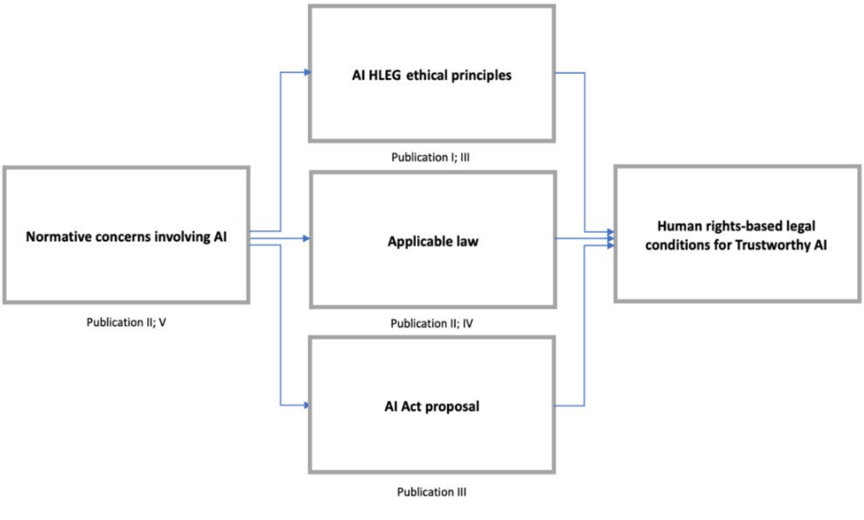

AI HLEG ethical principles
Publication I; III
Normative concerns involving AI
Publication II; V
Applicable law
Publication II; IV
Human rights-based legal conditions for Trustworthy AI
AI Act proposal
Publication III

Figure 1*. Illustration of the development of the human rights-based legal conditions for Trustworthy AI (source: author)*

***Table 2**. Explanation of Figure 1 (source: author)*

Description	**Explanation**
Normative concerns involving AI	End-user concerns identified in **II** and harms identified regarding the ransomware attack *WannaCry* in **V**
AI HLEG ethical principles	Respect for human autonomy, the prevention of harm, fairness, and explicability
Applicable law	Cybersecurity Act, NIS 1/NIS 2 Directives, Electricity Directive, GDPR
AI Act proposal	Up until the European Parliament legislative resolution of 13 March 2024
Human rights-based legal conditions for Trustworthy AI	Legal and institutional requirements for Trustworthy AI

Throughout the doctoral studies, the author analysed various aspects of the phenomenon under study. This also included regulating the development and use of AI for military purposes. In this regard, the author focused on how to establish meaningful human control over the critical functions of autonomous weapons systems. The author has not submitted these findings for publication but has used some of them as preparatory materials for co-taught courses at the home university. The feedback from the Ragnar Nurkse Department doctoral school seminar participants, especially the emphasis on sustainability in regulating ICTs, influenced **Publication III** and was integrated into its revised version.

3 Theoretical background

This chapter provides the theoretical background of the thesis. First, it examines the impact of public-sector AI applications on human rights, focusing on the core human rights-related normative concern of *discrimination*, and *data protection and privacy* and *cybersecurity* issues raised by AI. Two AI use cases further illustrate the impact: the case of *SyRI*, which refers to the use of AI in the administration of welfare, and the case of SM, which is related to the use of AI in the monitoring of energy consumption in residential households. Subsequently, the author presents the analytical framework of the thesis. The chapter then outlines existing legal proposals for AI regulation, including the two main emerging regulatory approaches to AI.

3.1 The impact of public-sector AI applications on human rights

The use of both rule- and learning-based AI in the public sector has increased significantly across Europe in recent years (Busuioc, 2021; Chiusi *et al.*, 2020; Kaun *et al.*, 2023; Redden *et al.*, 2022; Tangi *et al.*, 2022; Wolswinkel, 2022). AI is used in fields such as predictive justice, predictive policing, or the allocation of welfare benefits using risk indicators and impacting access to public services. Quantifiable, data-driven decision-making represents an emerging mode of control over and coordination in society with numerous implications for human agency in the relationship between the State as the duty bearer and the citizens as rights-holders under the rule of law (Hildebrandt, 2018; Hoffmann-Riem, 2020, 2021; Murray, 2021; Nemitz, 2021; Nemitz & Pfeffer, 2020; Yeung, 2018). The use of AI in the public sector is generally expected to result in more efficient and effective delivery of public services (Misuraca & van Noordt, 2020). Additionally, it may reduce error rates and bias, save time and resources, increase transparency, and improve accessibility and availability of public services (Coglianese & Lehr, 2017; Wirtz *et al.*, 2021; Zuiderwijk *et al.*, 2021). However, AI can also negatively impact people's agency and freedoms when they are subjected to an administrative decision, either informed or automated by an AI system (Alston, 2019; Jørgensen, 2023; Kaun, 2022; A/74/493, 2019; Wendehorst, 2020; Wieringa, 2023). In the public sector, this is crucial because the government usually has exclusive authority over public decision-making processes and must act under the rule of law while respecting, protecting, and fulfilling fundamental rights as ratified in international treaties and enshrined in its constitution (Greenstein, 2022; Nyman-Metcalf & Kerikmäe, 2020).

AI's main functionality is its ability to replicate human intelligence, including reasoning and decision-making, and to assist or replace humans by (partially) automating a diverse range of tasks through software running on substantial amounts of personal and non-personal data (Russell & Norvig, 2010). In this regard, distinguishing between the forms of AI's autonomy is helpful. Barocas *et al.* (2023) outline three forms of automation. The first concerns translating decision-making rules agreed upon by humans in policies into software (*Ibid.*). In this case, human decision-making is aided or automated by rule-based AI. The second relates to replicating informal judgments of humans and translating them into software (*Ibid.*). Thirdly, AI systems can establish new rules based on data for decision-making processes without relying on human-made, pre-existing rules (*Ibid.*). The latter two cases involve ML-based AI. Whereas ML-based AI can achieve better performance, accuracy, and effectiveness than rule-based systems, rule-based AI is less opaque and more transparent (Ebers, 2020; Waltl & Vogl, 2018).

However, all three forms can negatively impact human agency and the legal interests of individuals and groups (Barocas *et al.*, 2023; Niklas, 2020).

This thesis argues that if developed and deployed within a legal framework and institutional system that adequately protects and promotes fundamental rights, the potential of AI systems for the tasks they are designed for, such as facilitating or automating detection, prediction, and monitoring processes in areas ranging from energy to healthcare, law enforcement, employment, social welfare administration or the administration of justice and democratic processes, could be increased, and potential human rights harms to individuals and groups mitigated (Hoffmann-Riem, 2021; Nemitz, 2018; Smuha, 2020); **I-V**). However, realising adequate protection and promotion of fundamental rights concerning the development and public-sector use of AI is a complex challenge.

Delegating tasks traditionally assumed by humans to novel technologies is not new. It does not necessarily have to raise normative concerns that reach the fundamental rights realm and material scope. However, the ethical and normative implications of delegating tasks from humans to AI systems are unprecedented due to AI's human-designed capabilities for inferential statistics, data-matching, and profiling (A/HRC/38/35, 2018). Furthermore, they are unparalleled because AI is opaque, may exhibit bias, operates with partial autonomy, and is highly intrusive while widely deployed in society by private and public entities (Hoffmann-Riem, 2021; Yeung, 2019). Additionally, the speed of decision-making, the potential for the scalability of AI and the resulting complexity of the "black box" technology are unexampled (Burrell, 2016; Loi & Spielkamp, 2021; Pasquale, 2015; Yeung, 2019). Although environmental considerations have played only a minor role in the debate on regulating AI, considering them is at least equally important. As evidenced, energy-intensive computational methods and the increasing extraction of resources for their development are often neglected side effects of the green and digital transition (Santarius *et al.*, 2023; van Wynsberghe, 2021).

Consequently, AI raises several human rights-related normative concerns, including *discrimination*, *data protection and privacy*, *cybersecurity*, and *environmental concerns* (Yeung, 2019; **I**; **II**; **V**). With AI's potential to "continually and immanently mediate and pre-empt our beliefs and choices" (Cohen, 2017, p. 89), AI even presents significant challenges to the existing idea of fundamental rights and its underlying principle of human dignity (Cohen, 2019; Hildebrandt, 2015; Murray, 2021; Nemitz, 2018).

In turn, the introduction of AI has implications for the realisation and protection of all fundamental rights, encompassing civil and political, socio-economic, collective, absolute, and derogable rights, and the principles of legality, necessity, and proportionality, whether understood as being codified in treaties or considered as moral entitlements from an ethical standpoint (Cohen, 2017; High-Level Expert Group on Artificial Intelligence, 2019a; Yeung, 2019)[7]. This raises questions about handling the technology, including its

[7] See also the references to the fundamental rights enshrined in the Charter that the AI Act (Regulation 2024/1689, recital 48) seeks to protect: The "right to human dignity, respect for private and family life, protection of personal data, freedom of expression and information, freedom of assembly and of association, the right to non-discrimination, the right to education, consumer protection, workers' rights, the rights of persons with disabilities, gender equality, intellectual property rights, the right to an effective remedy and to a fair trial, the right of defence and the presumption of innocence, and the right to good administration, […] [the rights of the child] enshrined in Article 24 of the Charter and in the United Nations Convention on the Rights of the Child, further developed in the UNCRC General Comment No 25 as regards the digital environment,

potential misuse by malicious actors. Regulatory approaches to AI, the objective of which is to promote economic interests over the protection of fundamental rights, might not achieve their intended goals. If the goal is to attain Trustworthy AI related to the development and public-sector use of AI, then legislative frameworks for AI need to include requirements that prioritise long-term protection of human values, the environment, and the public interest over short-term economic interests (Nemitz, 2021; Taylor, 2023a; Whittaker, 2021; **III**).

The impact of public-sector AI applications on human rights can become clearer when examining the functionalities and internal logic of AI systems. Normative concerns primarily arise from AI's reliance on detecting patterns in data and correlating these data points to make either recommendations for human decisions or automate decision-making (Ebers, 2020; Hildebrandt, 2018; Hoffmann-Riem, 2021; Mayer-Schönberger & Cukier, 2013). The main problem with this statistical reasoning is that civil servants might rely on an AI system's output, a recommendation or prediction, as the sole basis for decision-making with legal implications for individuals and groups (Laux *et al.*, 2024; Niklas, 2020). In this regard, civil servants tend to treat AI systems as neutral technology (Niklas, 2020; Wagner, 2019). This can lead them to unreflectively trust the output of AI or result in a limited likelihood for them to override an AI system, known as "automation bias" (Green, 2022, p. 7; Laux, 2023; Niklas, 2020; Wagner, 2019). Yet, the quality of the data input impacts the output, as encapsulated by the wording "garbage in, garbage out" (Anastasopoulos & Whitford, 2019, p. 506; Kitchin, 2017). If the information and data on individuals and groups used for AI-informed decisions are inaccurate, it can result in biased decisions and discrimination (Crawford, 2013; Niklas, 2020). In turn, this can lead to human rights harms and potential violations of both substantive and procedural safeguards for individuals and groups, impacting, for example, the rights to *respect for private and family life*, *the protection of personal data*, *non-discrimination* or the right to *good administration* (Alston, 2019; Barocas *et al.*, 2023; Niklas, 2020; A/HRC/38/35, 2018; A/74/493, 2019). The likelihood of human rights harms can increase due to factors such as weak information on the functioning of the system, poor data handling, weak technical expertise, time constraints, poor working conditions, low motivation or strong hierarchies with little agency for public servants to rectify visible mistakes (Green, 2022; Laux, 2023; Niklas, 2020; Wagner, 2019). Understanding contextual factors concerning the development and public-sector deployment of AI is, therefore, crucial to mitigating discrimination and potential human rights harms.

The next sub-section investigates the human rights-related issue of discrimination in more detail. Thereafter, the issues of data protection and privacy and cybersecurity are addressed.

3.1.1 Discrimination

Bias or discrimination is a key normative concern about AI (Barocas *et al.*, 2023; Hacker, 2018; Lehr & Ohm, 2017; Martínez-Ramil, 2022; Zuiderveen Borgesius, 2020); **I**; **II**). From a technical standpoint, discrimination can be intentional or unintentional (Barocas, 2014). Unintentional discrimination often occurs due to statistical bias or inaccurate inferences from avoidable errors in the data mining process (Barocas, 2014; Barocas *et al.*, 2023). From a legal perspective, discrimination can be direct or indirect (Hacker, 2018;

[…] the fundamental right to a high level of environmental protection enshrined in the Charter and implemented in Union policies […]."

Zuiderveen Borgesius, 2020). To prove indirect discrimination, the crucial factor is not the intention but the impact of AI-informed decision-making on the individual or group (Zuiderveen Borgesius, 2020). This type of unintended discrimination occurs hidden (*Ibid*.). Still, it can have the same serious outcomes, causing primarily immaterial, psychological harm, depending on the type of service the affected individual or group expected to receive (Yeung, 2022). Proving that discrimination occurred concerning AI is highly difficult for affected persons (Yeung, 2019; Zuiderveen Borgesius, 2020). Often, an individual is unaware that an AI system has directly or indirectly discriminated against them based on race or ethnicity or indirectly through proxy discrimination, for example, based on location (Martínez-Ramil, 2022; Zuiderveen Borgesius, 2020). This can occur because individuals do not know that AI is deployed on them. If individuals become aware, they have hurdles to prove causation between the model's logic and its effects on them. Additionally, individuals often waive their rights when entering into contractual relationships or obligations by consenting to conditions that allow AI companies to, for example, claim trade secrecy and intellectual property rights (Yeung, 2019).[8] Furthermore, public authorities often withhold information from citizens about risk indicators and the algorithm's functioning for reasons that include "gaming the system", for example, in welfare allocation contexts (Case C-09-550982, 2020; Rachovitsa & Johann, 2022).

The right to information not only about the data input but also about the data output of AI, namely how public authorities further process personal data, is, therefore, crucial in mitigating potential discrimination (Wachter & Mittelstadt, 2019; Wachter *et al.*, 2017a). Consequently, legal frameworks for AI need to include requirements that mandate public authorities to provide explanations to citizens about how administrative decisions based on risk indicators, whether recommended or solely taken by AI, have been made to the extent that the overall functioning of the system is preserved. However, these requirements, which concern increasing transparency and explicability in calculating risk indicators with legal effect for individuals, can only be considered necessary but not sufficient mitigation measures (Ananny & Crawford, 2018; Busuioc, 2021). This is primarily but not exclusively due to the *opacity of AI*.

The opaque nature of AI, also known as the "black box" problem, makes detecting and mitigating discrimination challenging (Pasquale, 2015). According to Burrell (2016), opacity can entail an intentional dimension, namely the interests of organisations and governments in withholding information from citizens for corporate or State secrecy reasons (for example, national security concerns) (Burrell, 2016). Secondly, opacity can arise due to technical illiteracy (*Ibid.*). Third, opacity can result from the features of AI's complexity and, hence, the difficulty of deploying AI in a targeted way (*Ibid.*). In the second type, both the developers and deployers of the technology may not fully understand the inference process, the step between the input data and the output (Ebers, 2020; Kitchin, 2017). This lack of understanding can result in discriminating against specific individuals or groups and requires the introduction of measures to

[8] See also the broad exceptions to Article 22 in Article 22(2) of the GDPR (Regulation 2016/679), which present hurdles for citizens to address issues of discrimination and bias relying on extant data protection legislation: "Paragraph 1 shall not apply if the decision: is necessary for entering into, or performance of, a contract between the data subject and a data controller; is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or is based on the data subject's explicit consent."

increase the socio-technical digital literacy of developers, deployers and citizens (Burrell, 2016; Edwards, 2017-2018; Hasselbalch, 2021; Pasquale, 2015; **III**).

When considering explicability and transparency requirements as mitigation measures against potential discrimination, opacity can impact the quality of the explanations provided to individuals. Hence, depending on the context in which the technology is deployed, beyond the requirement for public authorities to explain to individuals which risk indicators were most influential for the final decision-making process, it is necessary that deployers also provide information on the systems' limitations (High-Level Expert Group on Artificial Intelligence, 2019a; **III**). One argument is that legal requirements for developers to create explainable models for AI should be weighed against the economic benefits of achieving transparency and explicability (Buiten, 2019). While financial considerations are important, especially from the perspective of small and medium-sized enterprises developing AI systems, when AI is used in sensitive application areas in the public sector such as in law enforcement or welfare contexts, financial costs should not pose obstacles for developing more explainable AI models and achieving greater transparency in its deployment (Yeung, 2022).

Moreover, while *human oversight* (Green, 2022; Laux, 2023), including its main approaches of "human-in-the-loop", "human-on-the-loop" and "human-in-command" (High-Level Expert Group on Artificial Intelligence, 2019a, p. 16), is considered a key measure to mitigate potential human rights harms, including discriminatory outcomes, a requirement for establishing effective human oversight can only be treated as necessary but not sufficient against the backdrop of the problem of opacity. Establishing effective control over the development and deployment of AI in public sectors, therefore, requires adopting complementary mechanisms. Yet even the possibility of contesting an AI-informed decision presents an insufficient mitigation mechanism. As shown above, existing legal safeguards provided by Article 22 of the GDPR are limited regarding AI (Niklas, 2020). Additionally, remedies such as a right to lodge a complaint or a right to receive an explanation on individual decision-making from deployers would not address other factors such as potential structural or organisational deficiencies, particularly where an organisational culture of "rubber-stamping" persists (Wagner, 2019, pp. 117-118).

The SyRI case

The analysis of the *SyRI* case exemplifies the human rights-related normative concern of discrimination (Case C-09-550982, 2020). For several years, the risk model SyRI was deployed only in disadvantaged neighbourhoods without citizens' knowledge. Its use was enabled by data aggregation and matching among public authorities, including tax and police authorities. The risk model has been refined and modified over the years based on previous implementations of similar systems. Additionally, the SyRI legislation[9] adopted by the Dutch Parliament in 2014 legitimised its deployment only until the Hague District Court ruled that it conflicted with Article 8(2) of the European Convention of Human Rights (ECHR) (Case C-09-550982, 2020).

The *SyRI* case shows that correlation does not mean causation and that the knowledge AI-informed consequential decisions are based on might cause immaterial individual and

[9] *SyRI legislation* here refers to the *Work and Income Implementation Organisation Structure Act* (*Wet structuur uitvoeringsorganisatie werk en inkomen*) (*SUWI Act*) as adopted under the SUWI Decree: Case C-09-550982, 2020.

collective harm in violation of applicable human rights standards (Yeung, 2022; **III**). Additionally, it highlights the importance of adopting legal and institutional requirements for AI as part of a precautionary rather than a permissive approach to AI regulation (Alston, 2020; Case C-09-550982, 2020; Yeung, 2022). While the detection of welfare fraud is in line with ethical principles, the deployment of AI in only disadvantaged districts but not others and the way the algorithm was designed to restrict access to social benefits demonstrated the potential of AI for differential treatment and discrimination causing both material and immaterial harms to a generally vulnerable group of people as individual and collective rights-holders (Alston, 2019; Chiusi *et al.*, 2020; Gantchev, 2019; Rachovitsa & Johann, 2022; A/74/493, 2019; van Bekkum & Borgesius, 2021). Whereas the State claimed to have used a rule-based system that primarily relied on decision trees, the Court eschewed classifying the system as an automated decision-making system within the scope of Article 22 of the GDPR due to a lack of information by the State about the exact functioning of the system (Case C-09-550982, 2020, *paras.* 6.47, 6.60, 6.90). Yet, based on the SyRI legislation, the Court still determined that SyRI's risk model could be based on deep learning and involve data mining to create risk profiles and generate risk reports on individuals receiving welfare benefits by linking governmental databases (Case C-09-550982, 2020, *paras.* 6.53, 6.63, 6.93, 6.102).

The *SyRI* case reveals the need to address gaps in the law and introduce institutional requirements for the development and public-sector use of types of AI systems that can inform administrative decision-making with legal effects for the individuals affected. It reveals how the lack of transparency in the deployment of SyRI, the deliberate choice of deploying the system primarily in disadvantaged areas, and weak communication between affected persons and the deployers of SyRI can impact fundamental rights beyond the right to *respect for private and family life* (Case C-09-550982, 2020, *paras.* 6.91-6.95). Additionally, since neither the risk indicators nor the functioning of the risk model was made public and existing legal remedies under the GDPR proved insufficient to provide information rights to affected persons related to the data output of the system, the case illustrates the need for effective redress mechanisms, improved monitoring of fundamental rights by national supervisory authorities and mechanisms that can increase the participation of affected persons in addressing potential discrimination. Even though the Court primarily relied on Article 8 of the ECHR and the right to *respect for private and family life, home and correspondence*, the ramifications of the deployment of SyRI show that the implementation of AI for administrative purposes can affect multiple civil-political and socio-economic rights at once, including people's *human dignity and autonomy*, the right to *non-discrimination*, the right to an *effective remedy*, and the right to *access to social security* in similar future applications (Case C-09-550982, 2020, *paras.* 6.91-6.95).

3.1.2 Data protection and privacy and cybersecurity issues

In addition to discrimination, data protection and privacy issues have been among the most highlighted normative concerns about AI, following up on earlier debates on regulating digital technologies (Citron & Solove, 2022; Etzioni, 2007; Hoffmann-Riem, 2021; Kriebitz & Lütge, 2020; Solove, 2024; Wachter & Mittelstadt, 2019; Wachter *et al.*, 2017a; Wachter *et al.*, 2017b, 2020; Zuiderveen Borgesius, 2020). Privacy is another form of protection of human dignity (Kriebitz & Lütge, 2020). The scarce yet emerging case law on AI systems in the public sector has revealed that Courts primarily rely on the right to *respect for private and family life* to restrict overly intrusive algorithmic decision-making based on the processing of personal data and require that substantive and procedural adjustments related to the data handling, management and communication be made (Case C-09-550982, 2020). Privacy concerns of citizens are also mentioned as the most crucial factors for governments to consider before implementing AI for public service delivery (Kleizen *et al.*, 2023). They primarily originate from AI's functionalities and techniques of tracing and inferring information about individual and group behaviour and characteristics from personal data, precisely the use of AI in privacy-sensitive areas such as residential households (**II**).

As revealed in **II**, end-users' privacy and cybersecurity concerns include the denial of access to services, the exhibition of user habits and lifestyle, the exhibition of illnesses and disabilities, the identification of home appliances, denial of personal mobility and discrimination. **V** shows that privacy and cybersecurity concerns are intricately linked and can impact people's fundamental rights and trust in advanced ICTs. Therefore, legal frameworks for AI focusing on the dimension of cybersecurity should secure both *interconnected information systems*, including data, information systems and networks, and the *aggregate interactions among human users, society and these information systems and networks* (**II**; **V**, p. 21; Papakonstantinou, 2022; Von Solms & Van Niekerk, 2013).

To perform its various functions in the public sector, AI relies on the provision of personal data by individuals (Busuioc, 2021). The principle of consent governs the protection of personal data and privacy, including data collection, between individuals and public or private entities (Yeung, 2019). Consent is the prime legitimate cause for public authorities' processing of personal data and sharing it with third parties, including the principles of lawfulness and data minimisation (Yeung, 2019). Previous case law, however, has shown that third parties might use individuals' personal data for purposes other than those initially envisaged by the individual, such as for commercial and data analytics practices and personalised price offers based on consumer profiles (Espinosa Apráez, 2022; **II**). Therefore, processing personal data remotely using AI without informed consent from the individual can threaten the fundamental rights to *respect for private and family life, home and correspondence,* and *the protection of personal data* (Kriebitz & Lütge, 2020).

Furthermore, AI shifts information rights from the individual to the collective, particularly by using the technology primarily for monitoring and predictive purposes based on a sample of a general population rather than an individual (Taylor *et al.*, 2017a; Taylor *et al.*, 2017b; Yeung, 2019). However, inferring information from a group's sample creates significant challenges for extant data protection regulations, which focus on protecting individual rather than collective rights (Yeung, 2019; Zuiderveen Borgesius, 2020).

In extreme cases, by sharing data with the State and third parties, the State can match databases and aggregate the data with sensitive AI applications such as remote biometric systems for law enforcement purposes. This has occurred in authoritarian States, particularly against ethnic minorities and political opponents (Case 11519/20, 2023; Kriebitz & Lütge, 2020; Niklas, 2020). To prevent these extreme cases of surveillance practices enabled by AI, Kriebitz and Lütge (2020) propose three rules: Data transfers from enterprises to public authorities should only be permitted if all parties provide consent in the context of asymmetric and dynamically changing power relations, consented data processing can only be applied for purposes aimed at reducing harms to other actors (harm principle), and the use of privacy-intrusive AI solutions should be weighed against the proportionality principle.

Consequently, AI and the data processing that enables its use can threaten the fundamental right to the *protection of personal data* in at least three aspects: First, public and private authorities can process data without the informed consent of the individual. Secondly, individuals might waive their right by consent for convenience. Thirdly, the inferential analytics afforded by AI enables inferring sensitive information about individuals and groups. However, extant data protection legislation also has limitations in protecting the privacy of groups as it primarily focuses on the information rights of the individual (Taylor *et al.*, 2017a; Taylor *et al.*, 2017b; Zuiderveen Borgesius, 2020). Additionally, as shown in the previous sub-sections, the opacity of AI complicates the understanding of how the fundamental rights to *respect for private and family life* and the right to the *protection of personal data* might have been breached (Niklas, 2020).

The case of smart metering systems

At the beginning of the debate on regulating AI, threats to the fundamental right to the *protection of personal data* were addressed only from a legal perspective (**II**). Yet, applying solely legal approaches to mitigate human rights harms concerning the development and deployment of AI has proven unsuccessful (see, for example, Hoffmann-Riem, 2021; Prabhakaran *et al.*, 2022). As shown in **II** as part of a case study on ML methods for monitoring energy consumption in residential households through SM, this has implications for data protection legislation effectiveness and enforcement concerning the use of AI. From a technical perspective, developers of ML techniques are rarely aware of end-users' concerns and potential threats to fundamental rights the development of AI can cause to individuals and groups (**II**). Additionally, developers are often unaware of their legal obligations, thereunder the responsibility to conduct a risk impact assessment before the implementation of AI in critical areas where sensitive data might be collected. Furthermore, legal guidelines for developers were lacking on how AI can be developed so that the deployment of AI is compatible with the fundamental rights to *respect for private and family life*, the *protection of personal data,* and existing cybersecurity obligations (**II**).

Hence, **II** addressed these technical and legal challenges from an interdisciplinary perspective. The regulation of SM is a widely discussed topic in the data protection literature (Cuijpers & Koops, 2013; Espinosa Apráez, 2022; Huhta, 2020; Lavrijssen *et al.*, 2022) since its functionalities enable remote real-time reading of energy data in residential households, unlike conventional meters. On the one hand, SM enables participation in generating renewable energy, more efficient home energy management predictions, and personalised price offerings (**II**). On the other hand, SM functionalities allow for fine-grained profiling of citizens based on personal data and information about electricity consumption in their private homes. Using ML techniques, particularly NILM techniques, enhances this functionality, yet developers of this technique often neglect privacy concerns (**II**). To pre-emptively address potential privacy harms related to the implementation of SM in public and private sectors, the publication proposes a visual tool in the form of a table that combines both the technical and legal analysis of end-user concerns. Before the Commission's Proposal for an AI Act, it aimed to guide developers on their legal obligations related to developing and deploying AI and inform citizens about their rights when breaches of applicable data protection standards occur. For the development of legal frameworks for AI, **II** suggested further research into the principles and requirements of the AI HLEG, mainly how the principles and requirements for Trustworthy AI can inform existing data protection impact assessment mechanisms.

More broadly, empirical evidence suggests that public authorities use digital solutions to monitor the electricity consumption of welfare recipients (Alston, 2019; A/74/493, 2019). Combined with the use of AI for administrative decision-making, the implementation of SM could incentivise deployers to use energy consumption data collected through SM in the long term. While the use of AI systems for detecting fraud or preventing criminal offences in general presents a legitimate public reason, the *SyRI* case shows how the deployment of the technology in residential households for these purposes can impact the fundamental right to *respect for private and family life* and the *protection of personal data* and lead to significant human rights harms to affected persons, particularly vulnerable groups in society (Alston, 2019; A/74/493, 2019; Yeung, 2022). Therefore, the combined use of SM and ML methods for administrative decision-making purposes poses significant challenges to the privacy rights of individuals and groups, and consideration of whether its combined use should be prohibited is important.

3.1.3 Summary

This section aimed to illustrate the current impact of public-sector AI applications on human rights and outlined existing mechanisms to address potential bias and discrimination, data protection and privacy, and cybersecurity issues raised by the development and public-sector use of AI.

Compared to the case of SM, the analysis of the *SyRI* case reveals a need to introduce not only legal but also institutional requirements for the development and public-sector use of AI. These requirements need to address potential human rights harms emanating not only from the AI model but also from the decision-making processes of deployers. Otherwise, AI systems might reinforce hidden cognitive biases, impacting citizens' trust towards both AI systems and deployers. This necessitates including participatory mechanisms that can account for power in the development and assessment of AI systems in the public sector. Additionally, the analysis of the *SyRI* case underscores the importance of national supervisory authorities to mitigate potential human rights harms by deployers and enable effective redress to affected persons. To realise Trustworthy AI, the capacity of national supervisory authorities in the monitoring of AI use cases and the enforcement of fundamental rights must be enhanced (**III**).

As argued in this thesis, primarily the developers of AI systems and their infrastructure impact the protection and enjoyment of fundamental rights of individuals and groups concerning AI technology. Therefore, an important requirement to mitigate potential human rights harms is to involve AI developers in interdisciplinary, participatory design processes. As further highlighted in **III**, these should be based on **critical, participatory systems thinking approaches of VSD and the inspection process for Ethical AI** (Umbrello, 2021, 2022; Umbrello & van de Poel, 2021; Zicari *et al.*, 2021a; Zicari *et al.*, 2021b). These approaches are intended to increase developers, deployers, and citizens' awareness of the **socio-technicity of AI**, the importance of design choices in AI innovation, and the long-term impact of designing context-specific values in AI *for* citizens (**III**). In practice, this requires that direct and indirect stakeholders together design and assess AI systems such as ML-based SM before their deployment in public sectors.

3.2 Foundations of human rights protection concerning AI

This section defines *technology*, *law*, and *human rights*, and outlines the analytical framework of the thesis.

3.2.1 Defining technology, law, and human rights

The conceptualisation of **technology** in this interdisciplinary study of law and technology and human rights, or “how law shapes technology and how technology shapes the law” (Crootof & Ard, 2021, p. 348), has been informed by a post-phenomenological approach. In this regard, this thesis treats AI as a technology that “co-determine[s]” human agency and “mediate[s] the intentional relation between humans and world” (Ihde, 2009; Verbeek, 2005, p. 116), or, further, “as an autonomous digital technology embedded into societal structures and contexts, mediated through digital devices” (**III**, p. 59). This understanding of AI highlights its socio-technicity, particularly how the design of AI and value choices can impact the protection and promotion of fundamental rights during the use of the technology. However, it also considers that an AI system has its own heuristics (Hoffmann, 2020), which impacts human oversight over the technology since AI can create (long-term) outcomes unintended by its designers (Cohen, 2017).

Therefore, human beings can shape technology, but technology also shapes human beings (*Ibid.*). This assumption emphasises the importance of incorporating holistic, iterative approaches to regulating AI, based on fundamental rights, which is also essential for enforcing AI laws and monitoring new AI use cases in the long term. Additionally, it implies that an AI system is “neither good nor bad; nor is it neutral” (Kranzberg, 1986, p. 545). This is because equally to the law, politics and power can influence the development and use of AI systems, requiring critical perspectives that extend beyond the traditional legal approach (Taylor, 2023a).

Law, drafted and promulgated by lawmakers, interpreted and practised by lawyers and judges, and consulted by individuals and groups, is commonly defined as “legal rules, together with the related institutions, decisions, principles and values expressed in normative language that is intended to guide human behaviour” (Ballin, 2020, p. 13). Considering the introduction of AI systems, the growing power of digital platforms and systemic socio-technical configuration by AI developers, however, traditional understandings of law need re-evaluation. Against this background, the thesis adopts Brownsword’s concept of law as *Law 3.0*, which views law as a normative tool alongside technological instruments (Brownsword, 2016, 2020, 2022). Governments increasingly use non-traditional technology regulations alongside the law to achieve specific regulatory objectives, particularly for law enforcement purposes (Brownsword, 2020; Yeung, 2022). Therefore, advances in technology have made technology both the object and the subject of regulation, also referenced by Lessig under the concept of “code is law” and by Brownsword as *technological management* (Brownsword, 2020; Hydén, 2020; Lessig, 1999, 2006), or more prominently in the field of regulatory governance studies as *algorithmic regulation* (Ulbricht & Yeung, 2022; Yeung, 2018, 2022). This understanding of law aligns with Cohen’s conceptualisation of fundamental rights (Cohen, 2017), highlighting the need to emphasise more the role of AI developers, who influence the ordering function of the law and its legal effects through design by embedding their values into technology (Nemitz, 2021; Nemitz & Pfeffer, 2020). Legal approaches to the regulation of AI, including the field of *law and technology* in general, are still emerging. Understanding law as *Law 3.0* requires adopting interdisciplinary, fundamental rights-based systems thinking approaches to the regulation of AI to realise Trustworthy

AI in the long term. The analytical framework of this thesis, therefore, combines legal perspectives, human rights theory and science and technology studies (STS), taking a human rights-based, socio-technical perspective on AI regulation.

While understandings of **fundamental rights** vary, human rights are most commonly treated as ethical demands or legal claims on the State and society by individuals to shape and constrain power (Henkin, 1990; Rodley, 2014; Sen, 2004). In the hierarchy of norms, national constitutions accord human rights the most prominent position (Katz & Sander, 2019; Kriebitz & Lütge, 2020). The subject matter of these rights is, therefore, contingent on certain threshold conditions, as only those rights that are of "special importance" or of "social influenceability" can receive this status (Kerikmäe & Nyman-Metcalf, 2012; Sartor, 2020; Sen, 2004). They are considered a mechanism for articulating and realising varied understandings of human dignity (Donnelly, 2013) or for protecting individuals' distinct types of civil liberties from State interference (Henkin, 1990; Kerikmäe *et al.*, 2016; Kerikmäe & Nyman-Metcalf, 2012). Others conceptualised fundamental rights as *institutions of society* and *institutions of the law* (Luhmann, 1965). As institutions of society, human rights provide continuous protection against external factors that can affect the autonomy of the individual and the community. The impact of these factors can be made explicit, for example, through technological advancements. As institutions of the law, fundamental rights are essential when translating novel ethical claims for protecting and promoting the freedom of the individual and groups into law (Luhmann, 1965, as cited in Graber, 2017, p. 224). In light of the unequal access to public goods and unequal distribution of resources, especially when comparing the Global North with the Global South, human rights discourses have also broadened to encompass issues of development and capabilities (Nussbaum, 2011; Sen, 2004).

While often divided into the categories of first-generation rights, which create negative obligations on the State not to intervene in the civil and political rights of individuals, second-generation rights, demanding that the State create better economic, social and cultural conditions for individuals, and third-generation rights, pertaining to communal rights such as solidarity and the right to development or a healthier environment (Vasak, 1977), this thesis further understands fundamental rights as "universal, indivisible and interdependent and interrelated" (A/CONF.157/23, 1993, p. 5). It implies that the State, as the traditional duty bearer, has the responsibility to "respect" (negative obligation), "protect" (positive obligation) and "fulfil" (positive obligation) the human rights of rights-holders so that minorities and vulnerable groups in society enjoy the same rights as the majority (Noorman *et al.*, 2019). The rationale for protecting the interests of minorities and vulnerable groups and empowering them becomes even more important in more sensitive application areas of AI systems, including in law enforcement contexts or the administration of welfare (Joamets & Chochia, 2021; Noorman *et al.*, 2019). Furthermore, as the UN Human Rights Council affirmed that "the same rights that people have offline must also be protected online" (A/HRC/RES/20/8, 2012; A/HRC/38/35, 2018, p. 2), it is essential to examine *how* human rights can be ensured at both individual and collective levels regarding the development and public-sector use of AI.

3.2.2 Human rights as civil liberties, capabilities, and affordances

Whereas the State traditionally holds the monopoly over physical force for protecting individuals or groups from each other (Weber, 2010), private entities as third parties, particularly digital platforms as developers of AI and owners of some of the infrastructure it relies upon, have increasingly challenged this role (Jasanoff, 2016; Morozov & Bria, 2018; Yeung, 2019; Zuboff, 2019). This has varied implications for the protection and enjoyment of fundamental rights. Empirical evidence has shown that previous conceptualisations of human rights can hardly address human rights harms concerning ICTs alone, enabled by private entities' development and use of AI (Yeung, 2019). Both civil rights and capabilities discourses predate the implementation of big data and digital technologies, particularly AI. In an era of ubiquitous computing with "smart digital technologies […] continually, immanently mediating and pre-empting our beliefs and choices" (Cohen, 2017, p. 89; Hildebrandt, 2015), Cohen (2017), therefore, advocated for extending earlier conceptualisations of fundamental rights into affordances.

To protect fundamental rights in this context, the theoretical perspective of **human rights as affordances** highlights the importance of integrating socio-technical considerations into fundamental rights theory and legal practice. Other approaches, particularly from data justice discourses, have advanced the capability approach of Sen, inquiring into how opportunity freedoms and process freedoms can be understood and realised concerning big data and ICTs (Taylor, 2017). While these approaches advocate for additional information rights for individuals and groups and particularly emphasise education on data use and participation in political decision-making over standards on ICTs, they focus less on the impact of architecture and the role of developers (Cohen, 2017; Taylor, 2017). Therefore, complementary to conceptualising fundamental rights as civil liberties and capabilities, the lens of fundamental rights as affordances helps to inquire into how the developer, mainly through the design of AI systems and its accompanying infrastructure, can either limit or enhance the protection and enjoyment of fundamental rights of individuals and groups regarding the development and public-sector use of AI in the first place (**I-III**). This is important, considering that AI has an inherent technological normativity that impacts human values, including the intellectual and physical freedom of human beings (Brownsword, 2019; Calo, 2022; Hydén, 2021). For instance, the development and use of SM for climate mitigation efforts could be fundamental rights compliant if the design of the technology and the infrastructure it relies on allows for the protection and enjoyment of fundamental rights (**II**). In more concrete practices concerning exercising the right to *privacy*, existing understandings of the fundamental right to *privacy* focus on the libertarian, individualistic concept of consent. However, theory and practice have shown that people often waive their right to the *protection of personal data* by consent due to a lack of knowledge, low transparency on how their personal data are processed, or for general convenience (Yeung, 2019), demonstrating that previous fundamental rights discourses cannot mitigate potential human rights harms posed by AI (Cohen, 2017; Taylor *et al.*, 2017a). Consequently, addressing individual freedoms such as *privacy* is a necessary condition but insufficient to protect and promote fundamental rights concerning AI.

Following Cohen (2017, 2019, 2023), we need to re-evaluate the material and operational conditions shaped by the process of systemic socio-technical configuration by the developers of AI, to ensure that the implementation of AI systems by humans is compatible with the fundamental rights of individuals and groups. This involves establishing new substantive and procedural human rights standards for the development and public-sector use of AI to enable human beings to actively shape and constrain technological normativity and resulting values iteratively (**III**; Drechsler & Kostakis, 2014; Taylor, 2023a). This is important because current legal frameworks for AI primarily give technology owners and developers the authority to determine AI standards, with limited incentives for them to involve other vital domain-specific experts from the public sector (such as doctors) with agency and expertise in establishing AI standards and monitoring AI practices under fundamental rights law (Vetter *et al.*, 2022; Zicari *et al.*, 2021a; Zicari *et al.*, 2021b; **III**).

Consequently, the human rights-based approach of this thesis rests on three pillars: the protection of civil liberties, particularly the right to effective redress (Henkin, 1990), the promotion of capabilities, focusing on the right to education on AI and access to lifelong learning on AI (Sen, 2004; Taylor, 2017), and the advancement of the right to participation concerning the development and assessment of AI systems in the public sector (Cohen, 2017). The third element, which rests on understanding fundamental rights as affordances, pertains to creating legal and institutional requirements that allow citizens to "co-determine" how values, principles and rules are "hardwired" into AI systems (Cohen, 2017, p. 87).

Participation in these processes is a critical factor in why citizens accept the use of AI systems, particularly in public services (Gesk & Leyer, 2022). This is not to be confused with the overly optimistic views on using AI for processes of participation to create value, which, in the context of previous digital technologies, has been found to involve risks for processes of co-creation and co-production in providing public services (Lember *et al.*, 2019). To translate human values, principles, and rules, operationalise them via technical and legal design requirements and realise Trustworthy AI, this thesis proposes introducing participatory design approaches based on the VSD method and the inspection process for Ethical AI (Friedman & Hendry, 2019; Zicari *et al.*, 2021b; **III**). This rationale aligns with previous accounts on the impact of architecture and design on the exercise of human agency. For instance, Lessig's four modalities of regulation showed how the normative effects of a developer's choices in designing technology could either constrain or reinforce an individual's choices (Lessig, 1999; **I**). Additionally, the lens of fundamental rights as affordances is consistent with VSD theory as it equally calls for employing and combining legal approaches with systems thinking approaches from engineers and technologists to operationalise human values, human rights standards and principles in practice (Aizenberg & van den Hoven, 2020; Cohen, 2017; Friedman *et al.*, 2021; Friedman & Hendry, 2019; Umbrello, 2022; **III**). Protecting and empowering individuals and groups in the age of AI requires the introduction of legal and institutional requirements enabling citizens and domain-specific experts to participate in the development of AI systems and their assessments in the public sector by "reimagin[ing] the linkages between information flows and human freedom" (Cohen, 2017, p. 87; **III**).

To concretise this human rights-based approach to AI, a second part of the analytical framework of the thesis is the four ethical principles and seven requirements in the AI HLEG framework for Trustworthy AI. Trustworthy AI can be understood as a new concept in human rights theory involving an interdisciplinary, holistic, systems thinking approach to AI development and use (Chatila *et al.*, 2021; Smuha, 2020; Smuha *et al.*, 2021). Specifically, it can viewed as a new concept for protecting human autonomy concerning the development and use of AI as a response to previous calls for creating a "counterweight to the pervasive surveillance and asymmetry of power which now confronts the individual" and the collective (De Gregorio, 2021; European Data Protection Supervisor, 2015, p. 12; Hasselbalch, 2021). To steer AI development and uptake in both public and private sectors for societal good, Trustworthy AI aims to address the main pitfalls of AI, namely its potential to "unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans" while promoting "responsible competitiveness" (High-Level Expert Group on Artificial Intelligence, 2019a, pp. 5, 12; **I**).

In this vein, the AI HLEG ethical principles help to concretise fundamental rights as civil liberties, capabilities, and affordances. Therefore, to determine the legal conditions necessary and sufficient for protecting and promoting the fundamental rights of individuals and groups concerning the development and public-sector use of AI, this thesis adopts an analytical framework **combining fundamental rights as civil liberties, capabilities, and affordances with the ethical principles of the AI HLEG**. Following the AI HLEG (2019), the analytical framework of this thesis is further grounded in applicable human rights law, particularly the Charter, which EU institutions and Member States are obliged to adhere to when implementing EU law (Charter, 2016; Kerikmäe, 2014; Schima, 2015; Seubert & Becker, 2021). Additionally, it is rooted in Article 2 of the Treaty of the European Union (TEU) (EU Treaty, 2016).

The following sub-section examines the role of the Trustworthy AI framework in the development of legal frameworks for AI in the context of the EU AI strategy.

3.2.3 Trustworthy AI: The human rights-based approach of the AI HLEG

The economic and societal promises of AI and its risks to fundamental rights stimulated a debate among EU policymakers on how to approach AI (European Commission, 2018, 2020b; Floridi *et al.*, 2018; High-Level Expert Group on Artificial Intelligence, 2019a, 2019b); **I**; **IV**). In a global AI competition between the USA and China (Kerikmäe & Pärn-Lee, 2021; Smuha, 2021b), the European Council acknowledged in 2017 the importance of a coordinated European approach to AI aimed at establishing a competitive advantage in AI innovation in the EU while upholding strong data protection standards and safeguarding digital rights and ethical values (European Council, 2017). It, therefore, requested the Commission to “put forward a European approach to artificial intelligence” (European Council, 2017, p. 7). After the Tallinn Digital Summit 2017, the Commission developed an AI strategy (European Commission, 2018; European Group on Ethics in Science and New Technologies, 2018). The AI strategy identified three goals, including the need to establish an ethical and legal framework for AI that aligns with the Union’s values and the Charter, as advised by the European Group on Ethics and Science in New Technologies (European Commission, 2018; European Group on Ethics in Science and New Technologies, 2018; **IV**). The EU AI strategy considers the creation of an “ecosystem of trust” and an “ecosystem of excellence” (European Commission, 2020b, p. 3). While the former is rooted in applied ethics to develop a human-centric, fundamental rights-based approach, the latter focuses on responsible investment, innovation, and implementation of AI (European Commission, 2020b, 2021; **I**). More broadly, the EU’s approach to AI combines the development of the digital single market with protecting and empowering people (De Gregorio, 2021; Newman, 2020; Troitiño, 2022; **I**). However, this poses challenges to adequately safeguard fundamental rights and empower citizens concerning the development and public-sector use of AI (De Gregorio, 2021; **III**). This is also visible in the AI Act, which represents an innovation-inspired framework tilting towards the creation of an ecosystem of excellence rather than an ecosystem of trust with economic and human values remaining in tension (**III**; Smuha *et al.*, 2021). Considering that the AI HLEG’s four ethical principles and seven requirements for Trustworthy AI are grounded in existing international human rights law standards and the Charter, **III** shows that the AI Act Proposal only partially translated them into legal requirements (Laux *et al.*, 2024; COM/2021/206 final; Smuha *et al.*, 2021).

To inform the implementation of the EU AI strategy, the Commission mandated a European expert group composed of stakeholders from industry and academia, including human rights scholars, to create AI-specific ethics guidelines and policy recommendations (High-Level Expert Group on Artificial Intelligence, 2019a, 2019b). As part of this multistakeholder approach, the AI HLEG developed the concept of Trustworthy AI and proposed a three-part framework to achieve it (Chatila *et al.*, 2021; Smuha, 2019; Thiebes *et al.*, 2021; **I**). To be deemed trustworthy, AI should be:

- *lawful* and in compliance with all extant laws and regulatory frameworks,
- *ethical* and thus safeguard and promote ethical principles and values in democratic societies, and
- *robust*, ensuring both technical and social robustness, especially considering AI's potential for unintended harm (High-Level Expert Group on Artificial Intelligence, 2019a, p. 5).

The fundamental rights-based approach of the AI HLEG consists of four core ethical principles (*respect for human autonomy* as enshrined in Articles 1 and 6 of the Charter, *prevention of harm* as stipulated in Article 3 of the Charter, *fairness* as codified in Article 21 of the Charter, and *explicability* as derived from Article 47 of the Charter) and seven interdependent ethical and technical requirements (*human agency and oversight*; *technical robustness and safety*; *privacy and data governance*; *transparency*; *diversity, non-discrimination, and fairness; societal and environmental well-being*; *accountability*), and is a response to the array of normative concerns raised by the development and use of AI systems (Chatila *et al.*, 2021). Each component and requirement is essential, but they only represent necessary conditions, not sufficient ones, for achieving Trustworthy AI (Chatila *et al.*, 2021; High-Level Expert Group on Artificial Intelligence, 2019a). To make these principles and requirements practical for developers, deployers and citizens, the AI HLEG developed the assessment list for Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2020).

The element of *trustworthiness* is considered a pre-condition for the uptake of AI systems for societal good since previous technologies such as planes, nuclear power plants or, medical products and foods could not be used in society if not trusted by the people involved and impacted by them (European Commission, 2022; High-Level Expert Group on Artificial Intelligence, 2019a, p. 5; Thiebes *et al.*, 2021). Additionally, achieving individual and collective trust in the actors and processes regarding the development and use of AI systems is essential when considering its use in the public sector, where public trust in government has been identified as a critical value in public administration theory (Hood, 1991). A high degree of trust in government has been linked to improved quality and acceptability of public service delivery, increased willingness for public participation, more active civil engagement in political affairs, lower corruption and crime rates, and enhanced economic development (Bjørnskov, 2018; van Ingen & Bekkers, 2015, as cited in Laux *et al.*, 2024, p. 3). Therefore, as a new concept in human rights theory, realising Trustworthy AI requires creating conditions for trust in AI technologies and continuously evaluating AI's widespread adoption in society against the four ethical principles of the AI HLEG (**III**).

Furthermore, the need for Trustworthy AI is rooted in the complexity of AI and the risks and opportunities AI holds for human beings (Chatila *et al.*, 2021; Thiebes *et al.*, 2021). The primary purpose of Trustworthy AI can, therefore, be understood as managing and reducing the complexity that arises from the introduction of the "autonomous, adaptive, and interactive" socio-technical system of AI into wider society (**III**; Luhmann, 2014; van De Poel, 2020, p. 400). This involves establishing trust in the developers, deployers and public institutions monitoring the development and use of AI systems.

This can be achieved by establishing *human rights-based* legal and institutional requirements for the development and public-sector use of AI, combining human rights and socio-technical perspectives. These requirements would mandate developers to develop AI in line with them, deployers to use AI following them, and citizens to participate in their implementation and be empowered by them (**I-III**; High-Level Expert Group on Artificial Intelligence, 2019a). The realisation of Trustworthy AI requires an inclusive multistakeholder approach, based on interdisciplinary, participatory systems thinking, including design- and ethics-based approaches (**I-V**; Friedman & Hendry, 2019; Zicari *et al.*, 2021b). Conceptualising fundamental rights as *civil liberties, capabilities* and *affordances* and linking these three elements with the AI HLEG framework for Trustworthy AI to operationalise fundamental rights regarding the development and deployment of AI closes the gap in existing human rights theory concerning AI regulation. In other words, it enables translating the AI HLEG ethical principles into new legal requirements specific to AI from a human rights-based perspective that is sensitive to socio-technical issues that AI raises. As shown in **II** and **III**, the translation of these ethical principles can be achieved through a combination of technical and legal methods, particularly based on the structured systems thinking approaches of VSD and the inspection process for Ethical AI (Aizenberg & van den Hoven, 2020; Cohen, 2017; Friedman & Hendry, 2019; van den Hoven, 2017).

Trustworthy AI in this thesis is further understood in line with the AI HLEG as achieving "the trustworthiness of all processes and actors that are part of the system's lifecycle" and not the AI system as such (High-Level Expert Group on Artificial Intelligence, 2019a, p. 38; Smuha, 2021b). In this regard, Trustworthy AI implies that only humans behind the corporations and the AI system and not the technology itself can be held accountable for violations of fundamental rights (**I**; **IV**). Therefore, regulating AI is related to the question of who will design AI systems, for what purpose, who owns and deploys them, and in which contexts AI will be applied (**I**; **III**). Earlier criticisms that humans should trust the developers and deployers of AI and its institutions only, rather than AI itself, hold general value (Bryson, 2018; Freiman, 2022). However, the criticism does not directly apply to the core understanding of the concept as such and can, therefore, be rebutted (Smuha, 2021b). A more relevant criticism, as mentioned by Laux (2023) and Rieder *et al.* (2021), is to place greater focus on the aspects of distrust when establishing legal and institutional requirements for Trustworthy AI. This is relevant if one considers that citizens could unreflectively trust and uncritically accept the widespread adoption of AI systems despite the technology's earlier discussed technical errors, limitations, and potential human rights harms. Additionally, the critique that citizens may perceive trust in AI as externally and technocratically "engineered" (Laux *et al.*, 2024, p. 7) is crucial. This is important to consider since realising Trustworthy AI is a *long-term process* (Kleizen *et al.*, 2023; Laux *et al.*, 2024; **III**). It requires mechanisms that enable the participation of citizens in AI regulation by "thinking in and through language and practice to reimagine the linkages between information flows and human freedom" (Cohen, 2017, p. 87; Laux *et al.*, 2024). Realising Trustworthy AI in the public sector requires thinking beyond a "checkbox mentality" (Kleizen *et al.*, 2023, p. 11), gradually addressing citizens' concerns and implementing adequate safeguards for individuals and groups concerning AI.

3.3 Legal options for regulating AI

This section outlines earlier proposals for legal frameworks for AI and presents the two main emerging schools of thought on AI regulation: risk- and rights-based approaches. Furthermore, it includes the contribution this thesis seeks to make, in particular, to the latter.

3.3.1 Proposals for legal frameworks

Regulating AI is related to the initial attempts of governments to regulate the Internet (Black & Murray, 2019). To unlock the economic potential of the Internet, governments embraced a market-oriented approach, refraining from intervening in the market and advocating for self-regulation (Savin, 2020). This initial approach eventually led to the formation of monopolies of digital platforms known as *GAFAM*[10] and has recently been reconsidered by the EU as part of the EU AI strategy (European Commission, 2018, 2020a, 2020b, 2021). These digital platforms are nowadays the primary owners and developers or providers of AI systems representing with their enormous economic and knowledge power in our information society one of the two immediate norm addressees in the regulatory debate on AI (Floridi, 2020b; Nemitz & Pfeffer, 2020). The economic success of big data analytics in e-commerce settings and the promise for greater efficiency in the context of austerity programmes after the 2008 fiscal crisis were crucial drivers for governments in implementing similar solutions for regulatory purposes in the delivery of public services (Yeung, 2018, 2022). Therefore, public authorities as deployers of AI represent the second norm addressee in AI regulation.

Achieving adequate protection and promotion of fundamental rights concerning the development and public-sector use of AI is a complex problem (Chatila *et al.*, 2021; Smuha, 2020; Smuha *et al.*, 2021; Yeung, 2019). Initially, international civil society organisations, including AI researchers, practitioners, and ethics committees, identified the need to take regulatory actions on AI, proposing several ethics principles to improve the explicability, transparency, oversight and broader accountability of AI (Fjeld *et al.*, 2020; Floridi *et al.*, 2018; Russell *et al.*, 2015). As prerequisites for the development of legal frameworks for AI, some of these initiatives took direct inspiration from the values and principles enshrined in existing international and European human rights frameworks, data protection legislation and the field of bioethics (European Group on Ethics in Science and New Technologies, 2018; Fjeld *et al.*, 2020).[11]

At the beginning of the debate on **legal frameworks for AI**, proposals were made to extend extant data protection laws and introduce a right to reasonable inferences (Wachter & Mittelstadt, 2019). Arguing that individuals require meaningful protection

[10] GAFAM stands for Google (Alphabet), Amazon, Facebook (now Meta), Apple, and Microsoft: the Big Tech companies.

[11] See, for instance, the *Asilomar AI Principles* by the Future of Life Institute (2017); *the IEEE's (Institute of Electrical and Electronics Engineers) Ethically Aligned Design Principles*: The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2016, 2019); the co-created principles by the University of Montreal as part of the *Declaration for a Responsible Development of Artificial Intelligence*: Université de Montréal (2018); the *Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems* developed by Amnesty International and Access Now and grounded in the International Human Rights Framework: Amnesty International & Access Now (2018); the *Universal Guidelines for AI* published along the Global Privacy Assembly Conference held at the European Parliament under the auspices of Giovanni Buttarelli: The Public Voice (2018).

against not only the inputs but "the outputs of data processing", these calls aimed to address the lack of safeguards and remedies in AI-informed decision-making, focusing on increasing transparency and explicability of AI (Wachter & Mittelstadt, 2019, p. 579; Wachter *et al.*, 2017a). Even though the yet scarce case law (Case C-634/21, 2023; Case C-09-550982, 2020) confirmed that existing principles and requirements for transparency stipulated in the GDPR indeed place obligations on deployers of AI, the case law also showed that fulfilling the principle of transparency alone would not mitigate the spectrum of challenges posed by the introduction of AI into society (van Bekkum & Borgesius, 2021). Deployers could easily adjust the transparency parameters of AI to comply with the Court's current interpretation of transparency (*Ibid*.). In turn, an AI system could be deemed transparent but remain discriminatory and opaque due to other factors, such as the context in which the technology is used or insufficient oversight and monitoring measures (Niklas, 2020; Smuha *et al.*, 2021; **II-III**).

Furthermore, while the recent judgment by the Court of Justice of the European Union (CJEU) provided clarification on the material scope of GDPR Article 22 regarding "Automated individual decision-making, including profiling" by stating that the automated processing of personal data for creating "risk scores" in public and private entities' use of AI systems is prohibited (Case C-634/21, 2023; Opinion of Advocate General Pikamäe, 2023; Silveira, 2023), relying solely on CJEU rulings and national case law related to the GDPR may not be sufficient to address the normative concerns AI raises. This also applies when looking at other intrinsic limits of the GDPR, particularly how personal data or information is conceptualised in the legal instrument (Gellert, 2022). As Gellert (2022) argues, the GDPR's conceptualisation of information can best be understood as *knowledge communication*. Yet, in the context of learning-based systems, particularly where profiling is applied beyond individuals to groups for AI-informed decision-making purposes, the regulatory target should not be *data processing per se* but the outcome of the learning process of AI. In other words, the processing of personal data and information by AI systems ought to be rather understood within the context of *knowledge production* since AI systems continuously improve by the personal data of individuals and groups (Gellert, 2022). Moreover, developers' lack of awareness about potential privacy harms an AI model can pose to individuals and groups and about legal obligations to potentially mitigate them present additional significant challenges (**II**).

As a result of these factors, GDPR Article 22 provides only limited safeguards for data subjects concerning AI, specifically as to the explicability and transparency of the processing of personal data. The consequence is a need to extend applicable legal frameworks and introduce legal and institutional requirements specific to AI (Cohen, 2017, 2019; Taylor, 2023a; Taylor *et al.*, 2017a; **III**).

Several legal scholars have advocated for creating a new legal regime for AI, calling for a "robust, holistic and coherent" regulatory approach that mandates *ex-ante* and *ex-post* requirements towards developers and deployers of AI systems (Black & Murray, 2019, p. 16; Murray, 2021; Nemitz, 2018, 2021; Nemitz & Pfeffer, 2020). While Black and Murray (2019) called on legislators to set red lines for the development of AI, which cannot meet *ex-ante* transparency requirements, they also conceded that transparency requirements are not sufficient to prevent harm in the context of increasing private ordering and the resulting power asymmetries for people. This also concerns how far obligations for disclosure of AI-informed decision-making go and how these disclosure requirements are designed so that they do not create (additional) burdens for citizens but genuinely inform them about how deployers make AI-informed decisions based on their personal data (*Ibid.*). Some assessed the usefulness of transparency requirements, including against the financial costs they might impose on companies and developers of AI systems (Buiten, 2019; Reed, 2018). These strands in the scholarship also contended that introducing transparency measures needs to be evaluated for their potential impact on the accuracy of algorithms, as excessive transparency requirements might reduce accuracy (see, for example, Buiten, 2019).

Drawing comparisons with the regulation of pharmaceuticals and high-frequency trading regulation, Martini (2020) argued for introducing new regulatory modalities of preventive regulation, risk management tools and *ex-post* measures such as algorithmic responsibility codes. Similarly, Ebers (2020) advanced the idea that AI should be regulated based on an innovation-friendly approach, that would follow a co-regulatory, multi-level, risk-based approach and include elements such as product licensing, auditing mechanisms, regulatory sandboxes and data-based experimental legislation. Likewise, Wendehorst (2020) asserted that a risk-based approach presents an appropriate response to the physical and human rights-related risks AI can pose to citizens. While necessary yet not sufficient, mandatory impact assessments have been proposed for developers and deployers to assess the implications of AI systems on fundamental rights before the use of the technology (Malgieri & Pasquale, 2024; Mantelero, 2018, 2022; Moss *et al.*, 2021).

Others focused on establishing new institutions to coordinate governance and legislative efforts on AI at both national and international levels, ensure oversight of AI, enable effective enforcement of AI regulation and monitoring of AI use cases (Chesterman, 2021; Cohen, 2019; Floridi *et al.*, 2018; Scherer, 2016). On national levels, for instance, Scherer (2016) proposed the creation of an agency as part of an Artificial Intelligence Development Act with an expert committee to evaluate and certify the safety of an AI system and establish procedural mechanisms for experimental AI testing. This proposal aligns with the AI4People initiative, which recommended assessing national and international institutional capacities, developing AI-focused educational programs, implementing redress mechanisms, and appointing AI ombudspersons (Floridi *et al.*, 2018). As part of Estonia's AI strategy, the expert group suggested creating a registry of robots and providing new competencies to existing national institutions, particularly the Technical Regulatory Authority (**IV**). On the international level, Chesterman (2021) suggested forming an International Artificial Intelligence Agency to coordinate efforts on AI for the benefit of humanity globally. Maas and Villalobos's (2023) mapping of 43 initiatives for international institutional development in the field of AI attests to the overall complexity of regulating AI and emphasises the need for transnational, interdisciplinary and interinstitutional collaboration on AI regulation.

In addition to the previous proposals, Zech (2021, 2023) mentioned the idea of the development of a social insurance fund to remedy harm *ex-post*. This partially aligns with earlier ideas voiced by Estonia's AI Task Force calling for the development of a national insurance fund for AI (**IV**). The primary function of the fund would be to compensate for damage caused by public entities' use of AI, which is covered by the insurance based on a previously agreed list with third parties (developers of AI and insurance companies). On the one hand, this would spare the injured party from proving individual causation and make determining the traditional burden of proof for strict liability unnecessary. On the other hand, it risks lowering the incentive for providers and deployers to comply with extant legal requirements if damages could be financially compensated. At this initial phase in AI regulation, which types of damages the fund could address is unclear, mainly whether it could compensate for immaterial harm. Additionally, an insurance fund as part of economic compensation might divert attention from the necessity to create legal frameworks for AI, iteratively adjust legal requirements over the long term and enable citizens to meaningfully participate in AI regulation (Noorman *et al.*, 2019).

3.3.2 Two emerging schools of thought: Risk- and rights-based approaches

Against this background, prior research on the regulation of AI can be divided into two schools of thought. While both aim to mitigate risks concerning the development and use of AI through the concept of liability, they diverge on the method to this end, both substantively and procedurally. Whereas the first pleads for a preventive top-down risk-based approach based on AI auditing mechanisms and self-conformity assessments for developers of AI with liability primarily limited to the category of high-risk AI systems (Buiten, 2019; Calo, 2017; Ebers, 2020; Martini, 2020; Reed, 2018; Scherer, 2016; Wendehorst, 2020), the second calls for the adoption of a bottom-up approach.

In contrast to the **risk-based approach**, which often invokes a pro-innovation stance, the **rights-based approach** prioritises public participation and deliberation, introduces legal safeguards for citizens through mandatory fundamental rights impact assessments for developers and deployers of AI systems, and considers internationally applicable and nationally enforceable human rights standards at the development, standard-setting, monitoring and enforcement stages (Bria, 2017; Latonero, 2018; Mantelero, 2022; McGregor *et al.*, 2019; Morozov & Bria, 2018; Nemitz, 2018; Niklas, 2020; Raso *et al.*, 2018; Smuha *et al.*, 2021; Yeung *et al.*, 2020). Moreover, as highlighted in **III**, a human rights-based approach to AI treats individuals and groups not only as *consumers* but primarily as *citizens*, particularly as participatory citizens in AI regulation, where human values prevail over economic values.

Mantelero (2022) adds the perspective of a **principle-based approach** rooted in ethics and international human rights principles. Even though this is not a legal approach *per se*, it belongs to the second school of thought. However, an approach to AI regulation based on ethics alone lacks normativity and the legally binding nature of the law for the State to force the providers of AI systems as the owners of the technology into compliance (A/HRC/38/35, 2018). Additionally, it does not provide legal certainty for developers and deployers, including effective redress mechanisms for affected persons. Another problem with ethics approaches alone is the risk of “ethics-washing” and “ethics-shopping” (Wagner, 2018, p. 4). Specifically, the AI industry might claim to adhere to ethics guidelines and yet engage in practices that undermine the fundamental rights of individuals and groups. This is also why existing ethics-based auditing approaches to AI (Mökander & Floridi, 2021) might not serve the purpose of containing human rights harms (AI Now Institute, 2023).

Complementary to the first and second schools of thought are **design-based approaches**, focusing on “hardwiring” values, principles and rules into AI systems through a combination of technical and legal design requirements (Aizenberg & van den Hoven, 2020; Friedman & Hendry, 2019; Hildebrandt, 2020; van De Poel, 2020; van den Hoven, 2017). However, more than technological fixes through design requirements alone are necessary to adequately protect and promote the fundamental rights of individuals and groups. It could even perpetuate the problem of opacity and bias with AI if the decision to design the technology remains at the developers’ discretion.

AI has an inherent technological normativity, implying that its code has a regulatory impact on individuals (Brownsword, 2019; Graber, 2017; Hydén, 2021). This impact is similar to the influence of legal norms and the coordinating and ordering role of the law in society (Hydén, 2020, 2021). Therefore, participatory approaches to the development of AI, including deliberation over the requirements of AI systems' design, are increasingly essential for regulating AI (**III**). This is particularly important, as coding often occurs in research clusters driven by commercial interests and in closed environments (Umbrello, 2022; **III**). This impacts the design of AI, including the value choices embedded in the AI system (Umbrello, 2022; **III**).

The codification of the law alone is not a panacea either, as risks remain, such as that actors with significant financial resources, including the AI industry, might manipulate the legal system by prolonging litigation until claimants potentially give up (Schrems, 2014). The inclusion of substantive and procedural rights, in particular effective redress for affected persons, and the potential creation of new epistemic rights, such as the right to information, are, therefore, critical goals in human rights discourses on AI regulation (Risse, 2021; Smuha *et al.*, 2021; Wachter & Mittelstadt, 2019; Yeung *et al.*, 2020). These issues are not specific to AI *per se* and have always existed, but they are becoming more apparent in the age of AI.

3.3.3 Summary

Both the risk- and rights-based approaches attempt to address risks concerning the development and use of AI pre-emptively. However, human rights scholarship has yet to provide a practically applicable alternative to the existing risk-based approach in the AI Act. This leaves open the criticisms that rights-based approaches are too abstract, impose unnecessary regulatory burdens, and might hinder innovation (McGregor *et al.*, 2019; Yeung *et al.*, 2020). The thesis argues that the absence of an alternative regulatory approach to the risk-based approach in the AI Act presents a significant weakness for realising Trustworthy AI in the long term. Even though no empirical evidence currently exists to determine whether a rights- or risk-based approach is more beneficial for society at this early stage in the adoption of AI, a one-size-fits-all legal framework for AI does not exist (Chesterman, 2021; Ebers, 2020; Hoffmann-Riem, 2021; Nyman-Metcalf & Kerikmäe, 2020). However, this thesis tilts towards a rights-based approach in a regulatory field characterised by many "unknown unknowns" (Floridi, 2020c, p. 13), in which the law constantly struggles to keep up with technological advancements (Drechsler & Kostakis, 2014).

To "open up the black box of AI" to society and effectively address tensions between conflicting values and principles in AI practices, an adequate approach involves iteratively translating abstract ethical principles and legal requirements through genuine democratic deliberation and contestation among different segments of society with varied expertise backgrounds. This essentially involves empowering individuals through human rights-based mechanisms that combine ethics, design-based approaches, and the law's force (**III**; High-Level Expert Group on Artificial Intelligence, 2019a; Nyman-Metcalf & Kerikmäe, 2020; Zicari *et al.*, 2021b). However, not all ethical demands in the context of AI regulation rise to the threshold of a human rights-related concern. Therefore, only the most significant normative concerns that the legal system has not addressed yet may require implementing new legal measures (Sartor, 2020).

In an attempt to respond to the need for holistic, critical, interdisciplinary approaches to regulating AI, as advocated by environmental and computer scientists and socio-legal and human rights scholars (Ananny & Crawford, 2018; Barocas *et al.*, 2023; Bria, 2017; Cohen, 2012, 2017, 2019; Gellert, 2021; Hoffmann-Riem, 2021; Hydén, 2021; Kitchin, 2017; Larsson, 2019; Lindgren & Dignum, 2023; Mantelero, 2022; Mittelstadt, 2019; Niklas, 2020; Prabhakaran *et al.*, 2022; Rahwan, 2018; Selbst *et al.*, 2019; Smuha *et al.*, 2021; Taddeo *et al.*, 2023; Taylor, 2017, 2023a; Umbrello, 2022; Yeung *et al.*, 2020; Zicari *et al.*, 2021a), the author suggests adopting a *bottom-up* instead of a top-down approach to AI regulation to realise Trustworthy AI in the long term.

As further elaborated in Chapter 4, the proposed *human rights-based* approach, including the legal conditions, goes beyond mere compliance with legal measures in existing civil liability frameworks and data protection regimes, yet combines human rights and socio-technical perspectives on AI regulation. Including both *ex-ante* and *ex-post* mechanisms that involve developers, deployers *and* citizens in their implementation, it considers the specific context in which AI is developed and how that development can impact the fundamental rights of individuals and groups at the use stage of the technology in the public sector (Cohen, 2017; High-Level Expert Group on Artificial Intelligence, 2019a; Prabhakaran *et al.*, 2022; Rahwan, 2018; Smuha, 2021a; Umbrello, 2022; Umbrello & van de Poel, 2021; Zicari *et al.*, 2021b; **III**). As such, grounded in systems thinking and participatory design, the *human rights-based* legal conditions are aimed at legislators, policymakers, *and* citizens alike since realising Trustworthy AI in the public sector requires the involvement of all relevant stakeholders in the development and assessment of the technology, beyond developers and deployers.

4 The contributions of the thesis

The thesis aims to contribute theoretically and analytically to the regulation of AI, particularly to human rights-based approaches.

First, the theoretical contribution identifies some significant limitations in existing EU legal frameworks. These limitations arise from a variety of factors. One is how the AI Act conceptualises AI systems (Ruschemeier, 2023). The thesis shows that a conceptualisation of AI treating AI as a neutral tool or neutral product has significant drawbacks. This leaves decisions about standards and measures to mitigate potential human rights harms mainly at the developers' discretion, excluding the possibility of involving public stakeholders, particularly citizens, in their role as recipients of public services, in the assessment and development of the technology. It is essential to consider that the design of AI systems can either constrain or promote fundamental rights (Aizenberg & van den Hoven, 2020; Cohen, 2017), significantly impacting citizens' choices and autonomy. A socio-technical perspective on AI regulation takes this into account, presenting an alternative approach to the existing understanding of AI in legal theory and practice (**I-III**). It can inform legal frameworks for AI by integrating legal and institutional requirements that advance public participation in the design of the technology and the assessment of its use in the public sector. Additionally, legal frameworks for AI based on fundamental rights provide affected persons with legal personality (AI subject) and substantive and procedural rights, an aspect the AI Act addresses only in a limited way (**III**). In turn, this involves creating legal and institutional requirements that enable citizens to contribute to developing and assessing AI systems in the public sector. In doing so, AI developers, deployers, and citizens would be enabled to mitigate potential human rights harms together. To this end, one of the legal requirements for AI is to provide citizens with a right to participate in developing and assessing the use of AI technologies in the public sector.

A second limitation of the AI Act is the risk-based approach. This regulatory method is generally considered a technocratic and expert-driven approach that restricts public stakeholders' participation (Kusch, 2007; Laux *et al.*, 2024; Smuha *et al.*, 2021; **III**). Additionally, imposing stringent legal obligations to high-risk AI systems only conflicts with the essence of fundamental rights as it risks prioritising economic benefits and national security interests over fundamental rights protection (Smuha *et al.*, 2021; **III**). While several AI systems in the public sector have minimal potential to create human rights harms (such as spam filters or language tools), drawing a line between *prohibited AI practices*, *high-risk AI systems* and AI systems other than high-risk, namely *certain AI systems* that arguably pose "only limited risk" or "lowered risk" (Regulation 2024/1689, recital 53), is challenging due to many borderline cases. This is because even chatbots, specifically learning-based chatbots, could harm citizens interacting with them in the public sector if only public authorities or developers' interests are addressed in the system's design phase and citizens' values are excluded (Coghlan *et al.*, 2023; Makasi *et al.*, 2022; Makasi *et al.*, 2021). At this early stage in the development of AI technologies, even technical and legal experts still need to understand the societal impact of widespread AI adoption in the public sector, including the effects on the fundamental rights of individuals and groups and the environment. As a result, the risk-based approach should only be seen as an initial step in protecting and promoting the fundamental rights of individuals and groups concerning AI (**III**).

Consequently, the thesis argues that the AI Act represents an innovation-inspired framework that tilts towards creating an ecosystem of excellence rather than an ecosystem of trust*,* with economic values remaining in tension with the fundamental rights-based approach of the AI HLEG (**III**). This applies particularly to aspects of public participation in the development of AI and the assessment of the use of AI in the public sector, the lack of legal requirements for developing citizens' capabilities on AI, and the lack of regulatory focus on increasing access to effective redress mechanisms for affected persons. Neither the Artificial Intelligence Liability Directive nor the Product Liability Directive accompanying the AI Act resolves these gaps (COM/2022/496 final; COM/2022/495 final). This is primarily because they focus on addressing material damages without adequately considering immaterial or social harms emanating from AI. Addressing them requires introducing legally enforceable rights and procedural safeguards for individuals and groups.

In turn, complementary to the AI Act, the thesis suggests adopting human rights-based legal and institutional requirements for the development and public-sector use of AI as part of a human rights-based approach. It includes technical, social, and legal perspectives on AI regulation both relating to the development of AI and the deployment of the technology in the public sector. Particularly, it aims to shift the focus of legal theory and practice, moving beyond simply regulating the safety features of technologies or mitigating potential human rights harms through technological management solutions alone (Lindgren & Dignum, 2023). Instead, the proposed legal conditions and requirements aim to address potential material and immaterial harm *towards human beings* and promote their fundamental rights.

Secondly, the thesis contributes theoretically to existing understandings of cybersecurity in AI regulation. As argued in **V**, existing approaches to cybersecurity are not enough to mitigate potential harms, including immaterial harm, to individuals and groups *interacting* with ICT-based systems. This is because vulnerabilities in the network and information systems of public and private entities are primarily addressed through technical solutions and existing approaches are siloed. EU cybersecurity legislation does only indirectly treat cybersecurity as a regulatory area of fundamental rights protection (Directive 2022/2555; Regulation 2019/881), and the AI Act (Regulation 2024/1689) is based on a similar regulatory approach, focusing on technical solutions rather than prioritising human aspects (Junklewitz *et al.*, 2023). It is also to be noted that the cybersecurity requirements stipulated in Article 15 apply primarily to high-risk AI systems (Regulation 2024/1689), and not to AI systems other than high-risk.

Yet, as shown in **II** and **V**, the nature and gravity of harms caused by ransomware attacks such as *WannaCry* to public and private entities, including individuals in essential services such as hospitals, and the normative concerns regarding the use of ML-based SM in residential households, such as tampering with the training and input data of AI, make it imperative to address vulnerabilities in AI through a fundamental rights prism, one that is interdisciplinary and participatory by default. This becomes even more important considering that cyberattacks can result in *decreased trust* in ICTs (**V**).

A conceptualisation of cybersecurity that focuses primarily on technical solutions has negative implications for the data protection and privacy and the cybersecurity of individuals and groups at the use stage of AI systems and can negatively impact the development and deployment of Trustworthy AI in the public sector in the long term. The goals, limitations, and the scope of legal frameworks for AI, therefore, need to be informed by an understanding of cybersecurity that goes beyond technical solutions.

Grounded in an approach that allows broader stakeholder participation in the assessment and setting of technical standards for cybersecurity in AI systems, the thesis argues for treating human beings as *assets* to be protected (Von Solms & Van Niekerk, 2013, p. 101). By understanding cybersecurity as the "protection of cyberspace itself, the electronic information, the ICTs that support cyberspace, *and the users* of cyberspace in their personal, societal and national capacity, including any of their interests, either tangible or intangible, that are vulnerable to attacks originating in cyberspace" (*Ibid.*), the thesis suggests that legal frameworks for AI should secure the *aggregate interactions* among human users, society and these information systems and networks (**V**). As such, the regulatory target should be the *aggregated interactions of individuals and society* with network and information systems. This goes beyond the existing understanding of cybersecurity as *information security*, which exclusively focuses on protecting the availability, authenticity, integrity and confidentiality of network and information systems and data. Mitigating vulnerabilities specific to AI, such as tampering with the training and input data of AI (**II**), therefore, requires mechanisms that can enable *citizens* to shape cybersecurity standards of AI systems beyond developers and deployers. Considering that cybersecurity is viewed as a societal challenge, a truly whole-of-society approach is needed to mitigate potential human rights harm ensuing from inadequate cybersecurity measures at individual and organisational levels concerning AI systems. Building up *societal resilience* (van Kranenburg *et al.*, 2023) against malicious actors, realising *robust AI* and, essentially, Trustworthy AI in the long term requires incorporating participatory mechanisms into emerging EU cybersecurity legislation, including the AI Act. To this end, the thesis suggests extending requirements concerning existing mechanisms of fundamental rights impact assessments and introducing participatory design approaches for *robust AI*.

Thirdly, the thesis contributes analytically and theoretically to the roles of developers and legal scholars in protecting and promoting fundamental rights concerning AI systems, particularly related to data protection and privacy standards and emerging cybersecurity obligations (**II-III**). Since technical and legal discourses address threats to privacy rights alone rather than combined (Prabhakaran *et al.*, 2022), the thesis attempts to show how engineers' systems thinking approaches can be combined with legal methods. By attending to end-users' concerns before the deployment of an AI system in the public sector, the tool proposed in **II** seeks to guide developers on their legal obligations related to developing AI while involving end-users and informing them about their rights regarding potential AI privacy infringements. Furthermore, by adopting interdisciplinary systems thinking approaches that include the VSD approach and the co-assessment inspection process for Ethical AI (Umbrello, 2022; Zicari *et al.*, 2021b), the thesis attempts to show how these participatory approaches can increase the protection and promotion of the fundamental rights of individuals and groups concerning AI.

The findings reveal the benefits of establishing legal mechanisms that enable lawmakers, developers, deployers and citizens, including domain-specific experts (such as doctors or governmental lawyers) to develop AI systems through the participatory design approach of VSD and assess the technology through the fundamental rights impact assessments and the co-assessment inspection process for Ethical AI (**III**). These interdisciplinary, multistakeholder approaches show an initial potential to reconcile the values of both direct and indirect stakeholders at the development and use stages of the technology in the public sector, mitigate potential human rights harms and promote the fundamental rights of individuals and groups regarding AI (**III**).

4.1 Individual findings in the publications

The main argument of the thesis for the need to regulate the development and public-sector use of AI with a **human rights-based approach that combines socio-technical and human rights perspectives** rests on three publications, two of which are co-authored and one single-authored:

Publication I, co-authored with Professor Kerikmäe as part of a book chapter, conducted a literature review of existing research on legal frameworks for ICTs, with a focus on regulating AI. The chapter provides a broad overview of AI, its main approaches, and its application areas. Additionally, it identifies the main normative issues raised by AI and analyses how they could be regulated. Lessig's four modalities of regulation and the VSD method are considered suitable for this purpose, adopting a socio-technical perspective on regulating ICT-based systems (Lessig, 1999, 2006; van den Hoven, 2017). Furthermore, during the review of EU primary and secondary sources, the chapter identified the AI HLEG framework for Trustworthy AI as crucial for developing legal frameworks for AI. The chapter suggests that the regulatory discussions on AI should be centred on the question of by whom and for which purpose AI systems will be designed, by whom they are owned and deployed, and in which contexts they will be applied. To address normative concerns concerning the deployment of AI systems, the authors argue for establishing legal requirements that can reconcile the interests of developers, deployers and citizens in AI regulation.

Publication II was co-authored with Tobias Häring, Dr Korõtko, Professor Rosin, Professor Kerikmäe and Professor Biechl as part of an interdisciplinary case study on the use of the ML technique of NILM in SM in residential households. Since end-users' perspectives on AI regulation remain underexplored, we mapped AI end-users' concerns regarding the deployment of an AI system in residential households. Some of the end-users' concerns include discrimination, denial of access to services or identification of home appliances. The combined technical and legal analysis examines how applicable EU legal frameworks, particularly the GDPR and the NIS Directive, deal with end-users' concerns. GDPR Article 22 presents an important provision for citizens when public entities process personal data for administrative decisions enabled by SM (Regulation 2016/679). For instance, when citizens receive an incorrect electricity bill, Article 22 should grant data subjects the right to rectify potential discrimination and impose an obligation on utilities and developers of AI systems to address this issue pre-emptively. However, the paper also shows limitations of the protection provided by GDPR Article 22. This is mainly due to the opacity of AI, which complicates the assignment of responsibility for privacy infringements or data breaches. The publication proposes a visual tool for utilities, AI developers, and citizens to mitigate potential data breaches concerning the development and public-sector use of AI together.

Publication III is a single-authored article and examines to what extent the Commission's Proposal for an Artificial Intelligence Act translated the ethical principles of the AI HLEG into legal requirements. The paper's findings reveal several shortcomings in the Commission's Proposal in achieving this goal. One of the shortcomings is that the AI Act treats AI systems merely as a neutral product or tool, which is the common understanding in legal theory and practice. However, the STS scholarship and VSD approaches have demonstrated that technologies can embody the values of their designers, influencing and restricting the choices available to citizens (Cohen, 2017; Friedman & Hendry, 2019; Umbrello, 2022; van den Hoven, 2017). This is important since innovation in AI mainly occurs in research clusters where AI companies often prioritise economic gains over societal benefits. Without the right for citizens to participate in the development and assessment of AI, developers might prioritise more privacy-intrusive approaches for financial gains instead of technical approaches that provide transparency and explicability for citizens. A legislative framework for AI that treats AI merely as a neutral product or tool risks giving AI companies and developers too much leeway to decide which values to embed into their products and how to audit them as part of self-conformity assessments. By adopting a socio-technical perspective on the regulation of AI, the article shows the benefits of integrating participatory design mechanisms into legal frameworks for AI as part of a fundamental rights-based approach, conceived as precautionary, participatory, and bottom-up.

Publication IV in the appendix is a report on legal frameworks for AI systems that focuses on legal liability questions, led by Professor Kerikmäe and conducted with Professor Nyman-Metcalf, Professor Hoffmann, Dr Minn, Dr Liiv, Professor Taveter, Dr Solarte Vasquez and Olga Shumilo. The author contributed with desk research on the AI policies of Germany, France, the UK, Sweden, Denmark, Finland, the USA, and the EU. The reviews of the EU policy documents informed **Publication I**. The report evaluated the extent to which Estonian legislation can address the legal challenges AI systems present to private and public sectors. The publication recommended following existing EU laws and upcoming legal developments on AI at the EU level rather than creating new legislation for AI at a national level and proposed several measures to the Estonian government: establishing a registry of robots operating in the physical world, extending the competencies of the Technical Regulatory Authority, establishing a national insurance fund for autonomous intelligent technologies, creating impact assessments and standards for AI based on the degree of risk of AI, and reviewing six categories of Estonian legal acts with a focus on the machine readability of legal acts, legal decision-making, and granularity of access control and responsibility.

Publication V in the appendix, co-authored with Dr Kasper, is a discussion paper and a normative contribution to the development of EU legal frameworks for cybersecurity. The paper argues that understanding the harms of cyberattacks is essential to prioritising the legal framework's goals, limits, and scope for cybersecurity. To this end, the author conducted desk research on the ransomware attack *WannaCry*, identified the harms it caused, and evaluated its implications for the development of cybersecurity legislation at the Union level. The article's conceptualisation of EU cybersecurity laws informed **Publication II**. The findings show that protective measures should address the following harms: potential and actual economic damages, decreased productivity, reputational damage, physical and intangible harms to citizens, reduced trust in computer systems, destabilisation of the physical world, and potential losses in sovereignty. It calls on policymakers and legislators to rethink existing approaches to cybersecurity, which should move beyond technological solutions and consider social and human aspects in developing legal frameworks for cybersecurity and AI (Von Solms & Van Niekerk, 2013). The analysis revealed that the EU's approach to cybersecurity lacks several legal mechanisms to promote the development of common standards, enhance cooperation in preventing cyberattacks on both public and private network and information systems, and mitigate potential human rights harms to citizens as a result of inadequate cybersecurity measures at the organisational and individual level.

4.2 Lawful Trustworthy AI

Based on the individual findings in the publications, including in the appendix papers, this section of the thesis determines the legal conditions necessary and sufficient to adequately protect and promote individual and collective fundamental rights concerning the development and public-sector use of AI. Several proposed conditions have been highlighted in the respective publications and further discussed in the previous sections. Legal scholars have extensively analysed some of these conditions. Therefore, the proposed conditions are interconnected and may overlap with them. All proposed conditions are theoretical and conceptual, as none have been empirically tested or confirmed. The conditions are exemplified based on the findings of the individual publications and only apply to AI practices not prohibited by the AI Act (Regulation 2024/1689). In essence, aiming to restore human agency vis-à-vis technological determinism based on a deontological, Kantian interpretation of morality in which humans are treated as ends in themselves rather than mere means to an end, the thesis proposes a human rights-based approach to AI.

Regulating the physical or safety risks of AI is not enough to realise Trustworthy AI. AI regulation needs to be grounded in an approach that puts the main regulatory focus on the social or human rights risks (**III**). While the AI Act aims to enhance transparency by mandating developers of high-risk AI applications and general-purpose AI models to implement technical documentation measures (Regulation 2024/1689), without providing effective redress to affected persons and involving citizens in fundamental rights impact assessments before adopting AI systems in the public sector, these requirements do not yet adequately protect and promote fundamental rights in the long term (**III**). This is especially important given that the AI Act relies on self-conformity assessments by providers of stand-alone high-risk AI systems and product safety monitoring by market surveillance authorities with yet limited expertise in human rights monitoring of AI use cases (Laux *et al.*, 2024; Smuha *et al.*, 2021; Taylor, 2023a; **III**).

Additionally, the obligation to conduct self-conformity assessments for high-risk AI systems under Article 43 is triggered only if the provider has determined independently that the AI system does not represent "a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision-making" (Regulation 2024/1689, Articles 6(3), 43 and 49(2)).[12]

[12] The exception to the conduct of conformity assessments for high-risk AI systems is enshrined in an additional clause in paragraph 3 of *Article 6: Classification Rules for High-Risk AI Systems* (AI Act, 2024):
"1. Irrespective of whether an AI system is placed on the market or put into service independently of the products referred to in points (a) and (b), that AI system shall be considered to be high-risk where both of the following conditions are fulfilled:
(a) the AI system is intended to be used as a safety component of a product, or the AI system is itself a product, covered by the Union harmonisation legislation listed in Annex I;
(b) the product whose safety component pursuant to point (a) is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment, with a view to the placing on the market or the putting into service of that product pursuant to the Union harmonisation legislation listed in Annex I.
2. In addition to the high-risk AI systems referred to in paragraph 1, AI systems referred to in Annex III shall be considered to be high-risk.
3. By derogation from paragraph 2, an AI system referred to in Annex III shall not be considered to be high-risk where it does not pose a significant risk of harm to the health, safety or fundamental

In turn, the discretion upon complying with obligations for high-risk AI systems under Articles 8 to 27 of the AI Act (Regulation 2024/1689), particularly with the quality management system and technical documentation requirements, including obligations on *data and data governance, transparency and provision of information to deployers, human oversight, accuracy, robustness and cybersecurity*, remains *primarily* with the developers and owners of AI systems. This is important to consider since conformity assessment presents the main mechanism for the Commission, market surveillance authorities, and, in the case of biometrics, independent conformity assessment bodies, to assess the compliance of providers with the obligations for AI systems currently deemed high-risk for the fundamental rights of individuals and groups. The AI Act is, therefore, an innovation-inspired legal instrument, where economic values conflict with human values, putting the overarching goal of Trustworthy AI potentially at risk in the long term (**III**, pp. 58-59).

Consequently, rendering the development and public-sector use of AI compatible with fundamental rights implies going beyond expert-driven AI governance approaches (Laux *et al.*, 2024; Taylor, 2023a). There is a crucial need to shift the focus of AI regulation towards humans and the potential human rights harms that advanced technologies may pose, particularly if not addressed collaboratively at an early stage in the implementation of AI technologies (**III**).

4.2.1 Fundamental rights impact assessments

First, to enable adequate protection of fundamental rights and empower individuals and groups concerning the development and public-sector use of AI, public entities are required to ensure a robust and resilient ICT infrastructure against external and internal cyber threats, taking AI-specific vulnerabilities into account (**II**; **V**). As shown in the case study of the ransomware attack *WannaCry*, cyberattacks can impact citizens' and society's trust in the ICT infrastructure and public institutions in which the attack materialises by disrupting operations of essential services such as hospitals (Klimburg, 2012; Von Solms & Van Niekerk, 2013; **V**). Ensuring a robust and secure ICT infrastructure is specifically crucial concerning AI as the outputs of AI systems can be both unintentionally and intentionally incorrect or biased with harmful effects on individuals and groups, potentially infringing on the fundamental rights to *respect for private and family life*, the *protection of personal data*, or *non-discrimination* (**II**).

rights of natural persons, including by not materially influencing the outcome of decision making. The first subparagraph shall apply where any of the following conditions is fulfilled:

(a) the AI system is intended to perform a narrow procedural task;

(b) the AI system is intended to improve the result of a previously completed human activity;

(c) the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review; or

(d) the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III. Notwithstanding the first subparagraph, an AI system referred to in Annex III shall always be considered to be high-risk where the AI system performs profiling of natural persons.

4. A provider who considers that an AI system referred to in Annex III is not high-risk shall document its assessment before that system is placed on the market or put into service. Such provider shall be subject to the registration obligation set out in Article 49(2). Upon request of national competent authorities, the provider shall provide the documentation of the assessment […]".

Specifically, when AI systems are used in critical sectors such as energy (**II**) or healthcare, in addition to financial and reputational damage, material and immaterial harm to citizens can occur from public entities' lack of preventive and responsive cybersecurity measures (**V**). As revealed in **II** concerning the use of SM in residential households, AI-specific vulnerabilities can include tampering with the training and input data of AI, which in turn can result in the denial of access to services, discrimination, the exhibition of user habits, and illnesses, disconnection of home appliances, denial of personal mobility, or burglary, and thus potential harms to citizens. To address these AI-specific vulnerabilities and harms, cybersecurity legislation must respond to the needs of citizens and recipients of public services, considering citizens as *assets* to be protected (Von Solms & Van Niekerk, 2013, p. 101; **V**).

Therefore, in addition to addressing AI-specific vulnerabilities and increasing AI system's resilience and robustness through conventional mechanisms such as the cybersecurity certification scheme, the creation of incident response teams, the assignment of clear responsibilities and roles for the mitigation of cybersecurity threats or reporting obligations for improving the communication, coordination, and information-sharing between the private and public sectors regarding cyber incidents and threats, the role of citizens in cybersecurity mitigation and response efforts should be strengthened (Regulation 2019/881). While the initial focus on implementing technical solutions for protecting network and information systems is important, citizens' interests remain underrepresented in cybersecurity legislation (Papakonstantinou, 2022; Von Solms & Van Niekerk, 2013; **V**). Arguably, citizens' involvement in critical preventive cybersecurity measures can improve societal resilience and robustness as part of a whole-of-society approach and allow for adequate protection and promotion of fundamental rights concerning AI. Therefore, EU cybersecurity legislation should secure not only the interconnected information systems and networks, including data, but also the *aggregate interactions* between human users, society and these information systems and networks within the EU, taking thus human and social aspects in the standardisation of cybersecurity, and collective preventive and responsive measures for critical infrastructure protection and essential services into account (**V**, p. 21).

While the NIS 2 Directive has increased focus on the role of citizens concerning cybersecurity measures in public sector entities, as it now includes cybersecurity obligations specifically addressed to public administrations, adjusted the criteria of determining cybersecurity incidents from quantitative to qualitative measures, such as focusing not only on the numbers of citizens affected, the geographical intensity and length of attacks but also material and immaterial harm of cyber threats and attacks (Directive 2022/2555; Vandezande, 2024), legal incentives remain lacking for public authorities to enable the participation of citizens in cybersecurity mitigation efforts for AI (**II**). Addressing this gap involves mechanisms that allow and enable citizens to participate in assessing and developing AI systems in public sectors before their use.

To enable meaningful participation of citizens in the development and deployment of AI systems, one of the mechanisms should include the involvement of citizens, and depending on the context in which the technology is applied, domain-specific experts beyond developers and deployers, in fundamental rights impact assessments before deploying AI systems in public sectors (Mantelero, 2022; Zicari *et al.*, 2021a; Zicari *et al.*, 2021b; **II-III**). To ensure the AI system's technical robustness and safety, deployers should be legally required to allow citizens, including vulnerable groups and domain-specific experts, to participate in the conduct of fundamental rights impact assessments for AI and thus “co-determine” or co-construct how values, principles and rules are “hardwired” into AI systems (Cohen, 2017, p. 87; **III**). Gender parity should also be included to ensure that gender interests are taken into account in the assessment process. This is particularly important considering that the use of AI technologies tends to represent male interests more than female interests and may reinforce existing biases (Buolamwini & Gebru, 2018; O’Connor & Liu, 2023).

Due to the evolving nature of AI, legal frameworks for AI should also provide citizens with the right to participate in periodic reviews of the criteria of fundamental rights impact assessments. This is particularly important regarding the joint evaluation of what constitutes material and immaterial harm and thus creates a detrimental causal impact on individuals and groups concerning AI (Barocas *et al.*, 2017; McGregor *et al.*, 2019; Moss *et al.*, 2021). In the long term, fundamental rights impact assessments for AI need to learn from other impact assessment mechanisms and combine their assessment mechanisms for quantitative and qualitative measurements of harm. Lessons can be specifically drawn from environmental impact assessments, where the participation of directly or indirectly impacted stakeholders has shown to be the highest among existing types of impact assessments (Moss *et al.*, 2021). This is important for addressing limitations in existing fundamental rights impact assessments, specifically when it comes to the evaluation of risks of energy-intensive AI to the environment. While these risks in the deployment of AI systems may currently be seen as minimal, from an environmental perspective, there might be AI systems which pose a higher risk to the exercise of fundamental rights of human beings in the long term than existing empirical knowledge might suggest. In this regard, fundamental rights impact assessments should not serve as a panacea in mitigating human rights harms concerning AI, including the dimension of cybersecurity with regard to *robust AI*. However, they present a crucial mechanism if they are meaningfully opened to citizens and their assessment criteria are open to meaningful review and revision.

In this regard, making meaningful participation of citizens in fundamental rights impact assessments legally mandatory before the deployment of AI systems in the public sector, particularly but not only for high-risk AI systems, responds to the AI HLEG ethical principles of the respect for human autonomy, prevention of harm, fairness, and explicability.

4.2.2 Effective redress mechanisms

A second necessary condition is ensuring effective redress mechanisms for affected persons, in line with Article 47 of the EU Charter (**III**). Previous mundane AI use cases, for example, for publicly legitimate purposes such as the detection of welfare fraud in the *SyRI* case have shown how, despite data governance and oversight measures, the deployment of AI can cause significant human rights harms to citizens (Alston, 2019; Rachovitsa & Johann, 2022; A/74/493, 2019; Wieringa, 2023; Yeung, 2022). As the *SyRI* case has also shown the potential to negatively impact citizens’ trust towards deployers of AI technologies, addressing unintended harms in public-sector use of AI systems through effective redress mechanisms is an important mechanism to enhance citizens’ trust in public institutions deploying AI in the long term in line with the values enshrined in Article 2 of the EU Treaty (**III**).

In this regard, GDPR Article 22 and the “data subject’s right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her” presents one part of measures to achieve effective redress (Regulation 2016/679; **II**). As earlier addressed, this information right needs to be complemented by additional measures due to the limitations of extant data protection legislation for protecting only individual and not collective rights with regard to public authorities' processing of personal data in AI-informed decision-making. Furthermore, citizens often waive their rights when entering into contractual relationships or obligations by consenting to conditions that allow public entities to withhold information about the functioning of the AI system. Additionally, not all public-sector AI-informed decisions will likely be "solely" based on automated processing. AI's issue of opacity and the fact that public authorities often withhold information about the risk indicators and, thus, how the AI system might have reached a decision presents additional hurdles for citizens to claim their rights concerning the data output of the AI system (Wachter & Mittelstadt, 2019). Without reasonable individual or collective access to AI's source code and information about how the system has reached a legally binding decision for citizens, under these accumulated conditions, its use creates an additional power asymmetry between the State and the citizen, specifically when the citizen is expected to prove causation between the model's logic and its harms caused, be it discrimination or the exposition of illnesses (**II**).

Therefore, as part of effective redress mechanisms for AI, deployers should be obliged both to make citizens aware when an AI system is used to aid administrative decisions with a legally binding effect for citizens and provide an explanation in clear and easily understandable language about the purpose of its deployment, including concerning which legal options and mechanisms under both EU and national law are provided for affected citizens to receive redress for potential immaterial or material human rights harm, independent of the current risk-categorisation of the AI system.

Additionally, legal frameworks for AI need to include provisions that require public authorities to provide reasonable information as part of explanations to citizens about how administrative AI-informed decisions based on risk indicators have been reached, to the extent that the overall functioning of AI and the purposes it was designed for are preserved. In this regard, a precautionary approach to AI assumes that the burden of proof is not on the side of affected persons but on the deployer as the more powerful actor in the relationship. While tensions might arise in using AI for legitimate public purposes (High-Level Expert Group on Artificial Intelligence, 2019a), particularly for the prevention of criminal offences, affected persons should have the right to reasonable explanations on a case-by-case basis, and national supervisory authorities should review their individual or collective requests promptly.

Concretely, effective redress implies respecting and enabling human autonomy and agency when adversely impacted by an AI-informed administrative decision in line with the AI HLEG interdependent ethical principles for the respect of human autonomy, prevention of harm, fairness, and explicability (**III**). While Article 86 in the AI Act introduces a right to explanation of individual decision-making, a right to reasonable explanations about the AI system's functioning could only be an effective redress mechanism if affected persons receive information about and access to redress mechanisms (Regulation 2024/1689). In this regard, legal frameworks for AI should also enable the lodging of collective complaints similar to the "right to mandate a not-for-profit body, organisation or association" stipulated in Article 80 of the GDPR (Regulation 2016/679; Smuha *et al.*, 2021). Awareness and access should also include the obligation for public entities to communicate in an understandable language to citizens about their right to contest AI-informed decisions. "By design" mechanisms could enable this (Aizenberg &

van den Hoven, 2020; Fanni *et al.*, 2023). In this regard, legal frameworks for AI should allow citizens to participate in deliberating upon the design of digital redress mechanisms and their creation. This could, for instance, be combined with the conduct of fundamental rights impact assessments, as shown under the first condition. In addition to digital formats, citizens should also be informed about their rights in paper formats and formats that allow vulnerable groups in society, such as visually impaired individuals, to enjoy their right to effective redress.

Enabling effective redress will be a complex task for legislators, including public authorities. However, to realise Trustworthy AI in line with the AI HLEG ethical principles and Article 47 of the Charter implies considering taking complementary measures to the GDPR Article 22, extending information rights to citizens, including vulnerable groups in society. Additionally, it implies informing citizens about their rights and thus enabling them to meaningfully contest AI-informed decisions by public entities that may have caused them immaterial or material harm.

4.2.3 Socio-technical digital literacy

Third, to enable conditions one and two, the capabilities of citizens in their roles as public servants, judges, engineers, prosumers, developers, cashiers, platform workers, doctors, and so forth should be increased so they can critically and meaningfully participate and thus co-determine how AI systems are designed and deployed in the public sector for citizens (Pohle & Thiel, 2020; **III**). One measure to ensure digital literacy on AI, precisely one that is socio-technical, participatory and reflective, is to allocate adequate funding for interdisciplinary education programs and curricula from preschool to higher education (Hasselbalch, 2021). Therefore, adequate funding should be tailored to improving access to education, including higher education programmes and curricula that focus on the ethical, legal, technical, societal and environmental aspects of AI systems, including “the possible limits of automation”, from an interdisciplinary perspective (European Group on Ethics in Science and New Technologies, 2021; Kerikmäe & Pärn-Lee, 2021; Nyman-Metcalf & Kerikmäe, 2020, p. 48; **III**). Addressed beyond lawyers and engineers, these curricula should include courses that introduce students to the concept of Trustworthy AI, the conduct of fundamental rights impact assessments, cybersecurity, the VSD approach for developers of AI, and other ethical and legal aspects of technology law and human rights studies.

Furthermore, from an organisational perspective, it should be a legal requirement for public entities that educational curricula be “co-created” with the lecturers’ and students’ interests in mind, in particular, but not only, in engineering education (Voll, 2023). This is essential for improving the effectiveness and quality of higher education programs and enhancing the long-term well-being of the students and lecturers.

A lifelong learning approach to AI should be enabled to increase citizens’ digital capabilities and critical thinking and allow for informed and conscious decisions about how, with whom, and for which purposes citizens share their data (Berlin Declaration on Digital Society, 2020). Interdisciplinary higher education programmes on AI should be open to citizens of all ages as part of open university programs.

While public-private collaboration in education on AI is essential, it is also important to take a critical distance and review tech partnerships sometimes, specifically to ensure the data protection rights of children, students and lecturers (Celeste, 2021; Compagnucci, 2022). Sometimes, opting for (European) open-source and privacy-protective technical solutions in schools and universities might be wiser than experiencing lock-in effects by commercial software. Considering the risk of AI industry "tech capture" (Whittaker, 2021) in academic institutions, in the long term, establishing the right for citizens to review tech collaborations at public universities to maintain intellectual freedom may also be necessary.

Ensuring socio-technical digital literacy through adequate government and private funding of interdisciplinary educational programs is essential for achieving Trustworthy AI. While not a panacea, education and access to education are key enabling factors for citizens to be able to meaningfully partake in co-determining AI standards in the development and use of AI systems, for example, through participation in fundamental rights impact assessments for robust AI or the development of effective redress mechanisms by design.

4.2.4 Monitoring and enforcement capacity of national supervisory authorities

Fourth, it is essential to ensure compliance with existing and novel human rights obligations and improve the capacity of national supervisory authorities to monitor and enforce them concerning public entities' use of AI systems (Smuha *et al.*, 2021; Zuiderveen Borgesius, 2020; **III**). Three criteria must be fulfilled in this regard.

First, Member States should provide adequate funding for national supervisory authorities since experience around the GDPR has shown that compliance with and enforcement of data protection laws are highly contingent on sufficiently resourced national supervisory authorities and trained human rights experts (NOYB – European Center for Digital Rights, 2024; Zuiderveen Borgesius, 2020; **III**). Secondly, compared to the Commission's Proposal in 2021, the final text of the AI Act (Regulation 2024/1689) stipulates in Articles 70 and 77 that Member States should secure funding for enforcing and monitoring obligations concerning high-risk AI systems. However, these requirements should also apply to AI systems other than high-risk (**III**). This is a vital legal requirement since these AI systems, including chatbots or deep fakes, might create similar human rights harms for citizens interacting with them, for example, when chatbots are deployed to inform administrative decisions with legal effects for individuals and groups. The risk-based approach in the AI Act creates the problem that the monitoring and enforcement of the obligations are curtailed to high-risk AI systems only (Laux *et al.*, 2024; Smuha *et al.*, 2021; **III**). However, the nature of fundamental rights makes it mandatory that its principles and protective guarantees are at least equally applicable to the interactions of individuals and groups with AI systems other than high-risk as to those currently classified as high-risk (**III**). Democratic societies are only beginning to understand the borderline cases of AI and, thus, the consequences of the widespread introduction of AI systems in public entities (**III**).

On the one hand, a risk-based approach might have the value of focusing often scarce resources on mitigating and monitoring the most imminent harms posed by AI and creating more demanding requirements and safeguards for the developer and deployer of AI systems (Wendehorst, 2020). On the other hand, the risk-based approach, as a list-based approach, has limitations since national competent authorities, including notifying authorities and market surveillance authorities, selected by the respective Member State to monitor the enforcement of the AI Act, could place the regulatory focus on high-risk AI systems only. Consequently, due to limited resources, they might overlook AI use cases other than high-risk that might be causing equal, if not more, severe human rights harms to citizens (**III**).

Since the risk-based approach has been adopted to realise Trustworthy AI, the approach must be continuously and iteratively contextualised and evaluated against the AI HLEG four ethical principles (**III**). To achieve this, fundamental rights impact assessments for the public-sector use of AI should be conducted with the participation of citizens, including regarding the revision of their parameters, as referred to under condition one. This may enable the contextualisation of human rights harms and align the risk-based approach with the "dignitarian" understanding of fundamental rights (Mantelero, 2022; Yeung & Bygrave, 2022, p. 143).

Furthermore, Article 112 is a vital mechanism in the AI Act regarding the potential future revision and extension of high-risk AI systems in Annex III. However, beyond the Commission, other public actors, particularly national human rights institutes and, on a case-by-case basis, citizens, should be allowed to make suggestions on adding additional AI systems to the list of high-risk AI systems (**III**). In this regard, more extensive collaboration among members of the EU AI Board, and the respective authorities protecting fundamental rights, including representatives of the advisory forum, is required. This can be achieved through meetings among the relevant stakeholders on a regular instead of a case-by-case basis, including in the sub-groups. Additionally, the agenda of all meetings of the EU AI Board must be published and made accessible to citizens.

Thirdly, legal frameworks for AI should encourage market surveillance authorities and national human rights institutes to increase collaboration and information exchange on AI human rights monitoring experiences. This collaboration and sharing of experiences between the agencies could be further institutionalised through AI registries in cities, built on the experience of public AI registers in Amsterdam and Helsinki (Floridi, 2020a). This is important since the primary expertise of market surveillance authorities does not lie in protecting fundamental rights (Smuha *et al.*, 2021; **III**).

In general, realising Trustworthy AI should not be limited to only a compliance exercise that can be achieved quickly. In line with Laux *et al.* (2024, p. 2), "trustworthiness is a longitudinal concept that necessitates an iterative process of controls, communication, and accountability to establish and maintain its existence across both AI technologies and the institutions using them." The question of whether the risk-based approach is incompatible with the concept of trustworthiness in general, as argued by Laux *et al.* (2024), requires the collection of empirical evidence. To realise Trustworthy AI and, thus, ensure adequate protection and promotion of fundamental rights concerning AI, relying solely on the existing risk-based approach is insufficient (**III**). As previously argued, the risk-based approach must be continuously reassessed and adapted. To this end, it is necessary to adopt complementary legal and institutional requirements that enable participation and contestation to maintain the individual and collective agency regarding the development and public-sector deployment of AI.

In summary, condition four implies that only adequately resourced and trained human rights experts in institutions can ensure compliance, monitor the public-sector use of AI, and address potential human rights harms. The principles of fundamental rights should apply equally to all AI systems and, thus, beyond high-risk AI systems, and collaboration between the national supervisory authorities, particularly market surveillance authorities and institutions monitoring and enforcing fundamental rights law improved, for instance through the mechanism of public AI registers.

4.2.5 Fora for participatory design and the inspection process for Ethical AI

Fifth, legal frameworks for AI should include requirements mandating public authorities to create fora in existing public institutions for translating the ethical principles of the AI HLEG and applicable legal requirements into AI systems with the participation of developers, deployers, citizens and domain-specific experts (lawyers, doctors, and so forth) (Aizenberg & van den Hoven, 2020; Cantero Gamito & Gentile, 2023; Umbrello, 2022; Zicari *et al.*, 2021a; Zicari *et al.*, 2021b; **II-III**). This context-sensitive, nuanced translation process should be based on co-design methodologies, particularly the fundamental rights-based, interdisciplinary systems thinking method of VSD and the inspection process for Ethical AI (Friedman & Hendry, 2019; Umbrello, 2022; Zicari *et al.*, 2021b; **III**). Essentially, these fora are aimed at embedding ethical, technical, and legal requirements into AI systems in the public sector while addressing potential value tensions at the development stage of AI technologies and facilitating stakeholders' understanding of the potential human rights harms the use of AI can create for individuals and groups in different application contexts.

In addition to the participation of citizens in fundamental rights impact assessments, these fora need to be designed such that they enable all stakeholders to actively shape the technological normativity of AI and its resulting values following the AI HLEG principles of the respect for human autonomy, prevention of harm, fairness and explicability.

As argued in **II** and **III,** the design of AI systems in its engineering or development process can either increase or constrain human agency in the use stage of the system. In other words, due to the inherent technological normativity of AI, the developer's design choices can impact citizens' intellectual freedom and choices (Aizenberg & van den Hoven, 2020; Cohen, 2017; Hydén, 2020; van den Hoven, 2017; **III**). Additionally, it can impact the citizens' long-term perception of the AI system and, thus, AI's trustworthiness, including in governmental AI projects (Kleizen *et al.*, 2023). Therefore, the thesis explicitly defines AI from a socio-technical perspective as "an autonomous digital technology embedded into societal structures and contexts, mediated through digital devices" (**III**, p. 59). Emphasising the socio-technicity of AI serves to establish the link between the role of developers' design choices and their implications for human agency (Umbrello, 2022; Umbrello & van de Poel, 2021; **III**). It thus serves to increase the role of design approaches for protecting and promoting the fundamental rights of individuals and groups in the context of innovation in AI products and services (**III**). This is important to consider since the AI Act primarily gives technology owners and developers the authority to determine AI standards in self-conformity assessments with limited incentives for developers to involve citizens or deployers in this auditing process (Laux *et al.*, 2024; Regulation 2024/1689).

Additionally, developers are often unaware of the potential human rights harms associated with the application context of AI technology (Friedman & Hendry, 2019; Umbrello, 2022; van den Hoven, 2017; **II-III**). Taking the growing information asymmetry among AI developers, deployers and citizens into account, the development of AI systems before their implementation in different domains in the public sector must be supported by the expertise of domain-specific experts such as doctors, lawyers, lecturers, and environmental experts (Vetter *et al.*, 2022; Zicari *et al.*, 2021a; Zicari *et al.*, 2021b; **III**), and also ordinary citizens ought to be allowed to participate in the development and assessment of AI systems in the public sector.

In the deliberation, development, and testing process of AI systems for use in Courts, it is essential to bear in mind that "legal protection by design" and not the "legal by design" principles should guide this process (Hildebrandt, 2020, p. 251 *ff*; **IV**). Considering that AI systems are imperfect, only AI technologies in public sectors that are by design contestable in the use stage should be used, with substantive and procedural safeguards for affected persons as part of effective redress mechanisms. This principle should also be applied regarding the second condition and the requirement of enabling effective redress mechanisms by design.

Furthermore, as the *SyRI* case revealed, the sources of human rights harms can lie beyond the design of the AI model and can emanate from deployers who purchase the AI system and select the purpose and location of its deployment. Opacity and bias in AI systems, including the deployer's hidden cognitive biases, can thus be potentially reinforced and mediated through AI, impacting the fundamental rights of citizens, such as their rights to *privacy* and *non-discrimination*, the right to an *effective remedy* and the right to *access to social security*. Yet, implementing the VSD approach and the co-assessment inspection process for Ethical AI before deploying AI systems in public sectors could also be a mechanism to mitigate human rights harms beyond the design of the AI system. The VSD approach and the co-assessment inspection process for Ethical AI present methods that, on a case-by-case basis, can also account for *power* concerning the development and deployment of AI systems, providing the first structured methods to resolve tensions between different stakeholders' values in this regard (Friedman *et al.*, 2021; Jacobs *et al.*, 2021; Zicari *et al.*, 2021b).

More recent implementations of AI systems in public entities have also shown the potential to increase *organisational errors*[13] and change institutional rules and procedures with harmful effects on citizens (see, particularly, Robodebt Scheme: Royal Commission into the Robodebt Scheme, 2023; Yeung, 2022). Beyond focusing primarily on the design of AI systems, the VSD approach, along with the co-assessment inspection process for Ethical AI, could also be applied to reduce errors rooted in *institutional design* that is yet insensitive to potential human rights harms emanating from the development and use of AI systems. Further empirical research is needed to understand how participatory design approaches can be applied to further account for power and address potential tensions among stakeholders in different application areas of AI in the public sector. In this regard, the thesis views the VSD method and the co-assessment inspection process for Ethical AI as the first structured approaches to organising *participatory AI governance* in the public sector.

Realising the fifth condition implies creating fora to implement the participatory design method of VSD and the co-assessment inspection process for Ethical AI in existing public institutions. These fora allow for a proactive contextualisation of potential human rights harms concerning the development and deployment of AI systems in the public sector. Given that AI developers are increasingly configuring public spaces and citizens' choices through their architecture, including the risks of AI for increasing organisational errors in public institutions, these fora can be viewed as one mechanism for "taking collective responsibility to attend to the socio-technical foundations of moral and democratic freedom" (Cohen, 2017; Yeung, 2019, p. 70). They are aimed at increasing

[13] *Organisational errors* refers here to "actions of multiple organisational participants that deviate from organizationally specified rules and can potentially result in adverse organisational outcomes" (Goodman *et al.*, 2011, p. 154).

human oversight in the AI system lifecycle and fostering public accountability in line with the ethical principles of the AI HLEG. Therefore, legal frameworks for AI should include requirements that mandate governments to establish these fora and enable the participation of developers, deployers, and citizens, including a diverse group of domain-specific experts, in the development and assessment of AI systems. Lastly, it is important to communicate the necessity of establishing these fora among both public and private stakeholders.

4.2.6 Visual summary

Figure 2 provides a visual summary of the findings of the thesis, illustrating the connections among citizens, developers, deployers, the AI HLEG ethics principles within the Trustworthy AI framework, the proposed legal conditions as part of *Lawful Trustworthy AI*, and existing applicable legislation for AI under *Lawful AI*. Furthermore, Table 3 on pages 60 to 64 summarises the proposed legal conditions (A through E) and recommendations to address the identified gaps in the AI Act for achieving Trustworthy AI in the long term.

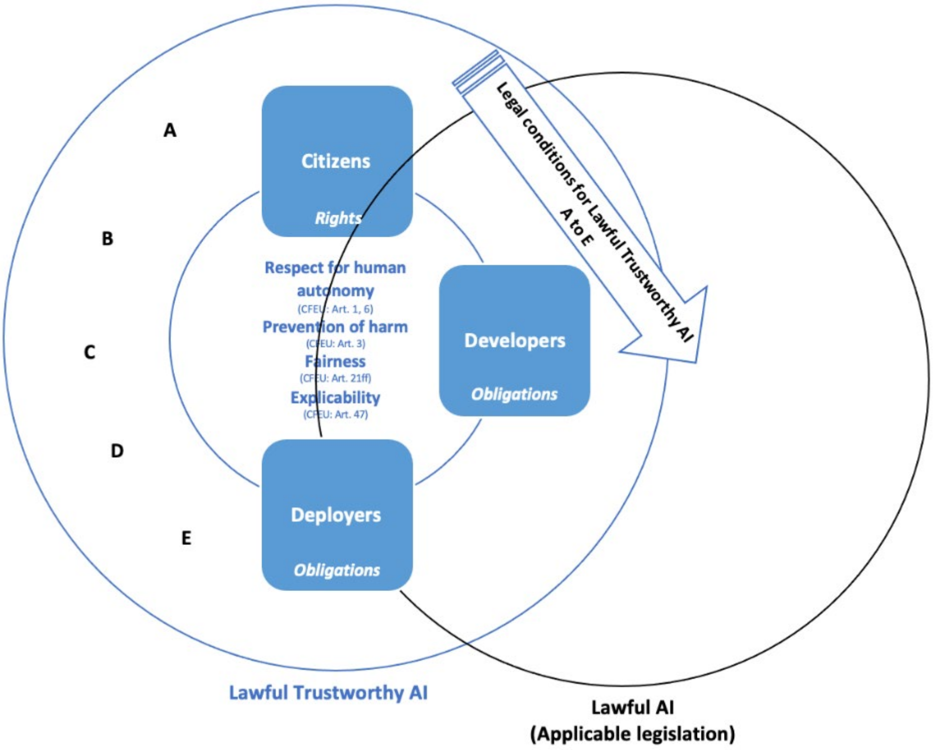


Figure 2. *Summary of findings (source: author)*

***Table 3**. Summary of findings and recommendations (source: author)*

Ethical and Robust AI	Lawful AI			Lawful Trustworthy AI			
AI HLEG ethical principles	**Addressee/s (AI Act)**	**Relevant articles (AI Act)**	**Main gaps in extant legislation**	**ID**	**Legal condition**	**Explanation/recommendations**	**Main stakeholder/s**
Respect for human autonomy: (CFEU: Articles 1 and 6) • Human agency and oversight *Prevention of harm:* (CFEU: Article 3) • Technical robustness and safety • Privacy and data governance • Societal and environmental well-being *Fairness:* (CFEU: Article 21*ff*) • Diversity, non-discrimination, and fairness • Societal and environmental well-being • Accountability *Explicability:* (CFEU: Article 47) • Transparency	Deployers	15 and 27; Recital 165	Applicable to high-risk AI systems only Fundamental rights impact assessments for cybersecurity Participation of citizens	**A**	Fundamental rights impact assessments for Robust AI	Inclusion of citizens in the conduct of fundamental rights impact assessments Conduct of fundamental rights impact assessments for AI systems other than high-risk Fundamental rights impact assessments for achieving appropriate level of cybersecurity and robust AI Participation of citizens in periodic reviews of the criteria of fundamental rights impact assessments Learning from environmental impact assessment mechanisms	Citizens, Deployers

(*Continued*)

Ethical and Robust AI	Lawful AI			Lawful Trustworthy AI			
AI HLEG ethical principles	**Addressee/s (AI Act)**	**Relevant articles (AI Act)**	**Main gaps in extant legislation**	**ID**	**Legal condition**	**Explanation/recommendations**	**Main stakeholder/s**
Respect for human autonomy: (CFEU: Articles 1 and 6) • Human agency and oversight *Prevention of harm:* (CFEU: Article 3) • Technical robustness and safety • Privacy and data governance • Societal and environmental well-being *Fairness:* (CFEU: Article 21*ff*) • Diversity, non-discrimination, and fairness • Societal and environmental well-being • Accountability *Explicability:* (CFEU: Article 47) • Transparency	Citizens, Deployers	50 and 85 to 86; Recital 165	Information rights on Individual decision-making applicable/ enforceable regarding high-risk AI systems only Collective complaints Participation of citizens	**B**	Effective redress mechanisms	Mechanisms complementary to Regulation (EU) 2016/679 (GDPR) Article 22 Increasing citizens' awareness about deployments of AI systems for administrative decision-making purposes with a legally binding effect Meaningful information about redress options and mechanisms under both EU and national law for affected citizens (applicable beyond only high-risk AI systems) Option for lodging collective complaints similar to Regulation (EU) 2016/679 (GDPR) Article 80 Citizen deliberation over and participation in the design of digital redress mechanisms (following the principle of "legal protection by design")	Citizens, Deployers

(*Continued*)

Ethical and Robust AI	Lawful AI			Lawful Trustworthy AI			
AI HLEG ethical principles	**Addressee/s (AI Act)**	**Relevant articles (AI Act)**	**Main gaps in extant legislation**	**ID**	**Legal condition**	**Explanation/recommendations**	**Main stakeholder/s**
Respect for human autonomy: (CFEU: Articles 1 and 6) • Human agency and oversight *Prevention of harm:* (CFEU: Article 3) • Technical robustness and safety • Privacy and data governance • Societal and environmental well-being *Fairness:* (CFEU: Article 21*ff*) • Diversity, non-discrimination, and fairness • Societal and environmental well-being • Accountability *Explicability:* (CFEU: Article 47) • Transparency	Deployers, Developers	4 and 70(3); Recitals 91 and 165	Higher education funding Socio-technical digital literacy of developers, deployers and citizens	**C**	Socio-technical digital literacy	Increasing capabilities of citizens to participate in the development and assessment of Trustworthy AI systems through *adequate* funding for interdisciplinary education programs and curricula on AI from preschool to higher education Enabling lifelong learning and promoting open universities Adoption of the method of co-creation in higher education Review of tech collaborations at public universities	Citizens, Deployers, Developers

(*Continued*)

Ethical and Robust AI	Lawful AI			Lawful Trustworthy AI			
AI HLEG ethical principles	**Addressee/s (AI Act)**	**Relevant articles (AI Act)**	**Main gaps in extant legislation**	**ID**	**Legal condition**	**Explanation/recommendations**	**Main stakeholder/s**
Respect for human autonomy: (CFEU: Articles 1 and 6) • Human agency and oversight *Prevention of harm:* (CFEU: Article 3) • Technical robustness and safety • Privacy and data governance • Societal and environmental well-being *Fairness:* (CFEU: Article 21*ff*) • Diversity, non-discrimination, and fairness • Societal and environmental well-being • Accountability *Explicability:* (CFEU: Article 47) • Transparency	Deployers (National competent authorities), Developers	6 to 27, 50, 65 to 67, 70, 74, 77 and 112; Recital 165	Risk-based approach Adequate funding for national supervisory authorities Information exchange between national competent authorities and national human rights institutes Amendments to list of high-risk AI systems in Annex III	**D**	Monitoring and enforcement capacity of national supervisory authorities	Adequate funding for national supervisory authorities Monitoring of AI systems other than high-risk AI systems Contextualisation of potential human rights harms through participation of citizens in fundamental rights impact assessments (applicable beyond high-risk AI systems) Collaboration and information exchange on AI human rights monitoring experiences between national human rights institutes and market surveillance authorities built on experience around public AI registers in cities Extensive collaboration among Commission, EU AI Board and Advisory Forum Amendments to list of high-risk AI systems in Annex III by national human rights institutes and, on a case-by-case basis, also citizens	Deployers (National competent authorities), Citizens, Developers

(*Continued*)

Ethical and Robust AI		Lawful AI				Lawful Trustworthy AI	
AI HLEG ethical principles	**Addressee/s (AI Act)**	**Relevant articles (AI Act)**	**Main gaps in extant legislation**	**ID**	**Legal condition**	**Explanation/recommendations**	**Main stakeholder/s**
Respect for human autonomy: (CFEU: Articles 1 and 6) • Human agency and oversight *Prevention of harm:* (CFEU: Article 3) • Technical robustness and safety • Privacy and data governance • Societal and environmental well-being *Fairness:* (CFEU: Article 21*ff*) • Diversity, non-discrimination, and fairness • Societal and environmental well-being • Accountability *Explicability:* (CFEU: Article 47) • Transparency	Developers, Deployers	Recital 165	Participatory design and co-assessment methodologies	**E**	Fora for participatory design and the inspection process for Ethical AI	Establish fora in existing public institutions and adopt co-design methodologies, particularly the structured method of value sensitive design and the inspection process for Ethical AI and enable the participation of citizens, developers and deployers, including a diverse group of domain-specific experts from both the public and private sector, in the development and assessment of AI systems in the public sector (beyond high-risk practises) Design these public fora such that all interested citizens, including disabled or visually and/or hearing impaired, are enabled to meaningfully participate in the development and assessment of AI systems in the public sector	Citizens, Deployers, Developers

Source: Author

4.3 Limitations

Several challenges to the realisation of the conditions and requirements exist. By only invoking the language of human rights, the interactions between citizens and AI systems, the institutions, and the human beings behind them will not become more fair, transparent, and accountable when deploying AI systems (Kleizen *et al.*, 2023). Trustworthy AI must be brought to action. Yet, risk aversion and blame avoidance among decision-makers in policy innovations (Hood, 2010; Howlett, 2014) could complicate the implementation of the proposed requirements, particularly since implementing them may take time and may not lead to immediate policy success. Similarly, overly constrained budgets in the Member States might, yet should not impede their realisation, specifically but not only when it comes to conditions three and five for increasing citizens' capabilities to critically partake in shaping the technological normativity of AI. Achieving effective communication to ensure that both public and private stakeholders understand the intentions of governmental AI projects and can participate in their implementation presents an additional hurdle (Kerikmäe & Pärn-Lee, 2021).

The proposed human rights-based conditions are not a panacea or a strive for perfection to address the complex landscape of challenges that the introduction of AI in the public sector carries, specifically since AI is a wicked problem (Austin & Háji, 2023). If one were to apply these requirements to other cultures and contexts, they ought to be adjusted *with* the involvement of the citizens, civil society, business communities, and public authorities, depending on their needs, capabilities, and capacities. Human rights approaches themselves need to be self-reflective to the extent that they do not impose a morality superior over other approaches to AI regulation (Karppinen & Puukko, 2020).

Additionally, conducting interdisciplinary normative research in the field of human rights and involving interdisciplinary perspectives may carry the risk of focusing on the law as it should be (*lex ferenda*) rather than as it is (*lex lata*) (Coomans *et al.*, 2010). While this argument is relevant, it must be evaluated in the context of recent legislative actions on data protection and privacy, cybersecurity, and AI. Considering the threats the development and use of AI can pose to democracy, the rule of law, and fundamental rights (Murray, 2021; Nemitz, 2018, 2021; Nemitz & Pfeffer, 2020), this thesis responds to the growing call in the legal research community to adopt an interdisciplinary approach to regulating AI (Hoffmann-Riem, 2021).

Thus, by combining socio-technical and human rights perspectives, the thesis attempts to show the current limits in the law for addressing potential human rights harms and promoting the fundamental rights of individuals and groups concerning the development and public-sector deployment of AI systems. It highlights the limits of traditional legal approaches for addressing these *lacunae*, specifically from the perspective of citizens interacting with AI. To protect fundamental rights in the context of *Law 3.0* (Brownsword, 2016, 2020, 2022), including interdisciplinary perspectives that can enhance public participation in the development and assessment of AI systems and allow citizens to influence the value choices made by AI developers is imperative. As shown in **I-III** and further discussed in this chapter, this necessitates implementing mechanisms that can involve developers, deployers, and citizens in joint assessment and development processes. These mechanisms include fundamental rights impact assessments, the VSD participatory design approach, and the co-assessment inspection process for Ethical AI. Additionally, realising Trustworthy AI involves improving subsequent enforcement and monitoring of AI legislation based on fundamental rights as well as providing adequate funding for interdisciplinary education programs and curricula on AI from preschool to higher education.

5 Concluding remarks and avenues for future research

To adequately protect and promote the fundamental rights of individuals and groups regarding the development and deployment of AI systems in the public sector, this thesis suggests **adopting a precautionary rather than a permissive approach to regulating AI**. Each of the proposed conditions, including the specific requirements, aims to address the normative concerns identified in this thesis. By combining human rights with socio-technical perspectives, the thesis proposes an inherently bottom-up, interdisciplinary, collaborative approach to AI regulation. Inclusive of the needs and capabilities of citizens, the proposed legal and institutional requirements are aimed at both protecting and promoting the fundamental rights of individuals and groups. In highlighting some significant limitations in the EU legal frameworks examined in this thesis, the author suggests that Trustworthy AI can be realised if governments take the following actions:

- Create a robust and resilient ICT infrastructure in the public and private sectors. This includes strengthening the role of citizens in cybersecurity mitigation and response efforts, particularly through citizen participation in fundamental rights impact assessments before deploying AI systems in the public sector.
- Establish effective redress mechanisms, specifically by extending information rights to citizens about AI-informed administrative decisions.
- Allocate adequate funding for interdisciplinary education programs and curricula to increase the socio-technical digital literacy of citizens in the long term.
- Provide adequate funding to national supervisory authorities to increase their capacity to monitor compliance with human rights obligations across all types of AI systems and enforce fundamental rights.
- Establish fora in existing public institutions and implement the systems thinking and participatory design methods of *value sensitive design* and *the inspection process for Ethical AI* to enable public and private stakeholders, developers, deployers, and citizens, including domain-specific experts (doctors, lawyers, and so forth), to translate and embed the ethical principles of the AI HLEG and applicable legal requirements into AI systems for various application areas in the public sector.

Based on the findings of the thesis, several topics for future research emerge.

First, further research is required into how the ethical principles of the AI HLEG and applicable legal requirements can be translated for different AI application contexts in the public sector, including for the implementation of SM in residential households (**II**). To this end, empirical studies are needed on the practicability of the inspection process for Ethical AI in different application areas and beyond the area of healthcare, where the assessment process has shown its practicality (Zicari *et al.*, 2021a).

Secondly, at the EU level, further case law is needed to clarify public entities' obligations in the use of AI systems and individuals' rights under extant data protection legislation. While the CJEU's recent judgment clarified that Article 22 contains a prohibition concerning decisions reached by an AI system, particularly the assignment of "scores" or the creation of "risk profiles" by automated processing (Case C-634/21, 2023; Opinion of Advocate General Pikamäe, 2023; Silveira, 2023), further legal interpretation is needed on which AI use cases might constitute automated individual decision-making under Article 22 of the GDPR. At the national level, further judgments from Supreme or Constitutional Courts are needed to specify which AI practices can be considered constitutional and which should be banned in addition to those outlawed by the AI Act (see also recent judgment by the First Senate of the Federal Constitutional Court of

Germany (*Bundesverfassungsgericht*) on the legislation in Hesse and Hamburg regarding the use of automated data analysis for predictive policing: Case 1 BvR 1547/19, 2023).

Third, while in the literature on the regulation of AI, the argument is often invoked that human rights-based approaches might hinder innovation and set additional bureaucratic hurdles for businesses, human rights scholarship has only conducted conceptual and theoretical studies to counter these arguments (McGregor *et al.*, 2019; Yeung *et al.*, 2020). Therefore, qualitative and quantitative empirical studies are needed to assess the economic effects of adopting human rights-based approaches to AI regulation. To this end, the practicality of the method of empirical legal studies should be evaluated (Altwicker, 2019; Ballin, 2020). Yet, statistics should be used with caution in this regard due to the problem of causality and its potential for not accounting for contextual factors (Alston, 1992). When assessing the adequacy of fundamental rights protection, including, for instance, how adequately the obligations and rights in the AI Act can protect individuals and groups, it should be assessed in combination with qualitative methods (Coomans *et al.*, 2010).

Fourth, more participatory regulatory approaches to AI beyond the risk-based approach in the AI Act need to be sought (**III**). The proposed legal and institutional requirements in this thesis are only a first contribution to this debate. Parallel to the implementation of the AI Act, further study of the borderline cases of AI applications and their relation to the categories of *prohibited AI practices, high-risk AI systems,* and *certain AI systems* other than high-risk is necessary. This could be done by employing fundamental rights impact and environmental risk assessment mechanisms.

Fifth, additional qualitative and quantitative data ought to be collected to study citizens' perceptions of Trustworthy AI, focusing on effective redress mechanisms, fundamental rights impact assessments, and participatory design methods (**III**). Similarly, more participatory data collection methods should be explored and implemented for broader stakeholder involvement beyond conventional data collection methods such as surveys (Bergold & Thomas, 2012; de Vos *et al.*, 2021). Equally important is collecting more data on the environmental impact of the development and use of AI systems, and their potential for climate mitigation efforts (**III**), including concerning the effectiveness of SM in this context (**II**).

Setting standards for developers and deployers of AI systems to align the use of the technology with human values, fundamental rights, and the rule of law is a societal task that requires the involvement of all scientific disciplines and relevant stakeholders equally (Nyman-Metcalf & Kerikmäe, 2020; **I-III**). To realise Trustworthy AI or adequate protection and promotion of fundamental rights concerning AI, this thesis, therefore, argues for implementing an inherently interdisciplinary, human rights-based approach, involving legislators, policymakers, developers, deployers, *and* citizens in AI regulation.

In essence, realising Trustworthy AI *wisely* implies that human rights should not be "confined within the judicial model [and the code of AI systems] in which [they could be] incarcerated" (Sen, 2004, p. 319; adapted by the author). It also means decentring technology sometimes and reflecting on whether other approaches could address a societal problem more effectively than relying on AI systems alone. It is also important sometimes to pose the right questions and leave room for them to be heard, for instance, about whether developing some AI systems in the first place is beneficial for society at all. More importantly, realising Trustworthy AI in the public sector wisely involves empowering citizens, being responsive to their needs, and attempting to collaboratively *enhance understanding* of the root causes of societal problems over the long term *before* relying primarily on AI technologies to *solve* them.

List of figures

List of tables

References

Ad Hoc Committee on Artificial Intelligence (CAHAI). (2020, December 17). *Feasibility Study*. (Report CAHAI(2020)23). The Council of Europe. https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

AI Now Institute. (2023, April 11). *Algorithmic Accountability: Moving Beyond Audits*. https://ainowinstitute.org/publication/algorithmic-accountability?fbclid=IwAR0s9-oJil0XWO3ONS0el7-2_M_s029NX1bhEj_uuj6pFSxcg6ywOlm0b44#footnote-list-27

Aizenberg, E., & van den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society, 7*(2). https://doi.org/10.1177/2053951720949566

Alston, P. (1992). *The United Nations and Human Rights: A Critical Appraisal*. Clarendon Press.

Alston, P. (2019, September 26). *Brief by the United Nations Special Rapporteur on extreme poverty and human rights as amicus Curiae in the case of NJCMc. s. De Staat der Nederlanden (SyRI) before the District Court of The Hague (case number: C/09/550982/HAZA18/388)*. United Nations. The Office of the High Commissioner for Human Rights. https://www.ohchr.org/sites/default/files/Documents/Issues/Poverty/Amicusfinalversionsigned.pdf

Alston, P. (2020, February 5). *Landmark ruling by Dutch court stops government attempts to spy on the poor - UN expert*. United Nations: The Office of the High Commissioner for Human Rights. https://www.ohchr.org/en/press-releases/2020/02/landmark-ruling-dutch-court-stops-government-attempts-spy-poor-un-expert?LangID=E&NewsID=25522

Altwicker, T. (2019). International Legal Scholarship and the Challenge of Digitalization. *Chinese Journal of International Law, 18*(2), 217-246. https://doi.org/10.1093/chinesejil/jmz012

Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 — C9-0146/2021 — 2021/0106(COD)), OJ C, C/2024/506, 23.1.2024, ELI: http://data.europa.eu/eli/C/2024/506/oj

Amnesty International, Access Now. (2018, May 16). *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems*. https://www.torontodeclaration.org/declaration-text/english/

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973-989. https://doi.org/10.1177/1461444816676645

Anastasopoulos, L. J., & Whitford, A. B. (2019). Machine Learning for Public Administration Research, With Application to Organizational Reputation. *Journal of Public Administration Research and Theory, 29*(3), 491-510. https://doi.org/10.1093/jopart/muy060

Austin, Z., & Háji, A. (2023). Regulation of Wicked Problems: Opportunities, Responsibilities, and Threats. *Journal of Medical Regulation, 109*(3), 6-11. https://doi.org/10.30770/2572-1852-109.3.6

Ballin, E. H. (2020). *Advanced Introduction to Legal Research Methods*. Edward Elgar Publishing.

Barocas, S. (2014). Data Mining and the Discourse on Discrimination. *Data Ethics Workshop organised in conjunction with the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14), New York City, August 24*. https://dataethics.github.io/

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. *9th Annual Conference of the Special Interest Group for Computing, Information and Society (SIGCIS), Philadelphia, October 29*. https://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press.

Beduschi, A. (2020, February). *Human Rights and the Governance of Artificial Intelligence* [Research Brief]. Geneva Academy of International Humanitarian Law and Human Rights. https://www.geneva-academy.ch/research/publications/detail/513-human-rights-and-the-governance-of-artificial-intelligence

Bergold, J., & Thomas, S. (2012). Participatory Research Methods: A Methodological Approach in Motion. *Forum: Qualitative Social Research, 13*(1), 1-35. https://doi.org/10.17169/fqs-13.1.1801

Berlin Declaration on Digital Society. (2020). *Berlin Declaration on Digital Society and Value-Based Digital Government at the ministerial meeting during the German Presidency of the Council of the European Union on 8 December 2020.* European Commission. https://ec.europa.eu/isa2/sites/isa/files/cdr_20201207_eu2020_berlin_declaration_on_digital_society_and_value-based_digital_government_.pdf

Bex, F. (2023). Transdisciplinary research as a way forward in AI & Law. *Conference on Cross-disciplinary Research in Computational Law (CRCL): Computational "Law" on edge, Brussels, November 20-21*. https://www.cohubicol.com/assets/uploads/crcl23/bex_position_paper_crcl23.pdf

Bjørnskov, C. (2018). Social Trust and Economic Growth. In E. M. Uslaner (Ed.), *The Oxford Handbook of Social and Political Trust* (pp. 535-556). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190274801.013.24

Black, J., & Murray, A. D. (2019). Regulating AI and Machine Learning: Setting the Regulatory Agenda. *European Journal of Law and Technology, 10*(3), 1-17. http://eprints.lse.ac.uk/102953/4/722_3282_1_PB.pdf

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bria, F. (2017). Public Policies for Digital Sovereignty. *Platform Cooperativism Consortium Conference, New York City, November 10-11*. https://www.academia.edu/19102224/Public_policies_for_digital_sovereignty

Brownsword, R. (2016). Technological management and the Rule of Law. *Law, Innovation and Technology, 8*(1), 100-140. https://doi.org/10.1080/17579961.2016.1161891

Brownsword, R. (2019). Legal Regulation of Technology: Supporting Innovation, Managing Risk, and Respecting Values. In T. L. Pittinsky (Ed.), *Science, Technology, and Society: New Perspectives and Directions* (pp. 109-137). Cambridge University Press. https://doi.org/10.1017/9781316691489.005

Brownsword, R. (2020). *Law 3.0: Rules, Regulation, and Technology*. Routledge. https://doi.org/10.4324/9781003053835

Brownsword, R. (2022). *Rethinking Law, Regulation, and Technology*. Edward Elgar Publishing. https://doi.org/10.4337/9781800886476

Brownsword, R., & Goodwin, M. (2012). *Law and the Technologies of the Twenty-First Century: Text and Materials*. Cambridge University Press. https://doi.org/10.1017/CBO9781139047609

Bryson, J. (2018, November 13). AI & Global Governance: No One Should Trust AI. *United Nations Centre for Policy Research*. https://unu.edu/cpr/blog-post/ai-global-governance-no-one-should-trust-ai

Buiten, M. C. (2019). Towards Intelligent Regulation of Artificial Intelligence. *European Journal of Risk Regulation, 10*(1), 41-59. https://doi.org/10.1017/err.2019.8

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In A. F. Sorelle & W. Christo (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, ACM, New York City, February 23-24* (Vol. 81, pp. 77-91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951715622512

Busuioc, M. (2021). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review, 81*(5), 825-836. https://doi.org/10.1111/puar.13293

BVerfG, Judgment of the First Senate of 16 February 2023 - 1 BvR 1547/19 -, paras. 1-178, https://www.bverfg.de/e/rs20230216_1bvr154719en.html

Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. *U.C. Davis Law Review, 51*(2), 399-436. https://heinonline.org/HOL/P?h=hein.journals/davlr51&i=413

Calo, R. (2022, April 9). The Scale and the Reactor. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.4079851

Cantero Gamito, M., & Gentile, G. (2023). Algorithms, rule of law, and the future of justice: implications in the Estonian justice system (Issue Brief No. 27). http://doi.org/10.2870/640834

Celeste, E. (2021). Digital sovereignty in the EU: challenges and future perspectives. In F. Fabbrini, E. Celeste, & J. Quinn (Eds.), *Data Protection Beyond Borders: Transatlantic Perspectives on Extraterritoriality and Sovereignty* (pp. 211-228). Hart Publishing (Bloomsbury). https://doi.org/10.5040/9781509940691.ch-013

Charter of Fundamental Rights of the European Union (2016) OJ C 202, 7.6.2016, pp. 389-405, ELI: http://data.europa.eu/eli/treaty/char_2016/oj

Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., & Yeung, K. (2021). Trustworthy AI. In B. Braunschweig & M. Ghallab (Eds.), *Reflections on Artificial Intelligence for Humanity. Lecture Notes in Computer Science, vol 12600* (pp. 13-39). Springer Cham. https://doi.org/10.1007/978-3-030-69128-8_2

Chesterman, S. (2021). *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law*. Cambridge University Press. https://doi.org/10.1017/9781009047081

Chiusi, F., Alfter, B., Ruckenstein, M., & Lehtiniemi, T. (2020). Automating Society Report 2020. *AlgorithmWatch & Bertelsmann Stiftung*. https://automatingsociety.algorithmwatch.org/

Citron, D. K., & Solove, D. J. (2022). Privacy Harms. *Boston University Law Review, 102*(3), 793-864. https://heinonline.org/HOL/P?h=hein.journals/bulr102&i=815

Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., & D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. *DIGITAL HEALTH, 9*. https://doi.org/10.1177/20552076231183542

Coglianese, C., & Lehr, D. (2017). Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal, 105*(5), 1147-1224. https://heinonline.org/HOL/P?h=hein.journals/glj105&i=1166

Cohen, J. E. (2012). *Configuring the networked self: Law, code, and the play of everyday practice*. Yale University Press.

Cohen, J. E. (2017). Affording Fundamental Rights: A Provocation Inspired by Mireille Hildebrandt. *Critical Analysis of Law, 4*(1), 78-90. https://heinonline.org/HOL/P?h=hein.journals/cclaysolw4&i=78

Cohen, J. E. (2019). *Between Truth and Power: The Legal Constructions of Informational Capitalism*. Oxford University Press.

Compagnucci, M. C. (2022). Danish DPA Banned the Use of Google Chromebooks and Google Workspace in Schools in Helsingor Municipality. *European Data Protection Law Review (EDPL), 8*(3), 405-411. https://heinonline.org/HOL/P?h=hein.journals/edpl8&i=422

Consolidated Version of the Treaty on European Union (2016) OJ C 202, 7.6.2016, p. 13, ELI: http://data.europa.eu/eli/treaty/teu_2016/oj

Coomans, F., Grünfeld, F., & Kamminga, M. T. (2010). Methods of Human Rights Research: A Primer. *Human Rights Quarterly, 32*(1), 179-186. http://www.jstor.org/stable/40390006

Crawford, K. (2013, April 1). The Hidden Biases in Big Data. *Harvard Business Review*. https://hbr.org/2013/04/the-hidden-biases-in-big-data

Creswell, J. W. (2012). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches* (3rd ed.). Sage.

Crootof, R., & Ard, B. J. (2021). Structuring techlaw. *Harvard Journal of Law & Technology (Harvard JOLT), 34*(2), 347-418. https://heinonline.org/HOL/P?h=hein.journals/hjlt34&i=361

Cuijpers, C., & Koops, B.-J. (2013). Smart Metering and Privacy in Europe: Lessons from the Dutch Case. In S. Gutwirth, R. Leenes, P. de Hert, & Y. Poullet (Eds.), *European Data Protection: Coming of Age* (pp. 269-293). Springer. https://doi.org/10.1007/978-94-007-5170-5_12

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Governance of AI Program. Future of Humanity Institute. University of Oxford. https://www.fhi.ox.ac.uk/govaiagenda

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical Data Studies: A dialog on data and space. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951716648346

De Gregorio, G. (2021). The rise of digital constitutionalism in the European Union. *International Journal of Constitutional Law, 19*(1), 41-70. https://doi.org/10.1093/icon/moab001

de Vos, A., Preiser, R., & Masterson, V. A. (2021). Participatory data collection. In R. Biggs, A. d. Vos, R. Preiser, H. Clements, K. Maciejewski, & M. Schlüter (Eds.), *The Routledge handbook of research methods for social-ecological systems* (pp. 119-134). Routledge. https://doi.org/10.4324/9781003021339-10

Deibert, R. J. (2018). Toward a Human-Centric Approach to Cybersecurity. *Ethics & International Affairs 32*(4), 411-424. https://doi.org/10.1017/s0892679418000618

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer Cham. https://doi.org/10.1007/978-3-030-30371-6

Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union, OJ L 194, 19.7.2016, p. 1–30, ELI: http://data.europa.eu/eli/dir/2016/1148/oj

Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) (Text with EEA relevance), OJ L 333, 27.12.2022, p. 80-152, ELI: http://data.europa.eu/eli/dir/2022/2555/oj

Donnelly, J. (2013). *Universal Human Rights in Theory and Practice*. Cornell University Press. http://www.jstor.org/stable/10.7591/j.ctt1xx5q2

Drechsler, W., & Kostakis, V. (2014). Should Law Keep Pace With Technology? Law as Katechon. *Bulletin of Science, Technology & Society, 34*(5-6), 128-132. https://doi.org/10.1177/0270467615574330

Ebers, M. (2020). Regulating AI and Robotics: Ethical and Legal Challenges. In M. Ebers & S. Navas (Eds.), *Algorithms and Law* (pp. 37-99). Cambridge University Press. https://doi.org/10.1017/9781108347846.003

Edwards, L. V., Michael. (2017-2018). Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking for. *Duke Law & Technology Review, 16*, 18-84. https://heinonline.org/HOL/P?h=hein.journals/dltr16&i=18

Espinosa Apráez, B. (2022). The challenges of sharing data at the intersection of EU data protection and electricity market legislation: lessons from the Netherlands. *Journal of Energy & Natural Resources Law, 41*(4), 403-429. https://doi.org/10.1080/02646811.2022.2143673

Etzioni, A. (2007). Are New Technologies the Enemy of Privacy? *Knowledge, Technology & Policy, 20*, 115-119. https://doi.org/10.1007/s12130-007-9012-x

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Commission. (2018, April 25). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe* (COM/2018/237 final). Publications Office of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

European Commission. (2019, December 11). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: The European Green Deal* (COM/2019/640 final). Publications Office of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2019%3A640%3AFIN

European Commission. (2020a, Feburary 19). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Shaping Europe's digital future* (COM/2020/67 final). Publications Office of the European Union. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52020DC0067

European Commission. (2020b, February 19). *White Paper on Artificial Intelligence: A European approach to excellence and trust* (COM/2020/65 final). Publications Office of the European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0065&qid=1716208867333

European Commission. (2021, April 21). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Fostering a European approach to Artificial Intelligence* (COM/2021/205 final). Publications Office of the European Union. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM%3A2021%3A205%3AFIN

European Commission. (2022, January 26). *European Declaration on Digital Rights and Principles for the Digital Decade* (COM(2022) 28 final). Publications Office of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022DC0028

European Council. (2017, October 19). *European Council conclusions on migration, digital Europe, security and defence*. https://www.consilium.europa.eu/en/press/press-releases/2017/10/19/euco-conclusions-migration-digital-defence/

European Group on Ethics in Science and New Technologies. (2018). *Statement on artificial intelligence, robotics and 'autonomous' systems*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/531856

European Group on Ethics in Science and New Technologies. (2021). *Values for the Future: The Role of Ethics in European and Global Governance*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/595827

Evas, T. (2024). The EU Artificial Intelligence Act. *Journal of AI Law and Regulation, 1*(1), 98-101. https://doi.org/10.21552/aire/2024/1/11

Exec. Order No. 13,960, 85 Fed. Reg. 78939 (December 3, 2020).

Exec. Order No. 14,110, 88 Fed. Reg. 75191 (October 30, 2023).

Executive Office of the President. (2016a, May). *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Executive Office of the President. (2016b, October). *Preparing for the Future of Artificial Intelligence*. National Science and Technology Council Committee on Technology. The White House Office of Science and Technology Policy. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Executive summary of Opinion No 4/2015 of the European Data Protection Supervisor, 'Towards a new digital ethics: Data, dignity and technology', OJ C 392, 25.11.2015, p. 9–10. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52015XX1125%2801%29&qid=1716392213528

Fanni, R., Steinkogler, V. E., Zampedri, G., & Pierson, J. (2023). Enhancing human agency through redress in Artificial Intelligence Systems. *AI & Society, 38*(2), 537-547. https://doi.org/10.1007/s00146-022-01454-7

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Berkman Klein Center for Internet & Society. http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

Floridi, L. (2020a). Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki. *Philosophy & Technology, 33*(4), 541-546. https://doi.org/10.1007/s13347-020-00434-3

Floridi, L. (2020b). The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU. *Philosophy & Technology, 33*(3), 369-378. https://doi.org/10.1007/s13347-020-00423-6

Floridi, L. (2020c). What the Near Future of Artificial Intelligence Could Be. In C. Burr & S. Milano (Eds.), *The 2019 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook.* (pp. 127-142). Springer Cham. https://doi.org/10.1007/978-3-030-29145-7_9

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Freiman, O. (2022). Making sense of the conceptual nonsense 'trustworthy AI'. *AI and Ethics, 3*(4), 1351-1360. https://doi.org/10.1007/s43681-022-00241-w

Friedman, B., Harbers, M., Hendry, D. G., van Den Hoven, J., Jonker, C., & Logler, N. (2021). Eight grand challenges for value sensitive design from the 2016 Lorentz workshop. *Ethics and Information Technology, 23*(1), 5-16. https://doi.org/10.1007/s10676-021-09586-y

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

Future of Life Institute. (2017, August 11). *Asilomar AI Principles*. https://futureoflife.org/open-letter/ai-principles/

G7. (2023). *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*. Ministry of Foreign Affairs of Japan. https://www.mofa.go.jp/files/100573473.pdf

G20. (2019). *G20 AI Principles*. Ministry of Foreign Affairs of Japan. https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf

Gantchev, V. (2019). Data protection in the age of welfare conditionality: Respect for basic rights or a race to the bottom? *European Journal of Social Security, 21*(1), 3-22. https://doi.org/10.1177/1388262719838109

Gasser, U., & Almeida, V. A. (2017). A Layered Model for AI Governance. *IEEE Internet Computing, 21*(6), 58-62. https://doi.org/10.1109/mic.2017.4180835

Gellert, R. (2021). The role of the risk-based approach in the General data protection Regulation and in the European Commission's proposed Artificial Intelligence Act: Business as usual? *Journal of Ethics and Legal Technologies, 3*(2), 15-33. https://doi.org/10.14658/pupj-JELT-2021-2-2

Gellert, R. (2022). Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms? *Regulation & Governance, 16*(1), 156-176. https://doi.org/10.1111/rego.12349

Gesk, T. S., & Leyer, M. (2022). Artificial intelligence in public services: When and why citizens accept its usage. *Government Information Quarterly, 39*(3), 101704. https://doi.org/10.1016/j.giq.2022.101704

Glukhin v. Russia, no. 11519/20, 4 July 2023, https://hudoc.echr.coe.int/?i=001-225655

Goodman, P. S., Ramanujam, R., Carroll, J. S., Edmondson, A. C., Hofmann, D. A., & Sutcliffe, K. M. (2011). Organizational errors: Directions for future research. *Research in Organizational Behavior, 31*, 151-176. https://doi.org/10.1016/j.riob.2011.09.003

Government of the People's Republic of China. (2019, June). *Governance Principles for New Generation AI: Develop Responsible Artificial Intelligence*. Ministry of Science and Technology, National New Generation AI Governance Expert Committee. https://digichina.stanford.edu/work/translation-chinese-expert-group-offers-governance-principles-for-responsible-ai/

Graber, C. B. (2017). Freedom and Affordances of the Net. *Washington University Jurisprudence Review, 10*(2), 221-256. https://heinonline.org/HOL/P?h=hein.journals/wujurisre10&i=233

Graber, C. B. (2021). How the Law Learns in the Digital Society. *Law, Technology and Humans, 3*(2), 12-27. https://doi.org/10.5204/lthj.1600

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review, 45*, 105681. https://doi.org/10.1016/j.clsr.2022.105681

Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artificial Intelligence and Law, 30*(3), 291-323. https://doi.org/10.1007/s10506-021-09294-4

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review, 55*(4), 1143-1185. https://doi.org/10.54648/cola2018095

Hasselbalch, G. (2021). *Data Ethics of Power: A Human Approach in the Big Data and AI Era*. Edward Elgar Publishing.

Henkin, L. (1990). *The Age of Rights*. Columbia University Press.

Hervey, T., Cryer, R., Sokhi-Bulley, B., & Bohm, A. (2011). *Research Methodologies in EU and International Law*. Bloomsbury Publishing.

High-Level Expert Group on Artificial Intelligence. (2019a). *Ethics Guidelines for Trustworthy AI*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/346720

High-Level Expert Group on Artificial Intelligence. (2019b). *Policy and Investment Recommendations for Trustworthy AI*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/465913

High-Level Expert Group on Artificial Intelligence. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/002360

Hildebrandt, M. (2015). *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*. Edward Elgar Publishing.

Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), 20170355. https://doi.org/10.1098/rsta.2017.0355

Hildebrandt, M. (2020). *Law for Computer Scientists and Other Folk*. Oxford University Press. https://doi.org/10.1093/oso/9780198860877.001.0001

Hoffmann, T. (2020). Heuristics in Legal Decision-Making. *Acta Baltica Historiae et Philosophiae Scientiarum, 8*(1), 62-71. https://doi.org/10.11590/abhps.2020.1.03

Hoffmann-Riem, W. (2020). Legal Technology/Computational Law: Preconditions, Opportunities and Risks. *Journal of Cross-disciplinary Research in Computational Law, 1*(1). https://journalcrcl.org/crcl/article/view/7

Hoffmann-Riem, W. (2021). *Recht im Sog der digitalen Transformation: Herausforderungen [The Law in the Maelstrom of the Digital Transformation. Challenges]* (Vol. 11). Mohr Siebeck.

Hood, C. (1991). A Public Management for All Seasons? *Public Administration, 69*(1), 3-19. https://doi.org/10.1111/j.1467-9299.1991.tb00779.x

Hood, C. (2010). *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government*. Princeton University Press. https://doi.org/10.1515/9781400836819

Howlett, M. (2014). Why are policy innovations rare and so often negative? Blame avoidance and problem denial in climate change policy-making. *Global Environmental Change, 29*, 395-403. https://doi.org/10.1016/j.gloenvcha.2013.12.009

Huhta, K. (2020). Smartening up while keeping safe? Advances in smart metering and data protection under EU law. *Journal of Energy & Natural Resources Law, 38*(1), 5-22. https://doi.org/10.1080/02646811.2019.1622244

Hydén, H. (2020). AI, Norms, Big Data, and the Law. *Asian Journal of Law and Society, 7*(3), 409-436. https://doi.org/10.1017/als.2020.36

Hydén, H. (2021). *Sociology of Law as the Science of Norms*. Routledge. https://doi.org/10.4324/9781003241928

Ihde, D. (2009). *Postphenomenology and Technoscience: The Peking University Lectures*. Suny Press.

Jacobs, M., Kurtz, C., Simon, J., & Böhmann, T. (2021). Value Sensitive Design and power in socio-technical ecosystems. *Internet Policy Review, 10*(3). https://doi.org/10.14763/2021.3.1580

Jasanoff, S. (2016). *The Ethics of Invention: Technology and the Human Future*. WW Norton & Company.

Joamets, K., & Chochia, A. (2021). Artificial Intelligence for Persons with Disabilities: Legal and Ethical Questions Concerning the Application of Trustworthy AI. *Acta Baltica Historiae et Philosophiae Scientiarum, 9*(1), 51-66. https://doi.org/10.11590/abhps.2021.1.04

Jørgensen, R. F. (2023). Data and rights in the digital welfare state: the case of Denmark. *Information, Communication & Society, 26*(1), 123-138. https://doi.org/10.1080/1369118X.2021.1934069

Judgment of 7 December 2023, SCHUFA Holding (Scoring), C-634/21, ECLI:EU:C:2023:957.

Junklewitz, H., Hamon, R., André, A., Evas, T., Soler Garrido, J., & Sanchez Martin, J. (2023). *Cybersecurity of Artificial Intelligence in the AI Act*. Publications Office of the European Union. https://doi.org/10.2760/271009

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. *Nature Climate Change, 12*(6), 518-527. https://doi.org/10.1038/s41558-022-01377-7

Karppinen, K., & Puukko, O. (2020). Four Discourses of Digital Rights: Promises and Problems of Rights-Based Politics. *Journal of Information Policy, 10*, 304-328. https://doi.org/10.5325/jinfopoli.10.2020.0304

Katz, A., & Sander, G. G. (2019). *Staatsrecht: Grundlagen, Staatsorganisation, Grundrechte [Constitutional Law: Basics, State Organization and Fundamental Rights]* (19th ed.). C.F. Müller.

Kaun, A. (2022). Suing the algorithm: the mundanization of automated decision-making in public services through litigation. *Information, Communication & Society, 25*(14), 2046-2062. https://doi.org/10.1080/1369118x.2021.1924827

Kaun, A., Larsson, A. O., & Masso, A. (2023). Automating public administration: citizens' attitudes towards automated decision-making across Estonia, Sweden, and Germany. *Information, Communication & Society, 27*(2), 314-332. https://doi.org/10.1080/1369118X.2023.2205493

Kerikmäe, T. (2014). EU Charter: Its Nature, Innovative Character, and Horizontal Effect. In T. Kerikmäe (Ed.), *Protecting Human Rights in the EU* (pp. 5-19). Springer. https://doi.org/10.1007/978-3-642-38902-3_2

Kerikmäe, T., Hamuľák, O., & Chochia, A. (2016). A Historical Study of Contemporary Human Rights: Deviation or Extinction? *Acta Baltica Historiae et Philosophiae Scientiarum, 4*(2), 98-115. https://doi.org/10.11590/abhps.2016.2.06

Kerikmäe, T., & Nyman-Metcalf, K. (2012). Less is More or More is More? Revisiting Universality of Human Rights. *International and Comparative Law Review, 12*(1), 39-56. https://doi.org/10.1515/iclr-2016-0077

Kerikmäe, T., & Pärn-Lee, E. (2021). Legal dilemmas of Estonian artificial intelligence strategy: in between of e-society and global race. *AI & Society, 36*(2), 561-572. https://doi.org/10.1007/s00146-020-01009-8

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14-29. https://doi.org/10.1080/1369118x.2016.1154087

Kleizen, B., Van Dooren, W., Verhoest, K., & Tan, E. (2023). Do citizens trust trustworthy artificial intelligence? Experimental evidence on the limits of ethical AI measures in government. *Government Information Quarterly, 40*(4), 101834. https://doi.org/10.1016/j.giq.2023.101834

Klimburg, A. (2012). *National Cyber Security Framework Manual*. NATO Cooperative Cyber Defense Center of Excellence. https://ccdcoe.org/uploads/2018/10/NCSFM_0.pdf

Kranzberg, M. (1986). Technology and History: "Kranzberg's Laws". *Technology and Culture, 27*(3), 544-560. https://doi.org/10.2307/3105385

Kriebitz, A., & Lütge, C. (2020). Artificial Intelligence and Human Rights: A Business Ethical Assessment. *Business and Human Rights Journal, 5*(1), 84-104. https://doi.org/10.1017/bhj.2019.28

Kusch, M. (2007). Towards a Political Philosophy of Risk: Experts and Publics in Deliberative Democracy. In T. Lewens (Ed.), *Risk: Philosophical Perspectives* (pp. 141-165). Routledge. https://doi.org/10.4324/9780203962596

Lane, L. (2023). Artificial Intelligence and Human Rights: Corporate Responsibility in AI Governance Initiatives. *Nordic Journal of Human Rights, 41*(3), 304-325. https://doi.org/10.1080/18918131.2022.2137288

Larsson, S. (2019). The Socio-Legal Relevance of Artificial Intelligence. *Droit et société, 103*(3), 573-593. https://doi.org/10.3917/drs1.103.0573

Larsson, S. (2020). On the Governance of Artificial Intelligence through Ethics Guidelines. *Asian Journal of Law and Society, 7*(3), 437-451. https://doi.org/https://doi.org/10.1017/als.2020.19

Latonero, M. (2018). *Governing Artificial Intelligence: Upholding Human Rights & Dignity*. Data & Society. https://apo.org.au/sites/default/files/resource-files/2018-10/apo-nid196716.pdf

Laux, J. (2023). Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI & Society*, 1-14. https://doi.org/10.1007/s00146-023-01777-z

Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance, 18*(1), 3-32. https://doi.org/10.1111/rego.12512

Lavrijssen, S., Espinosa Apráez, B., & Ten Caten, T. (2022). The Legal Complexities of Processing and Protecting Personal Data in the Electricity Sector. *Energies, 15*(3), 1088. https://doi.org/10.3390/en15031088

Legal Information Institute. (2024). *Condition*. The Wex Legal Dictionary and Encyclopedia of the Legal Information Institute at Cornell Law School. https://www.law.cornell.edu/wex/condition

Lehr, D., & Ohm, P. (2017). Playing with the Data: What Legal Scholars Should Learn about Machine Learning. *U.C. Davis Law Review, 51*(2), 653-718. https://heinonline.org/HOL/P?h=hein.journals/davlr51&i=667

Lember, V., Brandsen, T., & Tõnurist, P. (2019). The potential impacts of digital technologies on co-production and co-creation. *Public Management Review, 21*(11), 1665-1686. https://doi.org/10.1080/14719037.2019.1619807

Lenaerts, K., & Gutiérrez-Fons, J. A. (2013). To Say What the Law of the EU Is: Methods of Interpretation and the European Court of Justice. *Columbia Journal of European Law, 20*(2), 3-[vi]. https://heinonline.org/HOL/P?h=hein.journals/coljeul20&i=183

Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M., & Briggs, M. (2021). *Artificial Intelligence, Human Rights, Democracy, and the Rule of Law: A Primer*. The Council of Europe. https://rm.coe.int/primer-en-new-cover-pages-coe-english-compressed-2754-7186-0228-v-1/1680a2fd4a

Lessig, L. (1999). The Law of the Horse: What Cyberlaw Might Teach. *Harvard Law Review, 113*(2), 501-549. https://doi.org/10.2307/1342331

Lessig, L. (2006). *Code: And Other Laws of Cyberspace, Version 2.0*. Basic Books.

Lindgren, S., & Dignum, V. (2023). Beyond AI solutionism: toward a multi-disciplinary approach to artificial intelligence in society. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* (pp. 163-172). Edward Elgar Publishing. https://doi.org/10.4337/9781803928562

Loi, M., & Spielkamp, M. (2021). Towards Accountability in the Use of Artificial Intelligence for Public Administrations. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), Association for Computing Machinery, New York, May 19-21*, 757-766. https://doi.org/10.1145/3461702.3462631

Luhmann, N. (1965). *Grundrechte als Institution. Ein Beitrag zur politischen Soziologie [Basic Rights as an Institution: A Contribution to Political Sociology]*. Duncker & Humblot.

Luhmann, N. (2014). *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität [Trust: A Mechanism for the Reduction of Social Complexity]* (5th ed.). UVK Verlag. https://doi.org/10.36198/9783838540047

Maas, M. M., & Villalobos, J. J. (2023). *International AI Institutions: A Literature Review of Models, Examples, and Proposals*. Legal Priorities Project. https://www.legalpriorities.org/documents/Maas%20-%20Villalobos%20-%20International%20AI%20Institutions.pdf

Makasi, T., Desouza, K. C., Nili, A., & Tate, M. (2022). Public Service Values and Chatbots in the Public Sector: Reconciling Designer efforts and User Expectations. *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS-55), January 3-7*, 2334-2343. http://hdl.handle.net/10125/79625

Makasi, T., Nili, A., Desouza, K. C., & Tate, M. (2021). A Typology of Chatbots in Public Service Delivery. *IEEE Software, 39*(3), 58-66. https://doi.org/10.1109/MS.2021.3073674

Malgieri, G., & Pasquale, F. (2024). Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. *Computer Law & Security Review, 52*, 105899. https://doi.org/10.1016/j.clsr.2023.105899

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review, 34*(4), 754-772. https://doi.org/10.1016/j.clsr.2018.05.017

Mantelero, A. (2019). *Artificial Intelligence and Data Protection: Challenges and Possible Remedies* (Report T-PD(2018)09Rev). The Council of Europe. https://rm.coe.int/report-on-artificial-intelligence-artificial-intelligence-and-data-pro/16808e6012

Mantelero, A. (2022). *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*. T.M.C. Asser Press. https://doi.org/10.1007/978-94-6265-531-7

Marcus, J. S., Petropoulos, G., & Yeung, T. (2019). *Contribution to Growth: The European Digital Single Market. Delivering economic benefits for citizens and businesses*. Study for the Committee on the Internal Market and Consumer Protection. Policy Department for Economic, Scientific and Quality of Life Policies. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2019/631044/IPOL_STU(2019)631044_EN.pdf

Martínez-Ramil, P. (2022). Discriminatory algorithms. A proportionate means of achieving a legitimate aim? *Journal of Ethics and Legal Technologies, 4*(1), 3-24. https://doi.org/10.14658/pupj-JELT-2022-1-2

Martini, M. (2020). Regulating Algorithms: How to Demystify the Alchemy of Code? In M. Ebers & S. Navas (Eds.), *Algorithms and Law* (pp. 100-135). Cambridge University Press. https://doi.org/10.1017/9781108347846.004

Martini, M., Möslein, F., & Rostalski, F. (2024). *Recht der Digitalisierung: Legal Tech*. Nomos.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McGregor, L., Murray, D., & Ng, V. (2019). International Human Rights Law as a Framework for Algorithmic Accountability. *International and Comparative Law Quarterly, 68*(2), 309-343. https://doi.org/10.1017/S0020589319000046

Misuraca, G., & van Noordt, C. (2020). *AI Watch-Artificial Intelligence in public services: Overview of the use and impact of AI in public services in the EU*. Publications Office of the European Union. https://doi.org/10.2760/039619

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501-507. https://doi.org/10.1038/s42256-019-0114-4

Mökander, J., & Floridi, L. (2021). Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines, 31*(2), 323-327. https://doi.org/10.1007/s11023-021-09557-8

Morozov, E., & Bria, F. (2018). *Rethinking the Smart City: Democratizing Urban Technology*. Rosa Luxemburg Foundation. https://rosalux.nyc/wp-content/uploads/2021/02/RLS-NYC_smart_cities_EN.pdf

Moss, E., Watkins, E. A., Singh, R., Elish, M. C., & Metcalf, J. (2021). *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*. Data & Society. https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf

Murray, A. D. (2021). *Almost Human: Law and Human Agency in the Time of Artificial Intelligence - Sixth Annual T.M.C. Asser Lecture*. Annual T.M.C. Asser Lecture Series. (6). T.M.C. Asser Press. https://www.asser.nl/asserpress/books/?rId=13986

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180089. https://doi.org/10.1098/rsta.2018.0089

Nemitz, P. (2021). Democracy through law: The Transatlantic Reflection Group and its manifesto in defence of democracy and the rule of law in the age of "artificial intelligence". *European Law Journal, 29*(1-2), 237-248. https://doi.org/10.1111/eulj.12407

Nemitz, P., & Pfeffer, M. (2020). *Prinzip Mensch: Macht, Freiheit und Demokratie im Zeitalter der Künstlichen Intelligenz [The Human Imperative: Power, Freedom and Democracy in the Age of Artificial Intelligence]*. Dietz Verlag.

Newman, A. L. (2020). Digital Policy-Making in the European Union: Building the New Economy of an Information Society. In H. Wallace, M. A. Pollack, C. Roederer-Rynning, & A. R. Young (Eds.), *Policy-making in the European Union* (8th ed., pp. 275-296). Oxford University Press. https://doi.org/10.1093/hepl/9780198807605.003.0012

Niklas, J. (2020). Human Rights-Based Approach to AI and Algorithms: Concerning Welfare Technologies. In W. Barfield (Ed.), *The Cambridge Handbook of the Law of Algorithms* (pp. 517-542). Cambridge University Press. https://doi.org/10.1017/9781108680844.025

NJCM et al. v The Dutch State. (2020) Rechtbank Den Haag, C-09-550982 (English), 5 February 2020, ECLI: NL: RBDHA:2020:1878, https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:1878

Noorman, M., Keymolen, E., Schellekens, M., de Groot, A., de Conca, S., Leenes, R., Zhao, B., Dalla Corte, L., Bayamlioğlu, E., Taylor, L., Pierce, R., & Coenmans, R. (2019, January 31). *Response on the draft ethical guidelines for trustworthy AI produced by the European Commission's High-Level Expert Group on Artificial Intelligence*. https://www.tilburguniversity.edu/sites/default/files/download/NotesonHLEGDraftethicalguidelines_30012019_1.pdf

NOYB - European Center for Digital Rights. (2024). *GDPR: a culture of non-compliance? Numbers of evidence-based enforcement efforts*. https://noyb.eu/sites/default/files/2024-01/GDPR_a%20culture%20of%20non-compliance.pdf

Nussbaum, M. C. (2011). *Creating Capabilities: The Human Development Approach*. Harvard University Press.

Nyman-Metcalf, K., & Kerikmäe, T. (2020). Machines Are Taking Over - Are We Ready?: Law and Artificial Intelligence [Special Issue]. *Singapore Academy of Law Journal, 33*, 24-49. https://heinonline.org/HOL/P?h=hein.journals/saclj33&i=875

O'Connor, S., & Liu, H. (2023). Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & Society*. https://doi.org/10.1007/s00146-023-01675-4

OECD. (2021). *OECD Regulatory Policy Outlook 2021*. OECD Publishing. https://doi.org/10.1787/38b0fdb1-en

OECD. (2024). Recommendation of the Council on Artificial Intelligence, OECD/Legal/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Opinion of Advocate General Pikamäe delivered on 16 March 2023, SCHUFA Holding (Scoring), Case C-634/21, ECLI:EU:C:2023:220. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62021CC0634

Pałka, P., & Brożek, B. (2023). How not to get bored, or some thoughts on the methodology of law and technology. In B. Brożek, O. Kanevskaia, & P. Pałka (Eds.), *Research Handbook on Law and Technology* (pp. 82-98). Edward Elgar Publishing. https://doi.org/10.4337/9781803921327.00013

Papakonstantinou, V. (2022). Cybersecurity as praxis and as a state: The EU law path towards acknowledgement of a new right to cybersecurity? *Computer Law & Security Review, 44*, 105653. https://doi.org/10.1016/j.clsr.2022.105653

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Pohle, J., & Thiel, T. (2020). Digital sovereignty. *Internet Policy Review, 9*(4). https://doi.org/10.14763/2020.4.1532

Prabhakaran, V., Mitchell, M., Gebru, T., & Gabriel, I. (2022). A Human Rights-Based Approach to Responsible AI. *arXiv*. https://doi.org/10.48550/arXiv.2210.02667

Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM/2022/496 final, 28.9.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496

Proposal for a Directive of the European Parliament and of the Council on liability for defective products, COM/2022/495 final, 28.9.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0495

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, COM/2021/206 final, 21.4.2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Quintavalla, A., & Temperman, J. (Eds.). (2023). *Artificial Intelligence and Human Rights*. Oxford University Press. https://doi.org/10.1093/law/9780192882486.001.0001

Rachovitsa, A., & Johann, N. (2022). The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case. *Human Rights Law Review, 22*(2), 1-15. https://doi.org/10.1093/hrlr/ngac010

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5-14. https://doi.org/10.1007/s10676-017-9430-8

Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). *Artificial Intelligence & Human Rights: Opportunities & Risks*. Berkman Klein Center for Internet & Society Research Publication. http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439

Redden, J., Brand, J., Sander, I., & Warne, H. (2022). *Automating Public Services: Learning from Cancelled Systems*. Carnegie UK. https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2022/09/21101838/Automating-Public-Services-Learning-from-Cancelled-Systems-Final-Full-Report.pdf

Reed, C. (2018). How should we regulate artificial intelligence? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), 20170360. https://doi.org/10.1098/rsta.2017.0360

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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) (Text with EEA relevance), OJ L 119, 4.5.2016, p. 1-88, ELI: http://data.europa.eu/eli/reg/2016/679/oj

Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 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) (Text with EEA relevance), OJ L 151, 7.6.2019, p. 15-69, ELI: http://data.europa.eu/eli/reg/2019/881/oj

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) Text with EEA relevance, OJ L 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj

Rieder, G., Simon, J., & Wong, P.-H. (2021). Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums. In M. Pelillo & T. Scantamburlo (Eds.), *Machines We Trust: Perspectives on Dependable AI* (pp. 27-42). https://doi.org/10.7551/mitpress/12186.003.0007

Risse, M. (2021). The Fourth Generation of Human Rights: Epistemic Rights in Digital Lifeworlds. *Moral Philosophy and Politics, 8*(2), 351-378. https://doi.org/10.1515/mopp-2020-0039

Rodley, N. (2014). International Human Rights Law. In M. D. Evans (Ed.), *International Law* (4th ed., pp. 783-820). Oxford University Press.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., & Bengio, Y. (2023). Tackling Climate Change with Machine Learning. *ACM Computing Surveys, 55*(2), 1-96. https://doi.org/10.1145/3485128

Rotenberg, M. (2024). Human Rights Alignment: The Challenge Ahead for AI Lawmakers. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Eds.), *Introduction to Digital Humanism: A Textbook* (pp. 611-622). Springer Cham. https://doi.org/10.1007/978-3-031-45304-5_38

Royal Commission into the Robodebt Scheme. (2023, July 7). *Report of the Royal Commission into the Robodebt Scheme*. https://robodebt.royalcommission.gov.au/system/files/2023-09/rrc-accessible-full-report.PDF

Ruschemeier, H. (2023). AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal. *ERA Forum, 23*(3), 361-376. https://doi.org/10.1007/s12027-022-00725-6

Russell, S., Dewey, D., & Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine, 36*(4), 105-114. https://doi.org/10.1609/aimag.v36i4.2577

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson.

Santarius, T., Dencik, L., Diez, T., Ferreboeuf, H., Jankowski, P., Hankey, S., Hilbeck, A., Hilty, L. M., Höjer, M., Kleine, D., Lange, S., Pohl, J., Reisch, L., Ryghaug, M., Schwanen, T., & Staab, P. (2023). Digitalization and Sustainability: A Call for a Digital Green Deal. *Environmental Science & Policy, 147*, 11-14. https://doi.org/10.1016/j.envsci.2023.04.020

Sartor, G. (2020). Artificial intelligence and human rights: Between law and ethics. *Maastricht Journal of European and Comparative Law, 27*(6), 705-719. https://doi.org/10.1177/1023263x20981566

Savin, A. (2020). *EU internet law* (3rd ed.). Edward Elgar Publishing. https://doi.org/10.4337/9781789908572

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology, 29*(2), 353-400. https://heinonline.org/HOL/P?h=hein.journals/hjlt29&i=365

Schima, B. (2015). EU Fundamental Rights and Member State Action after Lisbon: Putting the ECJ's Case Law in Its Context European Union Law Issue. *Fordham International Law Journal, 38*(4), 1097-1134. https://heinonline.org/HOL/P?h=hein.journals/frdint38&i=1117

Schrems, M. (2014). *Kämpf um deine Daten [Fight for your Data]*. Edition a.

Schrepel, T. (2023). Law+ Technology. *The Journal of Law and Technology at Texas*, 1-19. https://jolttx.com/2023/01/15/law-technology/

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings on the Conference on Fairness, Accountability, and Transparency (FAT* ’19), Atlanta, USA, January 29-31*, 59-68. https://doi.org/10.1145/3287560.3287598

Sen, A. (2004). Elements of a Theory of Human Rights. *Philosophy & Public affairs, 32*(4), 315-356. https://doi.org/10.1111/j.1088-4963.2004.00017.x

Seubert, S., & Becker, C. (2021). The Democratic Impact of Strengthening European Fundamental Rights in the Digital Age: The Example of Privacy Protection. *German Law Journal, 22*(1), 31-44. https://doi.org/10.1017/glj.2020.101

Shackelford, S. J. (2019). Should Cybersecurity be a Human Right: Exploring the Shared Responsibility of Cyber Peace. *Stanford Journal of International Law, 55*(2), 155-184. https://heinonline.org/HOL/P?h=hein.journals/stanit55&i=171

Shahriari, K., & Shahriari, M. (2017). IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. *Proceedings of the 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), Toronto, Canada, July 21-22*, 197-201. https://doi.org/10.1109/IHTC.2017.8058187

Shelton, D. (2015). *Remedies in International Human Rights Law*. Oxford University Press. https://doi.org/10.1093/law/9780199588824.001.0001

Silveira, A. (2023). Automated individual decision-making and profiling [on case C-634/21 - SCHUFA (Scoring)]. *UNIO – EU Law Journal, 8*(2), 74-85. https://doi.org/10.21814/unio.8.2.4842

Smuha, N. (2019). The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence. *Computer Law Review International, 20*(4), 97-106. https://doi.org/10.9785/cri-2019-200402

Smuha, N. (2020). Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea. *Philosophy & Technology, 34*(S1), 91-104. https://doi.org/10.1007/s13347-020-00403-w

Smuha, N. (2021a). Beyond the individual: governing AI's societal harm. *Internet Policy Review, 10*(3). https://doi.org/10.14763/2021.3.1574

Smuha, N. (2021b). From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology, 13*(1), 57-84. https://doi.org/10.1080/17579961.2021.1898300

Smuha, N., & Morandini, A. (2022, October 11). Trustworthy AI through regulation? Sketching the European approach. *The Digital Constitutionalist*. https://digi-con.org/4-trustworthy-ai-through-regulation-sketching-the-european-approach/

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., Maclaren, J., Piselli, R., & Yeung, K. (2021, August 5). How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.3899991

Solove, D. J. (2024, February 1). Artificial Intelligence and Privacy. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.4713111

Taddeo, M. (2019). Three Ethical Challenges of Applications of Artificial Intelligence in Cybersecurity. *Minds and Machines, 29*(2), 187-191. https://doi.org/10.1007/s11023-019-09504-8

Taddeo, M., Jones, P., Abbas, R., Vogel, K., & Michael, K. (2023). Socio-Technical Ecosystem Considerations: An Emergent Research Agenda for AI in Cybersecurity. *IEEE Transactions on Technology and Society, 4*(2), 112-118. https://doi.org/10.1109/TTS.2023.3278908

Taddeo, M., McCutcheon, T., & Floridi, L. (2019). Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature Machine Intelligence, 1*(12), 557-560. https://doi.org/10.1038/s42256-019-0109-1

Tangi, L., van Noordt, C., Combetto, M., Gattwinkel, D., & Pignatelli, F. (2022). *AI Watch. European landscape on the use of Artificial Intelligence by the Public Sector*. Publications Office of the European Union. https://doi.org/10.2760/39336

Taylor, L. (2016). Data Subjects or Data Citizens?: Addressing the Global Regulatory Challenge of Big Data. In M. Hildebrandt & B. v. d. Berg (Eds.), *Information, Freedom and Property: The Philosophy of Law Meets the Philosophy of Technology* (pp. 81-106). Routledge. https://doi.org/10.4324/9781315618265

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society, 4*(2). https://doi.org/10.1177/2053951717736335

Taylor, L. (2021). Public Actors Without Public Values: Legitimacy, Domination and the Regulation of the Technology Sector. *Philosophy & Technology, 34*(4), 897-922. https://doi.org/10.1007/s13347-020-00441-4

Taylor, L. (2023a). Can AI governance be progressive? Group interests, group privacy and abnormal justice. In A. Zwitter & O. Gstrein (Eds.), *Handbook on the Politics and Governance of Big Data and Artificial Intelligence* (pp. 19-40). Edward Elgar Publishing. https://doi.org/10.4337/9781800887374.00011

Taylor, L. (2023b). In Search of Public Values in Private Systems: A Response to the Montesquieu Lecture by Karen Yeung. *Tilburg Law Review, 27*(2), 50-55. https://doi.org/10.5334/tilr.305

Taylor, L., Floridi, L., & van der Sloot, B. (2017a). Introduction: A New Perspective on Privacy. In L. Taylor, L. Floridi, & B. van der Sloot (Eds.), *Group Privacy: New Challenges of Data Technologies* (pp. 1-12). Springer Cham. https://doi.org/10.1007/978-3-319-46608-8_1

Taylor, L., & Mukiri-Smith, H. (2021). Human rights, technology and poverty. In M. F. Davis, M. Kjaerum, & A. Lyons (Eds.), *Research Handbook on Human Rights and Poverty* (pp. 535-549). Edward Elgar Publishing. https://doi.org/10.4337/9781788977517.00049

Taylor, L., van der Sloot, B., & Floridi, L. (2017b). Conclusion: What Do We Know About Group Privacy? In L. Taylor, L. Floridi, & B. van der Sloot (Eds.), *Group Privacy: New Challenges of Data Technologies* (pp. 225-237). Springer Cham. https://doi.org/10.1007/978-3-319-46608-8_12

The Global Partnership on Artificial Intelligence (GPAI). (2024). *About GPAI*. https://gpai.ai/about/

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2016). *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems*. IEEE. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v1.pdf

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition*. IEEE. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

The Public Voice. (2018, October 23). *Universal Guidelines for Artificial Intelligence*. https://thepublicvoice.org/ai-universal-guidelines/

The Republic of Korea & UK Government. (2024, May 21). *Seoul Declaration for Safe, Innovative and Inclusive AI by Participants Attending the Leaders´ Session of the AI Seoul Summit* [Policy Paper]. Ministry of Foreign Affairs, Ministry of Science and ICT. Foreign, Commonwealth & Development Office, Department for Science, Innovation and Technology. https://aiseoulsummit.kr/press/?uid=41&mod=document

The White House. (2022, October). *Blueprint for an AI Bill of Rights: Making Automated Systems Work For the American People* [White Paper]. The White House Office of Science and Technology Policy,. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447-464. https://doi.org/10.1007/s12525-020-00441-4

Transatlantic Reflection Group on Democracy and the Rule of Law in the Age of “Artificial Intelligence”. (2023). A Manifesto on Enforcing Law in the Age of ‘Artificial Intelligence’. *European Law Journal, 29*(1-2), 249-255. https://doi.org/10.1111/eulj.12474

Troitiño, D. R. (2022). The European Union Facing the 21st Century: The Digital Revolution. *TalTech Journal of European Studies, 12*(1), 60-78. https://doi.org/10.2478/bjes-2022-0003

UK Government. (2023, November 1). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023* [Policy Paper]. Prime Minister's Office, 10 Downing Street, Foreign, Commonwealth & Development Office, Department for Science, Innovation and Technology. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Ulbricht, L., & Yeung, K. (2022). Algorithmic regulation: A maturing concept for investigating regulation of and through algorithms. *Regulation & Governance, 16*(1), 3-22. https://doi.org/10.1111/rego.12437

Umbrello, S. (2021). Conceptualizing Policy in Value Sensitive Design: A Machine Ethics Approach. In S. J. Thompson (Ed.), *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence* (pp. 108-125). IGI Global. https://doi.org/10.4018/978-1-7998-4894-3.ch007

Umbrello, S. (2022). The Role of Engineers in Harmonising Human Values for AI Systems Design. *Journal of Responsible Technology, 10*, 100031. https://doi.org/10.1016/j.jrt.2022.100031

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics, 1*(3), 283-296. https://doi.org/10.1007/s43681-021-00038-3

United Nations Educational, Scientific and Cultural Organization (UNESCO) (2021). Recommendation on the Ethics of Artificial Intelligence. 23 November. SHS/BIO/PI/2021/1. https://unesdoc.unesco.org/ark:/48223/pf0000381137

United Nations, Advisory Body on Artificial Intelligence (2023). Governing AI for Humanity. United Nations publication. https://digitallibrary.un.org/record/4042280?ln=en&v=pdf

United Nations, General Assembly (2012). The promotion, protection and enjoyment of human rights on the Internet: resolution / adopted by the Human Rights Council. 29 June. A/HRC/RES/20/8. https://digitallibrary.un.org/record/731540?ln=en&v=pdf

United Nations, General Assembly (2024). Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development: draft resolution. 11 March. A/78/L.49. https://digitallibrary.un.org/record/4040897?v=pdf

United Nations, Human Rights Council (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David A. Kaye. 6 April. A/HRC/38/35. https://digitallibrary.un.org/record/1631686?v=pdf

United Nations, Human Rights Council (2019). Report of the Special Rapporteur on extreme poverty and human rights, Philip Alston. 11 October. A/74/493. https://digitallibrary.un.org/record/3834146?ln=en&v=pdf

United Nations, The World Conference on Human Rights in Vienna (1993). Vienna Declaration and Programme of Action: note / by the Secretariat. 25 June. A/CONF.157/23. https://digitallibrary.un.org/record/183139?ln=en&v=pdf

Université de Montréal. (2018). *Montréal Declaration for a Responsible Development of Artificial Intelligence*. https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl_IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf

van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security, 23*(4), 323-340. https://doi.org/10.1177/13882627211031257

van De Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines, 30*, 385-409. https://doi.org/10.1007/s11023-020-09537-4

van den Hoven, J. (2017). Ethics for the Digital Age: Where Are the Moral Specs? Value Sensitive Design and Responsible Innovation. In H. Werthner & F. v. Harmelen (Eds.), *Informatics in the Future: Proceedings of the 11th European Computer Science Summit (ECSS 2015), Vienna, October 2015* (pp. 65-76). Springer Cham. https://doi.org/10.1007/978-3-319-55735-9_6

van Ingen, E., & Bekkers, R. (2015). Generalized Trust Through Civic Engagement? Evidence from Five National Panel Studies. *Political Psychology, 36*(3), 277-294. https://doi.org/10.1111/pops.12105

van Kranenburg, R., Bohara, R., Yahalom, R., & Ross, M. (2023). Cyber Resilience, Societal Situational Awareness for SME. *2023 IEEE International Conference on Cyber Security and Resilience (CSR), Venice, July 31-August 2*. https://doi.org/10.1109/CSR57506.2023.10225011

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics, 1*, 213-218. https://doi.org/10.1007/s43681-021-00043-6

Vandezande, N. (2024). Cybersecurity in the EU: How the NIS2-directive stacks up against its predecessor. *Computer Law & Security Review, 52*, 105890. https://doi.org/10.1016/j.clsr.2023.105890

Vasak, K. (1977). Human Rights: A Thirty-Year Struggle: The Sustained Efforts to give Force of law to the Universal Declaration of Human Rights. *The UNESCO Courier, XXX*(11), 29-32. https://unesdoc.unesco.org/ark:/48223/pf0000048063

Verbeek, P.-P. (2005). *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Penn State University Press. http://www.jstor.org/stable/10.5325/j.ctv14gp4w7

Vetter, D., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Gallucci, A., Krendl Gilbert, T., Hagendorff, T., van Halem, I., Hickman, E., Hildt, E., Holm, S., Kararigas, G., Kringen, P., Madai, V. I., Wiinblad Mathez, E., Tithi, J. J., Westerlund, M., Wurth, R., V. Zicari, R., & Z-Inspection® Initiative. (2022). Lessons Learned from Assessing Trustworthy AI in Practice. *Digital Society, 2*(35). https://doi.org/10.1007/s44206-023-00063-1

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications, 11*(233). https://doi.org/10.1038/s41467-019-14108-y

Voll, C. N. (2023). *First signs of co-creation in engineering education*. https://backend.orbit.dtu.dk/ws/portalfiles/portal/345599784/First_signs_of_co-creation_in_engineering_education.png

Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. *Computers & Security, 38*, 97-102. https://doi.org/10.1016/j.cose.2013.04.004

Wachter, S., & Mittelstadt, B. (2019). A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review, 2019*(2), 494-620. https://doi.org/10.7916/cblr.v2019i2.3424

Wachter, S., Mittelstadt, B., & Floridi, L. (2017a). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76-99. https://doi.org/10.1093/idpl/ipx005

Wachter, S., Mittelstadt, B., & Russell, C. (2017b). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841-887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. *Computer Law & Security Review, 41*, 105567. https://doi.org/10.1016/j.clsr.2021.105567

Wagner, B. (2018). Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? In E. Bayamlioglu, I. Baraliuc, L. A. W. Janssens, & M. Hildebrandt (Eds.), *BEING PROFILED:COGITAS ERGO SUM: COGITAS ERGO SUM: 10 Years of Profiling the European Citizen* (pp. 84-89). Amsterdam University Press. https://doi.org/10.1515/9789048550180-016

Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet, 11*(1), 104-122. https://doi.org/10.1002/poi3.198

Waltl, B., & Vogl, R. (2018). Increasing Transparency in Algorithmic- Decision-Making with Explainable AI. *Datenschutz und Datensicherheit - DuD, 42*(10), 613-617. https://doi.org/10.1007/s11623-018-1011-4

Weber, M. (2010). *Politik als Beruf [Politics as a Vocation]* (11th ed.). Duncker & Humblot.

Wendehorst, C. (2020). Strict Liability for AI and other Emerging Technologies. *Journal of European Tort Law, 11*(2), 150-180. https://doi.org/10.1515/jetl-2020-0140

Whittaker, M. (2021). The steep cost of capture. *interactions, 28*(6), 50-55. https://doi.org/10.1145/3488666

Wieringa, M. (2023). "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case. *Data & Policy, 5*(e2). https://doi.org/10.1017/dap.2022.39

Wirtz, B. W., Langer, P. F., & Fenner, C. (2021). Artificial Intelligence in the Public Sector-a Research Agenda. *International Journal of Public Administration, 44*(13), 1103-1128. https://doi.org/10.1080/01900692.2021.1947319

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration, 43*(9), 818-829. https://doi.org/10.1080/01900692.2020.1749851

Wolswinkel, J. (2022). *Artificial Intelligence and Administrative Law: Comparative Study on Administrative Law and the Use of Artificial Intelligence and Other Algorithmic Systems in Administrative Decision-Making in the Member States of the Council of Europe*. (Report CDCJ(2022)31). The Council of Europe. https://coe.int/documents/22298481/0/CDCJ(2022)31E+-+FINAL+6.pdf/4cb20e4b-3da9-d4d4-2da0-65c11cd16116?t=1670943260563

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance, 12*(4), 505-523. https://doi.org/10.1111/rego.12158

Yeung, K. (2019). *Responsibility and AI: A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*. (Report DGI(2019)05). The Council of Europe. https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168096bdab

Yeung, K. (2022). The New Public Analytics as an Emerging Paradigm in Public Sector Administration. *Tilburg Law Review, 27*(2), 1-32. https://doi.org/10.5334/tilr.303

Yeung, K., & Bygrave, L. A. (2022). Demystifying the modernized European data protection regime: Cross-disciplinary insights from legal and regulatory governance scholarship. *Regulation & Governance, 16*(1), 137-155. https://doi.org/10.1111/rego.12401

Yeung, K., Howes, A., & Pogrebna, G. (2020). AI Governance by Human Rights–Centered Design, Deliberation, and Oversight: An End to Ethics Washing. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 76-106). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.5

Zech, H. (2021). Liability for AI: public policy considerations. *ERA Forum, 22*, 147-158. https://doi.org/10.1007/s12027-020-00648-0

Zech, H. (2023). How Should We Regulate AI? [Special Issue]. *Weizenbaum Journal of the Digital Society, 3*(3). https://doi.org/10.34669/WI.WJDS/3.3.7

Zicari, R. V., Ahmed, S., Amann, J., Braun, S. A., Brodersen, J., Bruneault, F., Brusseau, J., Campano, E., Coffee, M., Dengel, A., Düdder, B., Gallucci, A., Gilbert, T. K., Gottfrois, P., Goffi, E., Haase, C. B., Hagendorff, T., Hickman, E., Hildt, E., Holm, S., Kringen, P., Kühne, U., Lucieri, A., Madai, V. I., Moreno-Sánchez, P. A., Medlicott, O., Ozols, M., Schnebel, E., Spezzatti, A., Tithi, J. J., Umbrello, S., Vetter, D., Volland, H., Westerlund, M., & Wurth, R. (2021a). Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier. *Frontiers in Human Dynamics, 3*, 688152. https://doi.org/10.3389/fhumd.2021.688152

Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., Kararigas, G., Kringen, P., McCullough, M., Möslein, F., Mushtaq, N., Roig, G., Stürtz, N., Tolle, K., Tithi, J. J., van Halem, I., & Westerlund, M. (2021b). Z-Inspection®: A Process to Assess Trustworthy AI. *IEEE Transactions on Technology and Society, 2*(2), 83-97. https://doi.org/10.1109/TTS.2021.3066209

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Zuiderveen Borgesius, F. J. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights, 24*(10), 1572-1593. https://doi.org/10.1080/13642987.2020.1743976

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly, 38*(3), 101577. https://doi.org/10.1016/j.giq.2021.101577

Acknowledgements

I would like to thank my supervisors, Professor Dr Tanel Kerikmäe & Associate Professor Dr Thomas Hoffmann, and our wonderful colleagues at the department for their encouragement and support throughout the academic journey. I also wish to thank my students, including our Moot Court participants who represented us abroad at several international competitions. Many thanks are also due to the Ragnar Nurkse Department doctoral school seminar participants and the reviewers for their instructive feedback on earlier drafts. Last but not least, I wish to thank our PhD students in the department for learning from and with them.

This PhD project benefited from funding provided by the European Regional Development Fund (2014-2020.4.01.16-0032, ASTRA "TTÜ arenguprogramm aastateks 2016-2022"), the State Chancellery (Lep18096, "Elaborating legal Solutions related to autonomous intelligent technologies", 15.12.2018−30.04.2019), and the European Commission (VERT22032, "Legal and Regulatory Governance of Artificial Intelligence Systems", 1.09.2022−31.08.2025).

Abstract

Human Rights-Based Legal Conditions for Trustworthy AI

Regulating the advanced networked digital information and communications technology of artificial intelligence (AI) system has recently become a key priority for governments around the world. While on the one hand, AI promises to contribute to economic growth, help tackle societal challenges, and render the delivery of public services more efficient and effective, on the other hand, empirical evidence, such as the *SyRI* case, has shown how some mundane AI implementations in the public sector can create significant material and immaterial harm to individuals and groups impacting the EU values of democracy, the rule of law and human rights, including the trust of citizens. In response to the complex twofold challenge of mitigating potential human rights harm while leveraging the projected societal benefits of AI, governments are establishing regulatory instruments focused on enabling the implementation of *Trustworthy AI*. Informed by the fundamental rights-based approach of the High-Level Expert Group on Artificial Intelligence (AI HLEG) and its three-part framework for Trustworthy AI as *lawful*, *ethical,* and *robust AI*, this cumulative thesis conceives the realisation of Trustworthy AI as adequate protection and promotion of the fundamental rights of individuals and groups concerning the development and public-sector use of AI. Yet, only limited research has dealt with the question of the legal conditions under which Trustworthy AI can be realised in the public sector.

Providing conceptually and theoretically informed insights on this topic, this interdisciplinary thesis examines one main research question and three sub-questions:
Under what legal conditions can the fundamental rights of individuals and groups be protected and promoted regarding the development and public-sector use of AI systems?

- What policy and legal measures has the European Union implemented as part of its AI strategy to protect and promote fundamental rights?
- To what extent can the main legal principles and mechanisms under the GDPR address risks to the fundamental right of the protection of privacy concerning the development and use of AI?
- In which aspects does the AI Act fall short in guaranteeing adequate protection and promotion of fundamental rights regarding the development and public-sector use of AI?

The main aim of the thesis is to determine the legal conditions necessary and sufficient for protecting and promoting the fundamental rights of individuals and groups concerning the development and public-sector use of AI. In addressing the research questions, the thesis looks into two cases, involving the application of two different AI models in the public sector: the *SyRI* case, and the implementation of smart metering systems (SM), including the use of non-intrusive load monitoring techniques, in residential households. It focuses on regulating narrow AI systems and neither deals with AI that poses only minimal risk, such as spam filters, nor with AI applications for military purposes.

Based on four original, peer-reviewed publications and a contribution to the Report of Estonia's AI Taskforce, this thesis contributes to the growing body of knowledge on AI regulation, particularly to nascent critical streams in the literature on human rights-based approaches to AI. By combining insights from human rights and science and technology studies, it examines pertinent EU legal frameworks under the EU AI strategy, including

the Cybersecurity Act, the NIS 1, and the revised NIS 2 Directive, the GDPR, and the AI Act. Methodologically, the thesis adopts a “law in action” approach, taking an external viewpoint on the law and including critical perspectives beyond the traditional legal approach. By applying qualitative research using both deductive and inductive methods, it thereby aims to understand how the four ethical principles of the AI HLEG can be translated into *lawful AI*. In this regard, the thesis follows Cohen’s conceptualisation of fundamental rights as *civil liberties, capabilities,* and *affordances*, considering that developers significantly influence the ordering function of the law by embedding their values into the technology in the development of AI, thereby impacting the choices and autonomy of citizens at the use stage of the technology.

Building on findings in Publication IV, to which the author contributed with desk research on AI policies of several European countries, the USA, and the EU, Publication I establishes the main technical approaches of the technology and evaluates potential regulatory models to address normative concerns related to AI. It suggests that the regulatory discussions on AI should be centred on the question of by whom and for which purposes AI systems will be designed and related to that, by whom the technology is owned and deployed and in which contexts AI will be applied. Publications II and III show how interdisciplinary, fundamental rights-based *systems thinking* approaches can be applied to mitigate potential human rights harms and promote the fundamental rights of citizens as end-users of AI. Publication III particularly emphasises the introduction of *participatory* systems thinking approaches, based on the structured participatory design method of value sensitive design and the co-assessment inspection process for Ethical AI. The thesis treats these approaches as initial methods for organising *participatory AI governance* in the public sector, aimed at involving legislators, policymakers, developers, deployers, and citizens in the joint development and assessment of the technology. The findings of Publication V, particularly on the harms caused by the ransomware attack *WannaCry*, informed Publication II, which contains technical and legal analyses of identified privacy and cybersecurity concerns of end-users concerning the implementation of SM in residential households.

In synthesis, the normative analysis highlights some significant limitations in extant EU legislative frameworks and the AI Act for realising Trustworthy AI in the public sector. In an attempt to overcome them, the thesis suggests implementing *human rights-based* legal conditions as part of an inherently interdisciplinary, collaborative, bottom-up approach to AI regulation. Overall, the conditions proposed involve implementing *fundamental rights impact assessments* and *effective redress mechanisms,* allocating *adequate funding for interdisciplinary education programs* and *national supervisory authorities*, and establishing *fora in existing public institutions for participatory design* and *the inspection process for Ethical AI*. By conceptualising *AI as a socio-technical system* embedded in societal structures and contexts, mediated through digital devices, the thesis highlights the need to treat AI not as a neutral tool or product but as a technology that requires advancing AI regulation as a *collective responsibility*. Complementary to the existing risk-based approach in the AI Act, the thesis provides guidelines and suggests practically applicable tools for practitioners and scholars to test and implement participatory regulatory approaches to AI in the public sector to realise Trustworthy AI in the long term.

Lühikokkuvõte

Inimõigustel põhinevad õiguslikud tingimused usaldusväärse tehisintellekti loomisel

Tehisintellekti süsteemidel põhinev võrgustunud digitaalse info- ja kommunikatsioonitehnoloogia reguleerimine on viimasel ajal muutunud kogu maailma valitsuste jaoks oluliseks prioriteediks. Ühelt poolt lubatakse, et tehisintellektiga aidatakse kaasa majanduskasvule, ühiskondlike probleemide lahendamisele ning avalike teenuste osutamise tõhusamaks ja tulemuslikumaks muutmisele, kuid teiselt poolt on empiirilised tõendid, näiteks SyRI juhtum, näidanud, kuidas mõned argised avalikus sektori tehisintellektirakendused võivad tekitada üksikisikutele ja rühmadele märkimisväärset materiaalset ja immateriaalset kahju, mõjutades demokraatia, õigusriigi ja inimõiguste väärtusi, sealhulgas kodanike usaldust. Vastuseks keerulisele kahetasandilisele väljakutsele, milleks on võimalike inimõiguste kahjustuste leevendamine ja samal ajal tehisintellekti prognoositava ühiskondliku kasu ärakasutamine, on valitsused kehtestanud regulatiivsed meetmed, mis on suunatud usaldusväärse tehisintellekti rakendamise võimaldamisele. Tuginedes kõrgetasemelise tehisintellekti eksperdirühma (AIHLEG) põhiõigustel põhinevale lähenemisviisile ja selle kolmest osast koosnevale usaldusväärse tehisintellekti raamistikule – seaduslik, eetiline ja tugev tehisintellekt – on käesolevas kumulatiivses doktoritöös käsitletud usaldusväärse tehisintellekti realiseerimist kui üksikisikute ja rühmade põhiõiguste piisavat kaitset ja edendamist seoses tehisintellekti arendamise ja avaliku sektori kasutamisega. Siiski on seni vaid piiratud määral uuritud, millistel õiguslikel tingimustel saab usaldusväärset tehisintellekti avalikus sektoris realiseerida.

Käesolevas interdistsiplinaarses doktoritöös uuritakse ühte peamist uurimisküsimust ja kolme alamküsimust, mis annavad kontseptuaalselt ja teoreetiliselt põhjendatud ülevaate sellest teemast:

Millistel õiguslikel tingimustel saab üksikisikute ja rühmade põhiõigusi kaitsta ja edendada seoses tehisintellekti süsteemide arendamise ja avaliku sektori kasutamisega?

- Milliseid poliitilisi ja õiguslikke meetmeid on Euroopa Liit võtnud osana oma tehisintellekti strateegiast põhiõiguste kaitsmiseks ja edendamiseks?
- Mil määral saavad üldises isikuandmete kaitse määruses sätestatud peamised õiguslikud põhimõtted ja mehhanismid käsitleda riske, mis ähvardavad eraelu puutumatuse kaitse põhiõigust seoses tehisintellekti arendamise ja kasutamisega?
- Millistes aspektides ei suuda tehisintellekti seadus tagada põhiõiguste piisavat kaitset ja edendamist seoses tehisintellekti arendamise ja avaliku sektori kasutamisega?

Lõputöö peamine eesmärk on määrata kindlaks õiguslikud tingimused, mis on vajalikud ja piisavad üksikisikute ja rühmade põhiõiguste kaitsmiseks ja edendamiseks seoses tehisintellekti arendamise ja avaliku sektori kasutamisega. Uurimisküsimuste käsitlemisel vaadeldakse doktoritöös kahte juhtumit, mis hõlmavad kahe erineva tehisintellekti mudeli rakendamist avalikus sektoris: SyRI juhtum ja nutikate mõõtmissüsteemide rakendamine, sealhulgas mittesekkuvate koormuse jälgimise meetodite kasutamine kodumajapidamistes. Peamiselt keskendutakse kitsaste tehisintellekti süsteemide reguleerimisele ning ei käsitleta tehisintellekti, mis kujutab endast vaid minimaalset ohtu, nagu rämpsposti filtrid, ega tehisintellekti rakendusi sõjalistel eesmärkidel.

Käesolev doktoritöö, mis põhineb neljal eelretsenseeritud teadusartiklil ja Eesti tehisintellekti töörühma aruandel, annab oma panuse tehisintellekti reguleerimist käsitlevate teadmiste kasvavasse kogumisse, eelkõige tekkivatesse kriitilistesse suundadesse kirjanduses, mis käsitlevad inimõigustel põhinevaid lähenemisviise tehisintellektile. Ühendades inimõiguste ning teadus- ja tehnoloogiauuringute seisukohti, uuritakse selles ELi tehisintellekti strateegia raames asjakohaseid ELi õigusraamistikke, sealhulgas küberturvalisuse seadust, NIS 1 ja läbivaadatud NIS 2 direktiivi, üldist majandushuvi käsitlevat määrust ja tehisintellekti seadust (tehisintellekti seadus). Metodoloogiliselt lähtutakse doktoritöös õigus pariselus – law in action –, võttes õiguse suhtes välise vaatenurga ja kaasates kriitilisi vaatenurki, mis väljuvad traditsioonilisest õiguslikust lähenemisviisist. Kvalitatiivne uurimus kasutab nii deduktiivset kui ka induktiivset meetodit, mille abil püütakse mõista, kuidas AIHLEGi neli eetilist põhimõtet on võimalik tõlkida seaduslikuks tehisintellektiks. Sellega seoses järgib doktoritöö Coheni põhiõiguste kontseptsiooni kui kodanikuvabaduste, võimekuste ja võimaldatavus (affordances) kontseptsiooni, võttes arvesse, et arendajad mõjutavad oluliselt õiguse korralduslikku funktsiooni, integreerides oma väärtused AI arendamisel tehnoloogiasse, mõjutades seeläbi kodanike valikuid ja autonoomiat tehnoloogia kasutamisetapis.

Tuginedes IV artikli järeldustele, millesse põimiti mitmete Euroopa riikide, Ameerika Ühendriikide ja ELi tehisintellektipoliitikat käsitlev dokumendianalüüs, kehtestatakse I artiklis tehnoloogia peamised tehnilised lähenemisviisid ja hinnatakse võimalikke regulatiivseid mudeleid, et käsitleda tehisintellektiga seotud normatiivseid probleeme. Selles tehakse ettepanek, et tehisintellekti üle peetavad regulatiivsed arutelud peaksid keskenduma küsimusele, kelle poolt ja millistel eesmärkidel tehisintellekti süsteemid kavandatakse ja sellega seoses, kellele tehnoloogia kuulub ja kes seda kasutab ning millistes kontekstides tehisintellekti rakendatakse. Usaldusväärse tehisintellekti realiseerimiseks näitavad II ja III artiklis esitatud järeldused, kuidas saab rakendada interdistsiplinaarseid, põhiõigustel põhinevaid süsteemseid lähenemisviise, et leevendada võimalikku kahju inimõigustele ja edendada kodanike põhiõigusi. III väljaandes rõhutatakse eelkõige osalusel põhineva süsteemse mõtlemise lähenemisviiside kasutuselevõttu, mis põhineb struktureeritud osalusel põhineval väärtustundliku projekteerimise meetodil ja eetilise tehisintellekti kaashindamise kontrolliprotsessil. Doktoritöös käsitletakse neid lähenemisviise kui esialgseid meetodeid osaluslike tehisintellekti valitsemise korraldamiseks avalikus sektoris, mille eesmärk on kaasata seadusandjad, poliitikakujundajad, arendajad, kasutuselevõtjad ja kodanikud tehnoloogia ühisarendusse ja -hindamisse. V artikli järeldused, eelkõige *WannaCry* lunavararünnaku põhjustatud kahjude kohta, on aluseks II artiklis, milles viiakse läbi tehniline ja õiguslik analüüs lõppkasutajate tuvastatud eraelu puutumatuse ja küberturvalisuse probleemide kohta seoses arukate mõõtmissüsteemide rakendamisega kodumajapidamistes.

Kokkuvõttes toob normatiivne analüüs esile mõned olulised piirangud olemasolevates ELi õigusraamistikes ja tehisintellekti seaduses, mis on vajalikud usaldusväärse tehisintellekti rakendamiseks avalikus sektoris. Nende probleemide ületamiseks tehakse ettepanek rakendada inimõigustel põhinevaid õiguslikke tingimusi osana interdistsiplinaarsest, koostööl põhinevast, alt-üles lähenemisviisist tehisintellekti reguleerimisel. Üldiselt hõlmavad väljapakutud tingimused järgmist: põhiõiguste mõju hindamise ja tõhusate õiguskaitsemehhanismide rakendamine, piisava rahastamise eraldamine interdistsiplinaarsetele haridusprogrammidele ja riiklikele järelevalveasutustele ning foorumite loomine olemasolevates avalikes asutustes osaluslike kujunduste ja eetilise tehisintellekti kontrolliprotsessi jaoks. Mõtestades tehisintellekti kui sotsiaal-tehnilist süsteemi, mis on integreeritud ühiskondlikesse struktuuridesse ja kontekstidesse ning mida vahendavad digitaalsed seadmed, rõhutatakse doktoritöös vajadust käsitleda tehisintellekti mitte neutraalse vahendi või tootena, vaid hoopis tehnoloogiana, mis nõuab tehisintellekti reguleerimise edendamist kui kollektiivset vastutust. Doktoritöö täiendab tehisintellekti seaduses sätestatud riskipõhist lähenemisviisi ning pakub praktikutele ja teadlastele suuniseid ja praktiliselt rakendatavaid vahendeid, et katsetada ja rakendada avalikus sektoris tehisintellekti osalusel põhinevaid regulatiivseid lähenemisviise, et saavutada pikas perspektiivis usaldusväärne tehisintellekt.

Appendix: Publications I-III and IV-V

Publication I

3.1 Antonov, A., & Kerikmäe, T. (2020). Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU. In D. R. Troitiño, T. Kerikmäe, R. M. de la Guardia & G. Á. P. Sánchez (Eds.), *The EU in the 21st Century: Challenges and Opportunities for the European Integration Process* (pp. 135–154). Springer. https://doi.org/10.1007/978-3-030-38399-2_9

Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU

Alexander Antonov [1]

Email alanto@taltech.ee

Tanel Kerikmäe [1]

[1] Tallinn University of Technology, Ehitajate Tee 5, 19086, Tallinn, Estonia

Abstract

Artificial intelligence has become a frequent subject of discussion at different international forums in recent years. Classified by the European Union as one of the “most strategic technologies in the 21st century”, in 2018 the EU Commission mandated a 52-member strong High-Level Expert Group on AI to discuss the ethical, legal, economic and social impact of this promising technology. Providing a holistic view on the main ethical and legal questions surrounding AI-powered systems, this chapter intends to explore the latest EU initiatives on AI governance with a view to identifying the main challenges ahead for the EU to become a “leader in cutting-edge AI that can be trusted throughout the world”.

Keywords

European Union
Artificial intelligence
High-Level Expert Group on AI
Trustworthy AI
AI race

This chapter makes reference to three reports on autonomous intelligent technologies to which the authors had contributed as part of the project “Lep18096 ‘Elaborating legal Solutions related to autonomous intelligent technologies (15.12.2018–30.04.2019)’”, coordinated by Prof. Kerikmäe and financed by the Estonian government (EV Riigikantselei). Special thanks are due to Prof. Hoffmann (TalTech University, Department of Law) for his helpful comments on the manuscript.

The way we approach AI will define the world we live in.
European Commission—AI Strategy—April 2018.

1. Introduction

Having stood at a crossroads for years, in April 2018 the Juncker Commission finally devised a European strategy on Artificial intelligence (hereafter AI) paving its own, European way, guided by a coordinated, proportionate and human-centric approach towards a future regulatory regime for "trustworthy AI".[1] The EU Commission put forward incentives that allow the 27 post-Brexit Member States to join forces to govern the development of "trustworthy AI".

In the context of a global competition on AI, which is also often labelled an "AI race"[2] among the three AI global players,[3] the USA, China and the EU, the EU strategy aims to preserve the fundamental rights of its 500 million citizens in the digital era while giving its around 5000 top AI researchers,[4] 800 AI companies[5] and European public institutions enough leeway to leverage the potential of AI-powered systems.[6]

Likened to the invention of the steam engine or electricity, AI has been characterised by the EU as one of the "most strategic technologies in the 21st century".[7] It is not without reason why the latest OECD figures suggest a rise in global AI equity investment in the European Union. Over a period of 6 years, the share has increased from 1% in 2013 to 8% in 2017[8] lending credence to a study conducted by McKinsey Global Institute, which estimated that AI applications could potentially generate an additional global GDP growth of around 1.2 per cent annually until 2030.[9] Discussions on the ethical and social implications of AI have dominated the agenda of different international forums in recent years, ranging from the OECD[10] to G7 and G20 Summits.[11]

Having chosen the "ethical by design" approach, which implies that AI-powered products have to be regulated already at the early stage of their development,[12] European policymakers have been reflecting on the fundamental question as to how this technology can be "designed to augment, complement and empower human cognitive, social and cultural skills" rather than "unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans".[13] AI is therefore not only considered a general purpose but also a disruptive technology.[14]

While there is no internationally agreed definition of AI, the European Commission's Communication on AI refers to it as "systems that display intelligent behaviour by analysing their environment and taking actions—with some degree of autonomy—to achieve specific goals".[15]

1.1. Methodology: Purpose and Scope

Predicated on the assumption that the EU's competitiveness as regards the development of AI-powered systems by the private sector and research talent by the public sector depends on the unity among the post-Brexit 27 Member States, this chapter intends to explore the latest EU initiatives on AI governance.

Taking a holistic view on the ethical risks and legal questions related to AI-powered systems, the authors consult pertinent EU Communications, EU regulations and directives, specifically the GDRP, and the latest reports by the Joint Research Centre (JRC) of the European Commission on Artificial Intelligence (2018) and by the High-Level Expert Group on Artificial Intelligence (hereafter AI HLEG) (2019).

Since AI can be applied both for civil and military purposes, it is also called a dual-use technology. Lethal autonomous weapon systems would fall within the latter category. This chapter deals solely with the former.

Most estimates, e.g., on investments into AI research or on the availability of venture capital are taken directly from studies mainly conducted by McKinsey Global Institute and the Centre for Data Innovation.

While the figures should be taken with caution as most statistics on AI remain weak or speculative in nature,[16] we consult them mainly for two reasons. Whereas the EU Commission also referred to figures published by McKinsey Global Institute, the methodology of the latest report by the Centre for Data Innovation, a non-partisan research institute, affiliated with the Information Technology and Innovation Foundation proved to be most evidence-based among the studies conducted on AI. The comparative analysis, which relies upon, *i.e.,* statistical data used by the OECD, or crunch base, applies both quantitative and qualitative methods and analyses the performance of China, the European Union and the USA looking on the dimensions of AI "talent, research, enterprise development, adoption, data, and hardware" and employing absolute metrics.[17]

Additional inspiration on the legal debate on AI regulation was drawn from expert opinions of (legal) scholars mainly from Western Europe such as Easterbrook (1995), Lessig (1999; 2006), Buiten (2018), Franke (2019) and Floridi (2019).

Prior to mapping the EU's approach to trustworthy AI, it is worthwhile looking at the features and the associated ethical risks ensuring from employing this technology, which no matter of the amount of attention devoted by the media to AI remains to be intangible. Interestingly enough, the most acclaimed AI researchers have even struggled to define intelligence.[18]

2. What Is AI?

2.1. From Desktop Computers to AI-Powered Systems

The pace of technological changes and breakthroughs over the recent decades is unparalleled in history, and we are only witnessing the beginning of this transformative process longing from desktop computers to 3D printers and from voice-controlled Internet of Things devices to AI-powered machines. The progress in the development of Information and Communication Technologies (hereafter ICTs) seems to be unlimited with influential AI researchers already pondering the bleak scenario that we might see super-intelligent machines overtaking us in all dimensions of human intelligence in the future, alluding to the concept of artificial general intelligence (hereafter AGI).[19]

Computers (based on input, storage, processing and output of data) have become indispensable to our contemporary way of life, to how policymakers and managers organise economies and societies and make predictions of the future. They have transformed the global economic landscape and brought societies closer together spurring the development of previously inconceivable products and making information accessible to billions of people via the Internet.[20]

With the advent of a rapid increase in computational power and data, catalysed by the commercialisation of the Internet, one specific field in the application of computer programmes has recently grown to prominence: AI. The generic term refers to an emerging research field and technology, which allows for machines to replicate human-like reasoning, the ability to learn and solve complex tasks.[21]

Bearing great potentials reaching from making cost-effective predictions to detecting cancer cells, AI-powered systems have experienced an increasing uptake by different sectors.[22] The following non-exhaustive list provides an overview of the sectors, in which AI-powered systems have already been deployed:

– Transport sector

– Agriculture

– Financial sector

– Marketing and advertising

- Science
- Health care
- Criminal justice/law enforcement
- Securing the public sector
- Augmented and virtual reality.[23]

2.2. Birth of the Discipline of AI

It is a widely held view that the computer scientist, John McCarthy, coined the generic term AI in the 1950s, defining the discipline as a "science and engineering of making intelligent machines".[24] In summer 1956, he invited a selected group of leading researchers from the same field, there under Marvin Minsky and Ray Solomonoff, to Dartmouth, New Hampshire, to test the assumption of whether "every aspect of learning or any feature of intelligence can in principle be so precisely described that a machine can be made to simulate it".[25] What was then called the "Dartmouth Summer Research Project on Artificial Intelligence" is today considered the birth of AI as an academic discipline.[26]

From then on, the field of AI has gone through various cycles of successes ("AI Summer") and failures ("AI Winter").[27] A beneficial combination of a recent breakthrough in computation power and a rapidly growing data pool has catalysed an AI "renaissance".[28] Specifically a sub-field of AI, "machine-learning (hereafter ML)", has given the discipline new impetus.[29] This approach is grounded on the technical ability of a programmed machine to learn optimising its own processes and improving its own decisions (output) mainly based on large data sets (input).[30]

2.3. The Promising Sub-field of AI—Main Approaches to Machine Learning

ML revolutionised computing. Whereas classical programming is deductive in nature, implying that a machine processes data following a pre-defined algorithmic logic to answer (a) question/s posed to it, ML reverses this logic by creating its own rules mainly based on a vast amount of data and pre-given answers to the machine.[31]

The cognitive abilities of a human being, captured by reasoning and learning, are replicated by the machine either through *supervised learning* (e.g. used for picture recognition: feeding the machine with thousands of photographs so that it can later on identify persons and objects based on other pictures it has never seen before), *unsupervised learning* (e.g. applied for marketing purposes: AI-powered algorithms target customers with tailor-made advertisements after the computer categorised large data sets) and *reinforcement learning* (e.g. AlphaGo: giving the machine for each decision a signal that it made either a good or bad decision, independent of the size of the data it was tasked with processing).[32]

The most successful among the various methods associated with machine learning relates to *deep learning*, which, put in simplified terms, is an approach that is built upon different hidden layers between the input and output data. Deep learning systems replicate a neural network, which allows the computer, for example, to recognise in several mathematical steps an image or a voice. The algorithmic logic is hierarchical, which also implies that deep learning renders it possible for the computer to learn complex concepts such as pictures by creating a simpler version of them in successive steps.[33]

These rule-based systems are supposed to decrease the error rate or, e.g., in terms of image or facial recognition its *bias*. Nevertheless, independent of the amount of data, a certain error rate always remains.[34] Potential flaws in data can have serious consequences for the end consumer of a product, especially in automated decision-makings. Achieving a high degree of accuracy should therefore take centre stage in AI research.[35]

Even though it has been demonstrated that ML techniques can outperform human beings in certain tasks such as pattern recognition or games as chess or Go, as of today, these systems fall short of the promises enunciated by the Dartmouth participants. Neither do they embody a "common sense reasoning, {nor} self-awareness {nor} the ability {…} to define its own purpose".[36] It is for this reason, why the current systems fall within the category of so-called narrow/weak AI.[37] In other words, the algorithms for now are unable to think out of the box.[38] Despite its great potential, AI is neither magic nor a panacea.[39]

3. Ethical and Legal Aspects of AI

3.1. Ethical Risks of AI

It is a shared view among scientists and policymakers that those taking the lead in developing AI will not only harness the technology but also influence its further ethical trajectory.[40] The director of OpenAI, Jack Clark, aptly set the context of the great challenge ahead for policymakers to keep pace in developing and regulating the use of AI:

> AI let us encode values into systems that have been scaled against sometimes entire populations. If we fail here {to lead}, then the values that our society lives under are partially determined by whichever society wins in AI so that the values of that society encodes become the values of what we experience. So I think the stakes here are societal in nature and we should not think about this as a technological challenge but as how we as a society want to become better and the success here will be the ability to articulate values for the rest of the world.[41]

While the use of AI-powered systems comes along with various social and economic opportunities, examples of the abuse of this technology have demonstrated the need to evaluate the possible ethical, legal and technical challenges ensuing from deploying a technology that is conceived as a "black-box".[42] In this vein, the future regulation of AI-powered systems is compounded by the fact that even its designers or programmers struggle to fully understand the decision-making process of this opaque technology.[43]

Apart from projections about large-scale job losses from automation,[44] cases of documented bias,[45] nudging,[46] cyber-security risks[47] or deep-fakes,[48] large-scale applications of facial recognition technologies by law enforcement units,[49] some of which even intentionally deployed to identify individuals from a minority group,[50] undergird the argument that AI is a double-edge sword-bearing great economic potential but if not ethically governed can cause grave if not irreversible harm to the rights of human beings subjected to a potentially abusive use of this technology. Additionally, the fact that AI-powered systems can create legally binding decisions based on automated processing and profiling of a vast amount of personal data makes it a compelling case for a future regulatory regime.[51]

A unique approach to AI governance illustrating this ethical dilemma has been taken by Estonia, which created a special metaphor for AI, the so-called Kratt.[52] Drawing on an old Estonian mythological creature, Kratt symbolises AI, a technology that "is devoted to serving its master, but can become bad if left idle".[53]

3.2. Scholarly Debate and Regulatory Fields of AI

The commercialisation of the ever more converging ICTs has required policymakers to rethink traditional approaches to governance and regulations. While policymakers intend to limit the unpredictable effects of new technologies by creating regulatory frameworks that allow different actors, be it private citizens, public officials or companies to thrive on and harness advanced technological systems, the intangible nature of ICTs complicates the lawmaking process.

With the advent of the introduction of the Internet, a debate on the regulation of cyberspace between two Western legal scholars from the USA provided the starting point for a discussion on the question of how to regulate ICTs or in the case of this chapter, AI-powered systems. It is also illustrative for the current discussions at the EU level as which approach might be the most applicable one in regulating AI-powered systems.[54]

Judge Easterbrook (1996) compared the regulation of algorithm-based computing systems with legal problems related to horses making the argument that cyberspace ought to be studied through the lenses of traditional fields of law, be it administrative law, criminal law or property law.[55] More importantly, before creating new legislation for cyberspace, *de lege ferenda*, he advocated improving and clarifying *de lege lata*, the law as it stands now.[56]

Lessig (1999) took issue with this view, contending that there was "something when we think about the regulation of cyberspace that other areas would not show us", alluding to the unique architecture of cyberspace.[57] Contrary to the physical space, the novelty of cyberspace was due to its malleable code, which is written by a plethora of programmers.[58] In this regard, the changeable nature of code makes it impossible to apply the law of the horse to cyberspace.[59]

He therefore proposed a four-dimensional model, the "Four Modalities of Regulation in Real Space and Cyberspace", (1) law, (2) social norms, (3) market and (4) architecture, and suggested regulators to study the novel characteristics of cyberspace by means of his multidimensional model, the "net regulation".[60]

In other words, while Easterbrook believed that the technological evolution would not necessitate a new regulatory field, captured by the term cyberlaw, Lessig refuted this argumentation positing that given the changing nature of the architecture of code, legal norms alone could not help achieve the goal of creating legal values that guide a society through times of rapid technological changes.[61]

Another approach to regulating AI-powered systems can be drawn from the value-sensitive design framework espoused by, e.g., Van den Hoven (2017). In line with the view that the architecture of advanced technologies beard considerable impacts on the values of the society by whom the technology is employed, van den Hoven's approach would suggest that policymakers should create a regulatory framework that allowed ethical issues to be taken into account at the design process of the technology, “when value considerations can still make a difference”.[62]

Hence, the main focus of regulations should be placed on the work of the programmer or code writer. The conclusion is especially valid when considering Lessig’s assumption that “code {was} law”.[63] Applying this logic to the regulation of AI-powered systems would imply that policymakers are well advised to “open the black-box of AI”[64] and identify the conceivable challenges this technology poses to the society. As argued by Buiten (2018), this goal could be best attained if the causes of the risks were addressed and regulatory mechanisms were established with a view to increasing the societal capacity to absorb or even control those identified risks.[65] Even prior to that, the point of departure in regulating AI-powered systems should be questioning *by whom* and *for which purpose* this technology will be designed and related to that *by whom* they are owned and deployed and *in which contexts* they will be applied.[66]

Nevertheless, rather than “inventing the wheel” as regards future AI regulations, it is helpful to map the current applicable regulatory framework for ICTs (e.g. Product Liability Directive: Council Directive 85/374/EEC; Machinery Directive 2006/42/EC) at the European level and avoid national solo-efforts or silo thinking.[67] Legal initiatives on AI should therefore be coordinated by the EU Commission. The “Coordinated plan on Artificial Intelligence” which *i.a.* requests EU Member States to draft their own national AI strategies by mid-2019 constitutes a good starting point for a comprehensive and structured approach towards a future EU regulatory regime for trustworthy AI-powered systems.[68]

Apart from only reflecting upon regulating AI-powered systems, a second relevant regulatory field pertains to its raw material, namely data. A recent study by the International Data Corporation has found that the world has been seeing an annual increase of 61% in data volume and projected that by 2025 four times more data will exist than today.[69] Since Europe lags behind the USA and China in accumulating data,[70] European policymakers are well advised to boost initiatives that facilitate data sharing, its reuse and storage of both personal and non-personal data within the bounds of EU law.[71]

While sharing non-personal data is less problematic from the point of privacy laws, some researchers have already criticised the EU’s GDPR framework, specifically taking issue with Art. 5(1)(c), which they contend “created an artificial scarcity of data by making it more difficult for organizations to collect and share data” and put European companies and public agencies from the European Union at a disadvantage in developing and using AI-powered systems in comparison with the USA and China, where the processing of data, be it of personal or non-personal nature, is less strictly regulated.[72]

4. Trustworthy Artificial Intelligence

4.1. Europe's AI Strategy in the Context of a Global Competition on AI

The EU's comprehensive AI strategy was released one year after the publication of China's *New Generation Artificial Intelligence Development Plan* in which Beijing announced to "become the world's premier artificial intelligence innovation center" by 2030 and projected the value of its AI industry to reach US$150 billion in the next decade.[73]

What makes the European approach unique not only in comparison with the Chinese strategy but also the American AI Initiative, launched in February 2019 on an executive order by the US president,[74] is that the European debate on AI is primarily value-orientated and neither solely driven by business nor national security interests.[75] Most importantly, the European Union strives to regulate the use of AI-powered systems with a purpose similar to the creation of the GDPR framework.[76]

Respect for and protection of the rights of the individual in the digital age takes centre stage for EU policymakers. As stated in the EU's AI strategy, the goal is threefold:

1. "1. Boost the EU's technological and industrial capacity and AI uptake across the economy

2. Prepare for socio-economic changes and

3. Ensure an appropriate ethical and legal framework".[77]

Whereas the latest American AI Initiative focuses on similar goals as those mentioned in points 1 and 2 in the European AI strategy, the current US administration considers regulations to place "obstacles" for AI businesses and researchers because regulatory barriers might "cede the competitive edge to authoritarian governments who do not share {the American} values".[78] The American debate is therefore geared towards incentivising businesses and AI researchers to accelerate the development of AI by having less strict data protection rules in place, making considerably more investments into AI R&D programmes and American companies having more venture capital at their disposal than in the EU and China.[79] A study by McKinsey Global Institute (2018) estimated that both North America (EUR 12.1–18.6 billion) and Asia (EUR 6.5–9.7 billion) have surpassed the EU (EUR 2.4–3.2 billion) in private investments in AI R&D programmes already in 2016.[80]

Estimates also indicate that China has been made up considerable leeway on AI research, a field where the USA still remains in the leading position with home to above 10,000 top AI researchers.[81] For the EU to attract AI researchers or to at least entice its talented IT experts to stay in the EU,[82] the EU Commission intends to help establish AI research excellence centres in Europe and support Master and PhD programmes in AI willing to devote in total EUR 20 billion per year in the next decade to these goals.[83] First steps into this direction have been taken by increasing the funding scheme Horizon 2020 for research and innovation by 70%.[84] Until the end of next year, around EUR 1.5 billion are reserved for AI-related projects.[85]

Against this background, testing and experimentation facilities are considered to be especially effective in developing trustworthy AI for the transport, agriculture, manufacturing or healthcare sector.[86] Questions, however, remain as to how the operation of these testing centres can be squared with the strict data protection rules in the EU without preventing researchers and businesses from innovating with AI-powered technology.

Data has been described as the “raw-material” of AI-powered systems.[87] Kai-Fu Lee, a prominent Chinese venture capitalist and head of Google’s subsidiary in China in the previous decade, believes that data will play a significant role for AI innovation and the development of AI-powered systems as a whole.[88]

The main competitive advantage of the USA and China to the EU is the amount of data American private businesses such as Google, Facebook or Amazon, control, and the Chinese State possess. As stated by German Chancellor Angela Merkel: “In the US, control over personal data is privatised to a large extent. In China the opposite is true: the state has mounted a takeover”.[89] For the EU to succeed in the development of AI, German Chancellor Angela Merkel likes to see Brussels finding a middle course between these two approaches.[90]

The findings of an online consumer survey conducted by Northstar Research Partners (2017) with around 4000 participants from the USA, Europe and China give a first understanding to the question of why the EU is well advised to seek this middle ground and regulate the use of data by AI-powered systems in the future: “Asian respondents were the most bullish about the positive effect of AI in the future, while Europeans were the least optimistic”.[91] While additional studies should be carried out to shed more light on the still unexplored field of European citizens' perceptions towards AI, the report's conclusions indicate a general tendency of Europeans being more sceptical of the latest developments in the use of AI. In this regard, Franke (2019) cites one specific example, namely the coverage of the Cambridge Analytica scandal, which falls in line with other misuses of ICTs.[92]

Additionally, the fact that the EU has taken a firm stand against Google's and Facebook's dominance levying fines in the millions of dollars,[93] undergirds the argument that the EU Commission's actions not only raise awareness of data breaches among European citizens but also of large-tech companies responsibility for protecting European citizens data. This awareness of the potential for abuse of new technologies for economic purposes could additionally reflect in a rather more sceptical view among European citizens making it even more necessary to take regulatory initiatives that could increase trust in AI-powered systems.

4.2. Trustworthy AI—Reports by the EU High-Level Expert Group on AI

The EU's human-centric approach is reflected in two latest reports by the AI HLEG. Mandated by the EU Commission as part of its AI strategy to scrutinise the ethical, legal and societal implications of AI, a 52-member strong independent group comprising computer scientists, experts from industry and civil society and researchers with interdisciplinary background elaborated ethics guidelines[94] and developed a series of policy recommendations for "trustworthy AI".[95]

The underlying tenet of both documents is that "trustworthy AI" systems should be "lawful, ethical and robust, and fully aligned with fundamental rights" as enshrined in the EU Charter.[96] Both documents emphasise that "trustworthy AI" designed the EU ought to respect the human dignity in the digital age which can only be achieved if these systems are developed under the following principles:

(i) respect for human autonomy;

(ii) prevention of harm;

(iii) fairness; and

(iv) explicability.[97]

Having set a distant goal to become “a leader in cutting-edge AI”, the EU Commission aims to align social trust with economic competitiveness.[98] Establishing a unique trademark for “trustworthy AI”, AI-powered technology made in Europe could then not only be trusted among EU citizens but in countries around the world.[99]

In the light of the EU’s goal to facilitate the development of trustworthy AI, it is commendable to see the EU establishing a “European AI Alliance”, which allows a broad spectrum of actors to voice their opinions and shape the legal discourse on AI in the European Union.[100] As a complimentary initiative to the AI HLEG, ideas of around 500 participants have been incorporated in the work of the expert group.[101] The goal of developing AI for good and remaining competitive on the international plane requires taking the European citizens perceptions towards AI into account and evaluating them over a long time period. The proposed idea by the expert group to envisage a holistic 10-year plan to continuously assess the “opportunities and challenges of AI” and to monitor and “adapt {…} impactful actions on a short-term rolling basis” can be considered as a wise approach, since the “future of AI is full of unknown unknowns”.[102]

4.3. Regulating AI’s Raw Material—Data

AI-powered systems run on a myriad of both personal and non-personal data, which are for the time being only partially governed by a European regulatory framework initiated during the 5 years term of the Juncker Commission.[103]

While the GDPR laid the foundation for the protection of personal data controlled and processed by businesses and most public agencies servicing customers and citizens of the European Union, the Commission maintained its pace towards the creation of a Digital Single Market and recently complemented the GDPR with a regulation on the free flow of non-personal data.[104]

Projecting that the data economy will make up 5.4% of the EU’s annual GDP in 2025, these legislative measures have been agreed upon with a view to removing national barriers and to enabling the private and public sector to “make the most out of data and its opportunities”.[105] The latest regulation is aimed at facilitating EU wide data storage and processing within the bounds of EU law.[106] In this regard, it is also worthwhile highlighting the goal of the EU to create common data spaces as part of the Digital Single Market strategy. This initiative bears special importance for the collection of new data, which is necessary to train AI-powered systems and allows for the creation of innovate goods.[107]

5. The Way Forward—Concluding Remarks

Briefly before the presentation of the newly appointed EU Commissioners, information has surfaced about the EU's resolve to restrict any potential "indiscriminate use of facial recognition technology" not only by businesses but also by public authorities.[108] Granting the European citizens, the right to "know when {facial recognition} data is used" testifies to the fact that the EU continues to build on the principles of the GDPR.[109] Finding a prudent balance between business-friendly legislation, legitimate national security interests and the protection of EU citizens' fundamental rights as to privacy remains a delicate task in the digital era.

Whereas the Juncker Commission avoided legislating on AI-powered systems directly and instead brought forward legislation on data protection, cyber-security and telecommunication, the next president of the EU Commission, Ursula von der Leyen, did not shy away from making the bold promise for "put[ting] forward legislation for a coordinated European approach on the human and ethical implications of artificial intelligence drafting" after taking office in the beginning of November this year.[110]

von der Leyen's goal falls in line with previous statements made by the French president Macron and German chancellor Merkel, who both advocate regulating the use of AI-powered systems on the European level, to "recreate a European sovereignty in AI" and to elucidate that AI "serves humanity".[111]

The European Union has deliberately chosen the path towards regulating AI-powered systems and remains resolved to shape the future of the Digital Single Market. While the AI High-Level Experts has created the architecture for a future regulatory regime of trustworthy AI, fundamental questions remain as to how the ambitious goal of becoming the "leader in cutting-edge AI that can be trusted throughout the world"[112] can be attained within the next decade.

Whereas tailored investments into the digital infrastructure of EU Member States such as Germany are of utmost importance to remain competitive at world level, the European Union is well advised not to succumb to the temptation of believing that it should run an AI race together with the USA and China. It should rather use the current AI momentum at the European level and invest in AI research that will be the key to translate or encode the EU ethical and legal guidelines into AI-powered machines. Whether this core, technical question can be answered at all, remains a "black-box" that can only be opened if the programmers or designers of these systems find a common language with all relevant stakeholders, and most importantly, with the policymakers.

In the words of the AI HLEG: “In a competition amongst different economic entities, the single most important element is the capability to apply and learn fast and consistently over a long period of time”.[113]

Bibliography

Books, Journal Articles and Reports

Boulanin, V. (Ed.). (2019). *The Impact of artifical intelligence on stategic stability and nuclear risk* (Vol. 1). SIPRI: Euro-Atlantic Perspectives.

Buchanan, B., & Miller, T. (2017). *Machine learning for policymakers: What it is and why it matters*. Harvard Kennedy School—The Cyber Security Project.

Buiten, M. (2018). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation, 10*(1), 41–59.

Castro, L. D., & Lago, J. R. R. (2018). Educación y diplomacia cultural en la primavera de Europa (1948–1954) = Education and cultural diplomacy in the spring of Europe (1948–1954). *Revista de Educación, 383,* 63–84.

Castro, D., McLaughlin, M., & Chivot, E. (2019) Who is winning the AI Race: China, the EU or the United States? Available at: https://www.datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states/ . Accessed: September 6, 2019.

Coglianese, C., & Lehr, D. (2017). Regulating by Robot: Administrative decision making in the machine-learning era. *The Georgetown Law Journal, 105,* 1147–1223.

Craglia, M. (Ed.). (2018). Artificial intelligence—A European perspective. Publications Office of the European Union. EUR 29425 EN.

Färber, K. (2017). Mitterrand and the great European design—From the cold war to the European union. *Baltic Journal of European Studies, 7*(2), 132–147.

Floridi, L. (2019). What the near future of artificial intelligence could be. *Philosophy & Technology, 32,* 1–15.

Franke, U. (2019). Harnessing artificial intelligence. *European Council on Foreign Relations, 289,* 1–9.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: The MIT Press.

Hamulák, O. (2018). La carta de los derechos fundamentales de la union europea y los derechos sociales. *Estudios constitucionales, 16*(1), 167–186.

High-Level Expert Group on AI. (2019a). *Ethics guidelines for trustworthy AI*. European Commission.

High-Level Expert Group on AI. (2019b). *Policy and investment recommendations for trustworthy AI*. European Commission.

Jia, J., Zhe Jin, G., & Wagman, L. (2018). *The short-run effects of GDPR on technology venture investment* (pp. 1–42). NBER Working Paper No. 25248.

Kerikmäe, T., & Särav, S. (2017). Paradigms for automatization of logic and legal reasoning. In: D. Krimphove, G. M. Lentner (Eds.). *Law and logic: Contemporary issues* (205–222). Duncker & Humblot.

Kerikmäe, T., Troitiño, D. R., & Shumilo, O. (2019). An idol or an ideal? A case study of Estonian e-governance: Public perceptions, myths and misbeliefs1. *Acta Baltica Historiae et Philosophiae scientiarum*, *7*(1).

Kerikmäe, T. (2019) *Autonoomsed intelligentsed tehnoloogiad ja õigusruum. Eesti Krati tegevuskava koostamise lõpuseminar, TalTech*. Tallinn: Majandus- ja Kommunikatsiooniministeerium, Riigikantselei. May 28, 2019

Lee, K. F. (2018). *AI superpowers*. Boston: Houghton Mifflin Harcourt.

Lessig, L. (1999). The law of the horse: What cyberlaw might teach. *Harvard Law Review, 113*(2), 501–549.

Lessig, L. (2006). *Code: Version 2.0* (2nd ed.). New York: Basic Books.

Martín de la Guardia, R. M., & Pérez Sánchez, G. A. (2001). *Historia de la integración europea*. Barcelona: Ariel.

McCarthy, J., et al. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Available at: http://www-formal.stanford.edu/jmc/history/dartmouth.pdf . Accessed: September 16, 2019.

McKinsey Global Institute. (2018). *Notes from the AI Frontier: Modeling the* *impact of AI on the world economy*. New York: McKinsey & Company. Available at: https://www.mckinsey.com/~/media/McKinsey/Featured-Insights/Artificial-Intelligence/Notes-from-the-frontier-Modeling-the-impact-of-AI-on-the-world-economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-ec . Accessed: September 15, 2019.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 1–21

Northstar. (2017) *AI today, AI tomorrow: Awareness, acceptance and anticipation of AI: A global consumer perspective*. Available at: https://www.arm.com/solutions/artificial-intelligence/survey . Accessed: September 6, 2019.

OECD. (2019). *Artificial intelligence in society*. Paris: OECD Publishing.

Outeda, C. C. (2001). *El proceso de constitucionalización de la Unión Europea: de Roma a Niza* (No. 3). Univ Santiago de Compostela.

Pawlas, I. (2015, September). The Visegrad Countries and European Union membership-selected issues. In *Proceedings of International Academic Conferences* (No. 2704866). International Institute of Social and Economic Sciences.

Ramiro Troitino, D., & Pando Ballesteros, M. D. L. P. (2017). Churchill’s European integration model. *REVISTA DE OCCIDENTE, 433,* 57–71.

Reinsel, D., Gantz, J., & Rydning, J. (2018). The digitization of the world: From edge to core. *An IDC White Paper*. Available at: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf . Accessed: September 17, 2019.

Russell, S. J., & Norvig, P. (Eds.). (2016). *Artificial intelligence: A modern approach* (3rd ed.). Essex: Pearson education.

Scherer, M. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies and strategies. *Harvard Journal of Law & Technology, 29*(2), 353–400.

Solomonoff, G. (1956). *Ray Solomonoff and the Dartmouth summer research project in artificial intelligence*. Available at http://raysolomonoff.com/dartmouth/dartray.pdf . Accessed: September 16, 2019.

Spielkamp, M. (Ed.) (2019). *Automating society: Taking stock of automated decision-making in the EU. AlgorithmWatch*. Available at: https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf . Accessed: September 17, 2019.

Surden, H. (2014). Machine learning and law. *Washington Law Review, 89,* 87–115.

Tegmark, M. (2016). *Life 3.0: Being human in the age of artificial intelligence*. London: Penguin Books.

Troitiño, D. R. (2017). Jean Monnet before the first European Community: A historical perspective and critic. *Trames, 21*(3), 193–213.

Troitiño, D. R., Kerikmäe, T., & Chochia, A. (Eds.). (2018). *Brexit: history, reasoning and perspectives*. Berlin: Springer.

Van den Hoven, J. (2017) Ethics for the digital age: Where are the moral specs?—Value sensitive design and responsible innovation. In: H. Werthner, & F. van Harmelen (Eds.). *Informatics in the future* (pp. 65–76).

Von der Leyen, U. (2019) *A union that strives for more—My agenda for Europe—By candidate for President of the European Commission*. Political Guidelines for the Next European Commission 2019–2024.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2018). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration* 1–20.

Others

COM. (2018a). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe. 237 final

COM. (2018b). Coordinated Plan on Artificial Intelligence. 795 final

COM. (2019). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions—Building Trust in Human Centric Artificial Intelligence, 168

Dignum, V. (2019). *There is no AI—race and if there is, it's the wrong one to run*. Alliance on Artificial Intelligence. Available at: https://allai.nl/there-is-no-ai-race/ . Accessed: 6 September 2019)

e-Estonia. (2019). Estonia accelerates artificial intelligence *development*. Available at: https://e-estonia.com/estonia-accelerates-artificial-intelligence/ . Accessed: September 18, 2019.

European Commission. (2019a). *AI landscape and indicators—AI players*. Available at: https://ec.europa.eu/knowledge4policy/ai-watch/topic/ai-landscape-indicators_en . Accessed: September 6, 2019.

European Commission. (2019b). *AI players in Europe*. Available at: https://ec.europa.eu/knowledge4policy/visualisation/ai-players-europe_en . Accessed: September 6, 2019.

European Commission. (2019c). *A definition of AI: Main capabilities and scientific disciplines High-Level Expert Group on artificial intelligence*. 8 April. Available at: https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines . Accessed: September 18, 2019.

European Commission. (2019d). *Digital single market: Commission publishes guidance on free flow of non-personal data*. Brussels, 29 May. Available at: https://europa.eu/rapid/press-release_IP-19–2749_en.htm . Accessed: September 18, 2019.

European Commission. (2019e). *The European AI alliance*. Available at: https://ec.europa.eu/digital-single-market/en/european-ai-alliance . Accessed: September 18, 2019.

Future of Life Institute. (2019). *AI Policy—China*. Available at https://futureoflife.org/ai-policy-china/ . Accessed: September 6, 2019.

Kayali, L. (2019, 18 July). Next European Commission takes aim at AI Artificial intelligence will be the next GDPR. *Politico*. Available at https://www.politico.eu/article/ai-data-regulator-rules-next-european-commission-takes-aim/ . Accessed: September 18, 2019.

Khan, M. (2019, 22 August). EU plans sweeping regulation of facial recognition. *Financial Times*. Available at: https://www.ft.com/content/90ce2dce-c413–11e9-a8e9-296ca66511c9 . Accessed: September 14, 2019.

Knight, W. (2019). Facebook is making its own AI deepfakes to head off a disinformation disaster. *MIT Technology Review*. Available at: https://www.technologyreview.com/s/614269/facebook-is-making-ai-deepfakes-to-head-off-a-disinformation-disaster/ . Accessed: September 17, 2019.

Kratid. (2019). *Estonia will have an artificial intelligence, or so-called Kratt strategy*. Available at: https://www.kratid.ee/in-english . Accessed: September 18, 2019.

Kratsios, M. (2019, 22 May). Artificial intelligence: Next steps. *The Forum—Network*. Available at: https://www.oecd-forum.org/users/262053-michael-kratsios/posts/49175-artificial-intelligence-next-steps . Accessed: September, 18 2019.

Mozur, P. (2017, 20 July). Beijing wants A.I. to be made in China by 2030. *The New York Times*. Available at: https://www.nytimes.com/2017/07/20/business/china-artificial-intelligence.html . Accessed: September 6, 2019.

Mozur, P. (2019, 14 April). One month, 500,000 face scans: How China is using A.I. to profile a minority. *The New York Times*. Available at: https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html). Accessed: September 17, 2019.

O’Meara, S. (2019, 21 August). Will China lead the world in AI by 2030? *Nature*. Available at: https://www.nature.com/articles/d41586-019-02360-7 . Accessed: September 16, 2019.

Pearl, A. (2017, 2 June). Homage to John McCarthy, the father of artificial intelligence (AI). *Artificial Solutions*. Available at: https://www.artificial-solutions.com/blog/homage-to-john-mccarthy-the-father-of-artificial-intelligence . Accessed: September 16, 2019.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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) L118, 4 May 2016, pp. 1–88

Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union. PE/53/2018/REV/1.

Satariano, A. (2019, 21 January). Google is fined $57 million under Europe’s data privacy law. *The New York Times*. Available at: https://www.nytimes.com/2

019/03/20/business/google-fine-advertising.html . Accessed: September 18, 2019.

The Economist. (2018, 20 September). Big data, small politics: Can the EU become another AI superpower? *The Economist*. Available at https://www.economist.com/business/2018/09/20/can-the-eu-become-another-ai-superpower . Accessed: September 18, 2019.

The White House. (2019). Accelerating American Leadership in Artificial Intelligence. 11 February. Available at: https://www.whitehouse.gov/articles/accelerating-americas-leadership-in-artificial-intelligence/). Accessed: September 18, 2019.

Thierer, A, O'Sullivan, A., & Russel, R. (2017). *Artificial intelligence and public policy*. Mercatus Center—George Mason University. Available at https://www.mercatus.org/publications/technology-and-innovation/artificial-intelligence-and-public-policy . Accessed September 16, 2019.

U.S. House of Representatives Committee on Science, Space, & Technology. (2019). *Artificial Intelligence: Societal and Ethical Implications*. 26 June. Statement by Clark, J. Between minute 1:00:30 and 1:01:30. Available at https://science.house.gov/hearings/artificial-intelligence-societal-and-ethical-implications . Accessed: September 16, 2019.

Vincent, J. (2018, 27 November). This is when AI's top researchers think artificial general intelligence will be achieved. *The Verge*. Available at: https://www.theverge.com/2018/11/27/18114362/ai-artificial-general-intelligence-when-achieved-martin-ford-book . Accessed: September 16, 2019.

[1] COM (2018a) 237 final.

[2] Castro et al. (2019); for a critical account on the (term) "AI race", see: Dignum (2019).

[3] The term "AI players" includes research centres, academic institutions and businesses which have taken part in at least one project related to AI (such as R&D processes, industrial production and marketing, specific AI-related services). According to the EU Commission, "the EU is among the geographical zones with the highest number of players active in AI (25%), just behind the United States (26%), and just ahead of China (24%)" European Commission (2019b).

[4] For estimates for the year 2017, consider study by: Castro et al. (2019), p. 5.

[5] This number may be overestimated, as during the current hyperbole around AI, some tech businesses want to be identified as an AI company for economic reasons, see: Castro et al. (2019), pp. 27, 84.

[6] European Commission (2019a) and Craglia (2018), pp. 12–13.

[7] COM (2018a) 237 final, p. 2.

[8] OECD (2019), pp. 40–41.

[9] McKinsey Global Institute (2018), p. 3.

[10] See OECD principles on AI:

"1. AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.

2. AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards—for example, enabling human intervention where necessary—to ensure a fair and just society.

3. There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.

4. AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.

5. Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles" OECD (2019).

[11] OECD (2019), p. 3.

[12] Craglia (2018), p. 10.

[13] High-Level Expert Group on AI (2019a), p. 12.

[14] Craglia (2018), pp. 8, 63.

[15] COM (2018a) 237 final, p. 1.

[16] European Commission (2019a): "AI is rapidly applying to almost all economic sectors. However, {…} no official statistics are available yet, so AI is escaping traditional industrial and product classifications". the OECD's estimates on global private equity investment into AI remain one of the few official statistics on one specific aspect of AI, even though they are also based on figures by Crunchbase (July 2018).

[17] For a detailed account on the methodology of the study, see: Castro et al. (2019), pp. 13–15.

[18] Tegmark (2017), p. 49.

[19] See, e.g., Tegmark (2016); for a lucid overview of the different views on AGI among acclaimed AI researchers, see: Vincent (2018).

[20] Craglia (2018), p. 16.

[21] Russell and Norvig (2016), pp. 4–8.

[22] OECD (2019), p. 47.

[23] Ibid.

[24] Solomonoff (1956) and Pearl (2017).

[25] McCarthy et al. (1955).

[26] Russell and Norvig (2016), pp. 3, 17.

[27] Russell and Norvig (2016) pp. 16–27 and Boulanin (2019) pp. 14–15.

[28] COM (2018a) 237 final, p. 2 and Boulanin (2019), p. 3.

[29] Färber (2017).

[30] Surden (2014), pp. 87–88.

[31] Craglia (2018), p. 20.

[32] European Commission (2019c), p. 4.

[33] Goodfellow et al. (2016), pp. 1, 5, 8 and Boulanin (2019), p. 16.

[34] Wirtz et al. (2018), p. 9.

[35] Ramiro Troitino and Pando Ballesteros (2017).

[36] European Commission (2019c), p. 5, see also: Kerikmäe and Särav (2017), p. 219.

[37] Ibid.

[38] As also discussed in the context of AI applications in the field of Legal Tech: Kerikmäe et al. (2018), p. 92.

[39] Buchanan and Miller (2017), pp. 26, 46 and Coglianese and Lehr (2017), p. 1155.

[40] O'Meara (2019).

[41] US House of Representatives Committee on Science, Space, & Technology (2019).

[42] Scherer (2016) and Thierer et al. (2017), p. 31.

[43] Mittelstadt et al (2016), p. 3 and Buiten (2018), p. 56.

[44] For a lucid overview, consider: Press (2019).

[45] Buchanan and Miller (2017), pp. 32–36.

[46] Mittelstadt et al (2016), p. 9.

[47] Craglia (2018), pp. 89–90 and Buchanan and Miller (2017), pp. 39–40.

[48] Knight (2019).

[49] Outeda (2001).

[50] Mozur (2019).

[51] For cases of automated decision making in the European Union, see report by AlgorithmWatch: Spielkamp (2019), p. 35 *ff.*

[52] Kerikmäe et al. (2019), p. 1 and Kratid (2019).

[53] Kerikmäe (2019) and e-Estonia (2019).

[54] Troitiño (2017).

[55] Easterbrook (1996), p. 207: "{T}he best way to learn the law applicable to specialized endeavors is to study general rules. Lots of cases deal with sales of horses; others deal with people kicked by horses; still more deal with the licensing and racing of horses, or with the care veterinarians give to horses, or with prizes at horse shows. Any effort to collect these strands into a course on "The Law of the Horse" is doomed to be shallow and to miss unifying principles".

[56] Ibid.

[57] Lessig (1999), p. 502.

[58] Lessig (1999), p. 506.

[59] Ibid.

[60] Lessig (1999), p. 507.

[61] Lessig (1999), p. 546.

[62] Van den Hoven (2017), p. 70.

[63] Lessig (2006), p. 5.

[64] Buiten (2018), p. 49.

[65] Ibid.

[66] Pawlas, (2015).

[67] Kerikmäe et al. (2019), p. 10.

[68] COM (2018b) 795 final, p. 9.

[69] Reinsel et al. (2018).

[70] Castro et al. (2019), p. 42.

[71] Hamulák (2018).

[72] Ibid; Franke (2019), pp. 3–5; for a study on the negative effects of the GDPR on investments in European technology companies, see: Jia, Zhe Jin and Wagman (2018); for Art. 5(1)(c) of the GDPR (Principles relating to processing of personal data), see: Regulation (EU) 2016/679: “Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (‘data minimisation’)”.

[73] Future of Life Institute (2019) and Mozur (2017).

[74] The White House (2019).

[75] Martín de la Guardia and Pérez Sánchez (2001).

[76] Kayali (2019).

[77] COM (2018a) 237 final, p. 4.

[78] Kratsios (2019).

[79] Castro et al. (2019), pp. 2–3, 10, 18, 25.

[80] McKinsey Global Institute (2018) (as cited in COM (2018a) 237 final, p. 5).

[81] Castro et al. (2019), p. 5.

[82] According to the “2018 Silicon Valley Index” by Joint Venture Silicon Valley, only in Silicon Valley around 240,000 Europeans worked for tech businesses in 2018 (as cited in COM (2018b) 795 final, p. 5).

[83] COM (2018b) 795 final, pp. 5–6 and COM (2018a) 237 final, p. 10.

[84] COM (2019) 168, p. 1.

[85] Ibid.

[86] Ibid.

[87] COM (2018a) 237 final, p. 3.

[88] Lee (2018).

[89] The Economist (2018).

[90] Ibid.

[91] Northstar (2017), p. 12.

[92] Franke (2019), p. 5.

[93] Satariano (2019).

[94] Castro and Lago (2018).

[95] High-Level Expert Group on (2019a) and High-Level Expert Group on AI (2019b).

[96] High-Level Expert Group on AI (2019a), p. 49.

[97] High-Level Expert Group on AI (2019a), p. 12.

[98] COM (2019) 168, pp. 9–10.

[99] Ibid.

[100] European Commission (2019e).

[101] Ibid.

[102] High-Level Expert Group on AI (2019b), p. 49 and Floridi (2019), p. 13.

[103] Troitiño et al. (2018).

[104] Regulation (EU) 2018/1807.

[105] Vice-President for the Digital Single Market Andrus Ansip quoted in: European Commission (2019d).

[106] Ibid.

[107] Ibid.

[108] Khan (2019).

[109] Ibid.

[110] von der Leyen (2019).

[111] Franke (2019), p. 6 and Kayali (2019).

[112] COM (2019) 168, p. 9.

[113] High-Level Expert Group on AI (2019b), p. 49.

Publication II

3.1 Antonov, A., Häring, T., Korõtko, T., Rosin, A., Kerikmäe, T., & Biechl, H. (2021). Pitfalls of Machine Learning Methods in Smart Grids: A Legal Perspective. In Proceedings - 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC), 12-14 November 2021, Rome (pp. 248–256). Institute of Electrical and Electronics Engineers (IEEE). Danvers. https://doi.org/10.1109/ISCSIC54682.2021.00053

Pitfalls of Machine Learning Methods in Smart Grids: A Legal Perspective

Alexander Antonov
Department of Law
Tallinn University of Technology
Tallinn, Estonia
+372 6202002
alanto@ttu.ee

Tobias Häring
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Smart City Center of Excellence (Finest Twins)
Tallinn, Estonia
tobias.haring@taltech.ee

Tarmo Korõtko
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Smart City Center of Excellence (Finest Twins)
Tallinn, Estonia
tarmo.korotko@taltech.ee

Argo Rosin
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Smart City Center of Excellence (Finest Twins)
Tallinn, Estonia
argo.rosin@taltech.ee

Tanel Kerikmäe
Department of Law
Tallinn University of Technology
Tallinn, Estonia
tanel.kerikmae@taltech.ee

Helmuth Biechl
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Tallinn, Estonia
Smart City Center of Excellence (Finest Twins)
Institute of Electrical Power Systems (IEES), University of Applied Sciences Kempten
Kempten, Germany
biechl@hs-kempten.de

***Abstract*—The widespread implementation of smart meters (SM) and the deployment of the advanced metering infrastructure (AMI) provide large amounts of fine-grained data on prosumers. Machine learning (ML) algorithms are used in different techniques, e.g. non-intrusive load monitoring (NILM), to extract useful information from collected data. However, the use of ML algorithms to gain insight on prosumer behavior and characteristics raises not only numerous technical but also legal concerns. This paper maps electricity prosumer concerns towards the AMI and its ML based analytical tools in terms of data protection, privacy and cybersecurity and conducts a legal analysis of the identified prosumer concerns within the context of the EU regulatory frameworks. By mapping the concerns referred to in the technical literature, the main aim of the paper is to provide a legal perspective on those concerns. The output of this paper is a visual tool in form of a table, meant to guide prosumers, utility, technology and energy service providers. It shows the areas that need increased attention when dealing with specific prosumer concerns as identified in the technical literature.**

Keywords-Machine Learning; GDPR; Cybersecurity; EU; Smart City; Smart Grid

I. INTRODUCTION

Within the context of the Third Energy package and the latest Clean Energy for all Europeans Package, the EU made the roll out of smart meters (SM) mandatory to enable residential end-users a better overview of their energy consumption and raise energy efficiency [1]. However, the transformation of energy systems raises various legal concerns, specifically in terms data protection, privacy and cybersecurity [1]. While the deployment of SM allows for real-time tracking of individual households' energy consumption, it might bear reverse effects on their autonomy and potentially affect their fundamental rights in the areas of data protection and privacy.

This is especially evident in applications such as pattern-recognition and profiling which machine learning (ML) facilitates. Latest increases in malicious cyber operations by state proxies against states' critical infrastructure or "essential services" [2], which includes electricity grids, pose an additional challenge to the application of SM.

Smart appliances and home energy management systems (HEMS) are gaining popularity in smart grids in the EU. Renewable energy sources of buildings are typically connected to a HEMS, which shifts the building from a passive role as electricity consumer into an active role as prosumer [3] [4]. To facilitate prosumer needs for auxiliary electricity services, the distribution system operator (DSO) is required to install SM.

Compared to legacy metering equipment, SMs enable improved measurements at shorter sampling intervals and provide additional functionality. Along with enhanced data collection and analysis tools, SMs are part of the advanced metering infrastructure (AMI), which is an essential component of modern electricity grids and smart cities. The fine-grained measurements and increased amounts of data enable the implementation of machine learning (ML) based analytical tools for various purposes e.g. energy flexibility analysis [5], non-intrusive load monitoring (NILM) [6] etc.

Although the analysis of AMI data enables efficient optimization methods, it is also recognized to raise numerous privacy and security issues [7], [8], [9]. Widespread use of ML algorithms further increases end-user concerns, since technical publications about machine learning approaches to NILM, e.g. Factorial Hidden Markov Models (FHMM) [10] or Neural Networks (NN) [11], rarely take privacy or cyber security aspects into account. Some publications even suggest the breach of end-user privacy through the implementation of additional occupancy monitoring measures [12].

Against this backdrop, it is identified that there is a need to map electricity prosumer concerns towards the AMI and its ML based analytical tools and analyze how these concerns could be addressed from a legal perspective with a view to raising ethical and legal awareness about potential pitfalls of ML methods, specifically from the perspective of accountability for potential data and cybersecurity breaches. Taking the latest regulatory initiatives of the EU in the areas of data protection, privacy and cybersecurity into account, the General Data Protection Regulation (GDPR) in particular, the paper is predicated on the assumption that the EU's approach towards the governance of new technologies such as SM presents a unique case in addressing these concerns.

Having mapped prosumer concerns towards the AMI, a technical analysis of the identified prosumer concerns in terms of the ML based analytical tools is conducted. The concerns identified in the technical literature are then analyzed from a legal perspective. For this purpose, pertinent EU legislative frameworks and deliverables by the European Commission's Smart Grid Task Force 2 (SGTF) are consulted [13], [14]. The authors suggest a visual tool in form of a table to provide guidance to prosumers, utility, technology and energy service providers for identifying and addressing prosumer concerns mapped in the technical literature.

The terms prosumer and active customer are applied interchangeably in this paper. The latter term is defined in Electricity Directive (ED), Art. 2(8) [15]. This paper treats prosumers and active customers as a special category of consumers.

The paper is organized as follows: The analysis of general user concerns for AMI are presented in Section 2. In Section 3 the technically relevant concerns are identified and then connected to relevant regulatory frameworks in Section 4. Finally, the conclusions with general recommendations are presented in Section 5.

II. ANALYSIS OF USER CONCERNS FOR AMI AND ML IN GENERAL

The AMI is a common application of electricity smart grids, which spreads across all its fields and domains and integrates relevant technologies for bidirectional communication between utilities and prosumers [16], [17], [18], [19], [20]. The AMI provides services for customers, suppliers and network operators and is used for automated meter reading, billing, information provision, event management, device configuration etc. A common configuration of the AMI is depicted on Figure 1. Common components of the AMI include SMs, hierarchically disposed communication networks, Meter Data Management Systems (MDMS) and Head-End Systems (HES). The HES is a central data system for exchanging data of various meters in its service area. The communication network of the AMI is primarily divided into three sections: home area networks (HAN), wide area networks (WAN), and the utility network. The MDMSs act as meter data concentrators and as gateways between the WAN and utility network. SMs are the coupling points of users into the AMI, which provide enhanced metering capabilities, data communication and optional auxiliary functions, e.g. the adjustment of energy use based on cost and availability [21], [22], [23]. SMs are used to report, measure and monitor power quality metrics, as well as loading conditions and power flows, which make them essential operational components and data sources for analytics.

The availability to utilize ML algorithms on fine-grained data at different parts of the AMI raises numerous concerns for residential prosumers. A literature survey was carried out to gain insight into the concerns of electricity end-users regarding the AMI and ML based analytical tools and more prominent concerns are outlined in Table I. Additional concerns of electricity prosumers, which do not utilize ML algorithms, include theft of data, eavesdropping, denial of ICT services, compromise of data integrity, hijacking of home appliances, energy theft, tampering of SMs and denial of power.

To address individual concerns, it is necessary to identify their origin. The AMI is a complex technological system, which reveals several surfaces for intrusion or other forms of cyber-attacks. For the classification of the origin of prosumer concerns, surfaces for cyber-attacks in the AMI, identified in [24], are adopted. The following surfaces of the AMI for cyber-attacks are distinguished in Table II.

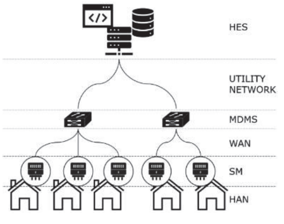


Figure 1. Common AMI configuration.

TABLE I. RESIDENTIAL PROSUMER CONCERNS REGARDING THE AMI AND ML ALGORITHMS [19], [20], [23], [25], [26], [27], [28], [29], [30]

Prosumer concern	Description	ID
Privacy		
Price discrimination	Variance in consumer pricing based on consumer profile	P1
Denial of consumer services	Denied access to consumer services due to unsuitable consumer profile	P2
Target to excessive advertisements	Increased advertisements, since consumer identified as target group by consumer profile	P3
Identification of home appliances	Unwanted identification of individual home appliances through NILM	P4
Exhibition of user habits and lifestyle	Exposure of sensitive data regarding consumer habits through NILM	P5
Exhibition of illnesses and disabilities	Exposure of sensitive health data through NILM	P6
Personification of anonymous data	The personification of data deemed to be collected anonymously through ML algorithms	P7
Cyber Security		
Disconnection of home appliances	The manipulation of demand response (DR) programs through the tampering of ML training and input data	C1
Burglary, arson, vandalism etc.	Increased threat through occupancy information gained by NILM	C2
Attractive target to burglary	Increased likelihood of burglary due to identification of attractive appliances through NILM	C3
Target to kidnapping	Possibility to use NILM for identifying persons in vulnerable situations	C4
Denial of personal mobility	The manipulation of DR programs through the tampering of ML training and input data to deny charging of electric vehicles	C5

TABLE II. SURFACES OF AMI

Abbr.	Description
HAN	the consumer side of the AMI. A consumer gateway acts as a bridge between the smart meter and the consumer's home devices
SM	the primary point of data collection for power grid energy consumption. Physical access to the meter is considered a vulnerable attack surface
SM data collector (SMDC)	a hardware computing device aggregating real-time data from multiple smart meters and providing a data collection and management point for the utility. An integral part of the MDMSs
AMI comm. interfaces and network	the network along with used communication interfaces linking the smart meter and the SMDCs. The AMI communications network exists alongside the power grid and can be scaled to serve millions of smart meters
AMI comm. protocols and software	the communication links and protocols utilized by the AMI
HES	the AMI management platform at the utility installation. Provides data warehousing for collected data and centralized management of the AMI

An estimation about the relevance of each listed surface regarding each individual problem is provided in Table VIII. To evaluate the user concerns stemming from increased use of ML algorithms in the AMI, the technical process enabling such actions needs to be studied.

III. ANALYSIS OF TECHNICAL PROCESS OF ML IN AMI AND IDENTIFICATION OF RELATED PROSUMER CONCERNS

The basic process behind the disaggregation of load patterns from smart meter data, or NILM, is shown in Table III.

It is the same for all different proposed ML methods, like FHMMs, NNs or Support Vector Machines, the main differences can be found in the amount, resolution and detail of the collected data, the amount auf auxiliary data measurements of additional information, and the way the obtained data is intended to be used after the identification of the loads.

A. Differences in Data Collection

For many publications on NILM different public datasets are used. A detailed overview of the differences is shown in [31]. Some publications rely on their own measurement data, which makes comparisons of the performance more difficult. Table IV shows an overview of the used datasets in selected recent publications.

If a typical percentage of 60-70% of the datasets was used for the training of the ML algorithms, it can be seen from Table IV that in most publications the data amount is large. Months and years of training data with small resolutions of less than 5min, thus high detail on the time of the energy consumption, are used. Only few datasets contain less than a month of data and/or a resolution of more than 5min. It should be noted that none of these technical papers discuss privacy concerns about the collected data and their use.

B. Additional Data Acquisition and Additionally Proposed Features

Some of the recent publications on NILM present the use of some additional data measurements to improve the disaggregation results. In [32] an additional voluntary user feedback about the disaggregated data is added. Authors of [12] propose the use of cameras, motion sensors and smartphone apps, to track the occupancy of the household. An additional smartphone application is developed in [33] to display the results to the prosumers in a structured way. In [10] a cloud based on-line monitoring approach is presented. The authors of [34] show a novelty detection function for their ML method for new appliances. Future research of [35] includes classifying the prosumer activities for better accuracy and in [36] the authors' future goal is to influence the prosumers' behavior to increase energy efficiency. Privacy and cyber security are not discussed in any of these publications.

C. Comparison Based on Metrics

Since the metrics for measuring the accuracy of the different NILM methods is not unified and the publications use different datasets for training and testing of their proposed methods, direct comprehensive comparisons can be more difficult. Additionally, different devices in the datasets result in different accuracies.

TABLE III. NILM PROCESS STAGES [37]

Stage	Description
Metering	Data is collected from smart meters and sometimes additional measurement equipment, typically with a low frequency (including current, voltage and power data)
Event detection	Events are detected within the data sets: e.g. an appliance changed its state
Feature extraction	Every appliance has a certain load signature and features, by which it can be distinguished from others
Classification	Load identification by a classification procedure to determine the times or periods a device was operating
Analysis of classification	Based on the application the NILM-process is used for, the classification can be analyzed

TABLE IV. OVERVIEW OF TRAINING DATASETS

Dataset	Duration/Resolution	Publication
Pecan Street	4Y/1min	[38], [39], [40]
REDD	2-4W/<=4s	[10], [36], [40]
UK-DALE	655D/<=6s	[35], [41], [10]
ECO	8M/1s	[11], [42]
BLUED	1W/<=1s	[34]
Challekere Campus	7D/2min	[33]
Private Dataset	1M/10s	[12]
Private Dataset	1M/30min	[32]

Measuring privacy is not unified as well. It usually has qualitative and quantitative aspects which makes it difficult to use some simple scoring system. Literature proposes either complicated quantitative methods or qualitative methods for privacy evaluation [43].

Therefore, a simplified scoring system has been developed to provide a rough overview of the correlation between the accuracy of ML methods and their privacy. The framework is not based on specific standards but aims to provide a quick categorization of ML techniques for NILM.

The privacy score is designed to have 6 levels from -10 to -35. The best achievable privacy score is -10 and the worst is -35. The privacy level is estimated by the amount of used data for training the algorithm and additional data acquisition methods. A low amount of used data is considered to have a lower impact on the prosumers' privacy. Therefore, the score is -5. If the used amount of data is higher, then the score is -10. The threshold for this is chosen to be 1 month of data. Many prosumers do not like their data to be processed in a cloud, so this gives an additional score of -5. Additional occupancy monitoring with cameras is considered a huge violation of privacy and therefore gets an additional score of -10.

The accuracy of the ML methods is usually shown as an accuracy value (ACC) or F1 score (F1). The two values are shown with different colors in Figure 2 as they are being calculated differently and therefore cannot be compared directly.

The metrics are shown in Table V and the simple privacy score in connection with a simplified accuracy and F1 score is shown in Figure 2.

As a result for the general process of NILM, the two figures show clearly that the accuracy of the NILM methods is directly correlated to the reduction of privacy. A higher amount of data that can be used as a training set improves the accuracy of ML methods but reduces the privacy and of the prosumer as more data is stored. Additional available data can also improve the accuracy but for the example of occupancy monitoring [12] it reduces the privacy level.

TABLE V. PRIVACY METRICS FOR COMPARISON

Measure	Privacy Score	Otherwise
Dataset < 1 Month	-5	-10
Resolution of Data > 5min	-5	-10
Occupancy Monitoring	-10	0
Cloud Processing	-5	0

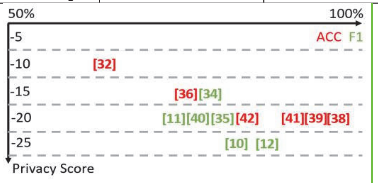


Figure 2. ACC and F1 score compared to proposed privacy score for selected publications.

D. Proposed Applications for NILM

NILM methods are used for different purposes and applications in Smart Grids. HEMS, ambient assisted living (AAL), recommender systems (RS) and fault diagnostics (FD) are the most common implementations [37]. The goal and purpose of NILM is different for each of these applications. Sometimes power on/off detection or power estimations are necessary [38], sometimes predictions for more efficient home energy management are needed [44]. Sometimes the goal is a recommendation on more efficient energy consumption or faults and unusual behavior can be detected in the ambient assisted living context [37]. For all these specific applications the privacy and cyber security concerns are identified individually, based on the stages of the NILM ML process considering implementation on different surfaces of AMI. This is shown in Table VIII.

IV. LEGAL VIEW ON CONCERNS IDENTIFIED IN TECHNICAL LITERATURE: THE EU REGULATORY FRAMEWORK

Against the backdrop of the concerns identified in the technical literature, the following analysis is geared to address two questions: How does current EU legislation protect the prosumer's data and privacy rights? How does the EU regulatory framework address the prosumer's concerns in the area of cybersecurity? For the first dimension, GDPR [45] and ED [15] are consulted; for the second, GDPR [45], ED [15], the NIS Directive (NIS) [2] and the Cybersecurity Act (CA) [46].

A. Data Protection and Privacy

Since SM help aggregate vast amounts of personal data of prosumers, data protection is a prevalent concern. As of 25 May 2018, GDPR governs the processing of an EU citizen's personal data. Potential personal data breaches by controllers or processors ensuing from the processing of a natural person's data can fall within the scope of GDPR [45].

This paper applies the definition of SM stipulated in ED, which establishes common rules for the EU internal market for electricity [15]. ED also includes the protection of prosumer rights and in the context of this paper is to be read together with GDPR [15], [45]. In this regard, a SM is defined as "an electronic system that is capable of measuring electricity fed into the grid or electricity consumed from the grid, providing more information than a conventional meter, and that is capable of transmitting and receiving data for information, monitoring and control purposes, using a form of electronic communication"[15].

Pursuant to Art. 4(1) of GDPR, prosumers in private households can be considered "natural persons", thus falling within the scope of "data subjects" [45]. In this case, any information processed by SM, which helps identify a natural person directly or indirectly by an identifier such as name, an identification number, location data, an online identifier or by other identifiers pertaining to the physical, psychological, genetic, mental, economic, cultural or social identity of that natural person, classifies as "personal data" [45].

ML generates profiles of prosumers. Without obtaining granular consent for the processing of personal data for "one or more specific purposes" in electronic communication or in form of an electronic or written contract from the data subject, GDPR renders processing of personal data generally illegal, except for situations allowed by law (Art. 6(1)(c-f); Art. 23(1)) [45]. The preconditions of receiving consent are stipulated in Art. 6(1), Art. 7 and Art. 12 [45]. Recital 32 clarifies consent as "a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject's agreement to the processing of personal data" [47]. Against this backdrop, the controller would be required to explain the prosumer in an electronic or written contract "using clear and plain language" for which purposes SM gather personal data and which measures are taken by the operator to safeguard the prosumer's rights in compliance with the GDPR [45].

GDPR Art. 5 is instrumental in understanding the key principles regarding the processing of personal data. Without respecting these principles, SM would infringe upon the prosumer's autonomy (for an overview of GDPR principles, see Table VI) [45].

GDPR makes a distinction between data controllers (Art. 4(7):"natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data") and data processors (Art. 4(8):"natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller") [45], where different obligations for each of these two actors are set out in Art. 24-43 (for an overview of the rights of the data subject and the obligations of the controller and processor, see Table VII). The multitude of actors involved in the design and operation of the smart grid system, however, complicates a clear identification of both data controller and data processor, thus posing challenges in terms of the attribution of duties and ensuing accountability requirements set out by GDPR [45] and ED [15] (for an overview of potential operators, consider [13], p. 9).

TABLE VI. DATA PROTECTION AND PRIVACY (I). GDPR: PRINCIPLES [45]

Principles	Article
Lawfulness, fairness and transparency	5(1)(a)
Purpose limitation	5(1)(b)
Data minimisation	5(1)(c)
Accuracy	5(1)(d)
Storage limitation	5(1)(e)
Integrity and confidentiality	5(1)(f)
Accountability	5(1)(g)

TABLE VII. GDPR: RIGHTS OF THE DATA SUBJECT AND OBLIGATIONS OF THE CONTROLLER AND PROCESSOR [45]

Rights	Article(s)
Transparent information, communication and modalities	12
Information and access to personal data	13;14;15
Rectification and Erasure	16;17;18;19;20
Right to object and automated individual decision-making	21;22
Obligations	**Article**
Responsibility of the Controller	24
Processor	28
Security of processing	32

GDPR Art. 22 ("Automated individual decision-making, including profiling") [45] presents a key prosumer right by obligating data controllers to implement measures that allow data subjects to intervene in automated decision-making procedures. In the context of this paper, this implies that a prosumer is granted the right to contest any automated decision facilitated by SM that entailed legal consequences for the data subject. However, the complexity of actors raises questions in terms of identifying and establishing accountability for GDPR breaches in cases such as denial of services, target to excessive advertisements or exhibition of prosumer habits and lifestyle. This could equally apply to scenarios in which e.g. electricity bills are sent out automatically to the prosumer based on potentially flawed data processed by SM, which result in e.g. price discrimination (for a legal view on all identified prosumer concerns, see Table VIII). It follows that national supervisory authorities play a central role in identifying operators and processors to be able to allocate their legal responsibilities in the smart grid.

B. Cybersecurity

According to GDPR Art. 5(1)(f), personal data must be processed in a manner which ensures appropriate security [45]. Here, security is mainly understood as the controller's duty to implement mechanisms which can appropriately mitigate a "personal data breach", more precisely "accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data transmitted, stored or otherwise processed" [45]. In the context of this paper, the term security refers to the security of personal data processed by SM in smart grids. Since this process takes place in the information and communication technology environment, the security of data would be generally governed by the framework of cybersecurity. Hence, NIS [2] and the latest adoption of CA [46] are instrumental in understanding how data security applies to SM. Consequently, cybersecurity forms one part of the understanding of security spelled out in ED, which refers to "security" as the "security of supply and provision of electricity and technical safety" [15].

Art. 2(1) of CA defines cybersecurity as "activities necessary to protect network and information systems, the users of such systems, and other persons affected by cyber threats'' [46]. SM can be considered network and information systems. This can be deduced from NIS Art. 4(1), which delineates the parameters of "network and information systems" [2]. A threat against i.a. SM is described as "any potential circumstance, event or action that could damage, disrupt or otherwise adversely impact" (Art. 2(8), CA) these systems [46].

Cyberthreats against network and information systems in energy systems can be mitigated provided operators/processors of personal data are able to secure "the ability of network and information systems to resist, at a given level of confidence, any action that compromises the availability, authenticity, integrity or confidentiality of stored or transmitted or processed data or the related services offered by, or accessible via, those network and information systems" (NIS Art. 4(2)) [2].

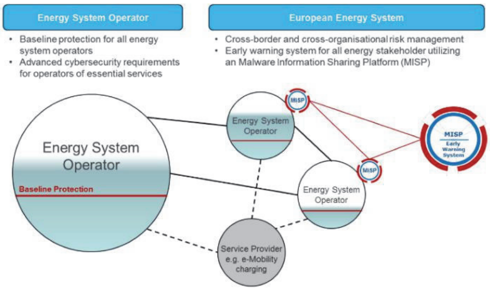
The terms availability, authenticity, integrity and confidentiality are initially derived from the concept of the "CIA Triad" [[definitions of C,I,A based on [48]],[49]]. Applying the general understanding of these terms individually to the operation of SM, operators/processors of data (i) are obliged to prevent disclosure of data to unauthorized third parties in this process (confidentiality) and (ii) to secure that the information contained in the data and gathered by SM is not altered in transit from the prosumer to the operator/processor, thus remaining authentic (integrity and authenticity) [2]. (iii) Additionally, according to NIS it is incumbent upon national authorities to establish mechanisms that can protect against e.g. distributed denial of service attacks conflicting i.a. with the principle of availability of data (availability) [2].

It is worthwhile mentioning that NIS creates mechanisms for the identification of operators of essential services (OES), which includes energy operators (NIS Directive, Art. 4(4), Art. 5(2), Annex 2) [2]. By the same token, NIS Art. 1(2)(e) obliges OES to inform a National Competent Authority (NAS) about potential cybersecurity incidents, broadly defined in NIS Art. 4(7) as "any event which has an actual adverse effect on the security of network and information systems" [2]. Establishing accountability for data breaches in SM remains problematic due to the great diversity of actors in the smart grid. Hence, the role of all relevant actors needs to be clearly identified and the list of actors continuously updated by NAS to understand for which actions and at what stages an operator/processor can incur responsibility for potential cybersecurity breaches outlined in Table VIII.

SGTF2 suggests the implementation of a Network Code on cybersecurity (c.f. Figure 3). It advocates for a minimum baseline protection [14]. In accordance with ISO/IEC 27001:2013, it would entail duties for operators to continuously adjust the cybersecurity mechanisms to be able to anticipate and identify cybersecurity threats against their infrastructure [14]. For this purpose, SGTF2 additionally recommends operators to utilize the EU cybersecurity certification scheme [14], [46].

V. CONCLUSIONS

When developing an application that makes use of NILM or operates at any surface of the AMI, cybersecurity and data protection and privacy needs to be considered, which can be done using the GDPR and following the CIA triad. This paper presents a tool in the form of a table (Table VIII) that can be used to identify key sections of the GDPR and the CIA triad in order to prioritize respective activities when developing or implementing technology. A sample workflow is presented in Figure 4 to provide an example for the use of the provided table.



- Baseline protection for all energy system operators
- Advanced cybersecurity requirements for operators of essential services

- Cross-border and cross-organisational risk management
- Early warning system for all energy stakeholder utilizing an Malware Information Sharing Platform (MISP)

Figure 3. SGTF network code on cybersecurity [14].

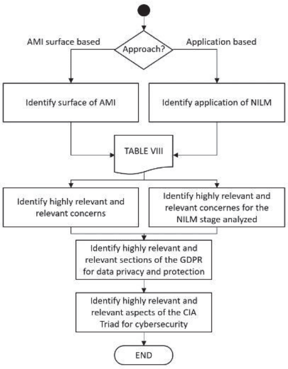

Approach?
AMI surface based
Application based
Identify surface of AMI
Identify application of NILM
TABLE VIII
Identify highly relevant and relevant concerns
Identify highly relevant and relevant concernes for the NILM stage analyzed
Identify highly relevant and relevant sections of the GDPR for data privacy and protection
Identify highly relevant and relevant aspects of the CIA Triad for cybersecurity
END

Figure 4. Workflow for using Table VIII to determine to filter more important sections of the GDPR and CIA triad for a specific implementation.

Implementations of ML methods for NILM rarely consider privacy aspects of prosumers. The identified prosumer concerns are relevant for all stages of the NILM process, considering possible implementations on different AMI surfaces, and depend on the proposed application in a HEMS, AAL, recommender systems or fault diagnostics context. Future research activities include the validation of the developed workflow and proposed mapping using real-life use-cases of ML applications in the electric smart grid.

VI. RECOMMENDATIONS

Although all surfaces of the AMI are relevant when addressing concerns of residential prosumers, some of them stand out. The HANs and SM are components of the AMI, which are highly relevant for all distinguished privacy and cyber-security concerns of residential prosumers. Utilities and manufacturers are encouraged to emphasize and promote cyber-security and privacy aspects of SM, while end-users are advised to secure their HANs by applying suitable measures and secure technologies. Utility companies are advised to provide insight into their HESs, since it is regarded as a component of the AMI, which is highly relevant in terms of end-user privacy.

Bearing the novelty of SM technology in mind, both the designers of SM and operators of the smart grid system are well advised to ponder how the principle of "data protection by design" underlying the GDPR framework can be fulfilled [45].

A Data Protection Impact Assessment (DPIA), laid down in GDPR Art. 35 [45], provides a suitable tool to address the prosumer concerns mapped in this paper. This mechanism makes it mandatory for operators to assess any data security, privacy or cybersecurity risk which is "likely to result in a high risk to the rights and freedoms of natural persons" [13], [45]. Ideally, this procedure is to be carried out prior to the wide-scale application of a new technology, which makes use of personal data. In general, a DPIA can be described as an accountability mechanism and "a process for building and demonstrating compliance" with GDPR [13]. This mechanism would help operationalize the policymakers' expectations towards SM for the benefit of the climate and the protection of the rights of the prosumer.

Additionally, the DPIA could be guided by the seven key requirements of the High-Level Expert Group on AI [50], which were reaffirmed in the EU Commission White Paper on AI [51]. While these requirements chime with the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [52], the authors recommend further research towards the operationalization of the seven key requirements, proposed by the High-Level Expert Group on AI, in electric smart grids.

TABLE VIII. MAPPING OF ML ANGLES VIA PROSUMER CONCERNS BASED ON RELEVANCE: TECHNICAL AND LEGAL VIEWS

			Prosumer Concerns											
			Price discrimination	Denial of prosumer services	Target to excessive advertisements	Identification of home appliances	Exhibition of user habits and lifestyle	Exhibition of illnesses and disabilities	Personification of anonymous data	Disconnection of home appliances	Burglary, arson, vandalism etc.	Attractive target to burglary	Target to kidnapping	Denial of personal mobility
Technical	Surfaces of AMI	Home Area Network	0	0	++	++	++	++	++	++	++	++	++	++
		Smart Meter	+	0	+	++	++	++	++	+	++	++	++	++
		Smart Meter Data Collector	0	0	++	+	+	+	+	+	+	+	+	+
		AMI Networks	0	0	+	+	+	+	+	+	+	+	+	+
		AMI Protocols	0	0	+	+	+	+	+	+	+	+	+	+
		Head-End Management System	++	++	++	++	++	++	++	+	+	+	+	+
	Applications of NILM	Home Energy Management System	α	α, ε	α, ε	γ	ε	0	α, ε	α, β, γ, δ, ε	δ, ε	γ, δ	δ, ε	α, β, γ, δ, ε
		Ambient Assisted Living	0			0		ε						
		Recommender System	α			γ		0						
		Fault Diagnostics	0			0		0						
Legal	Data Protection and Privacy (I)	GDPR Art. 5(1)(a)	++	++	++	++	++	++	++	+	+	+	+	+
		GDPR Art. 5(1)(b)	++	++	++	++	++	++	++	+	+	+	+	+
		GDPR Art. 5(1)(c)	++	++	++	++	++	++	++	+	+	+	+	+

		GDPR Art. 5(1)(d)	+	+	+	+	+	+	+	+	+	+	+	+
		GDPR Art. 5(1)(e)	+	+	+	+	+	+	+	+	+	+	+	+
		GDPR Art. 5(1)(f)	+	+	+	++	++	++	+	++	++	++	++	++
		GDPR Art. 5(1)(g)	++	++	++	++	++	++	++	++	++	++	++	++
	Data Protection and Privacy (II)	GDPR Art. 12	++	++	++	++	++	++	++	+	+	+	+	+
		GDPR Art. 13, 14, 15	++	++	++	++	++	++	++	+	+	+	+	+
		GDPR Art. 16, 17, 18, 19, 20	++	++	++	+	+	+	+	0	0	0	0	0
		GDPR Art. 21, 22	++	++	++	++	++	++	++	0	0	0	0	0
	Data Protection and Privacy (III)	GDPR Art. 24	++	++	++	++	++	++	++	++	++	++	++	++
		GDPR Art. 28	++	++	++	++	++	++	++	+	+	+	+	+
		GDPR Art. 32	+	+	+	++	++	++	+	++	++	++	++	++
	Cybersecurity: CIA Triad	Confidentiality	0	0	0	0	0	0	0	0	++	++	++	0
		Integrity/Authenticity	0	0	0	0	0	0	0	++	0	0	0	0
		Availability	0	0	0	0	0	0	0	0	0	0	0	++

++ = highly relevant; + = relevant; 0 = not relevant/applicable; α = Metering NILM stage; β = Event detection NILM stage; γ = Feature extraction NILM stage; δ = Classification NILM stage; ε = Analysis of classificationα NILM stage

ACKNOWLEDGMENTS

This work was supported by the European Commission through the H2020 project Finest Twins (grant No. 856602). Additionally, funding was provided by ASTRA "TTÜ arenguprogramm aastateks 2016-2022" (Doctoral School in Economics and Innovation Project code: 2014-2020.4.01.16-0032) as part of the European Regional Development Funding Scheme (EU programming period of 2014-2020).

REFERENCES

[1] Directorate-General for Energy (European Commission), ''Clean energy for all Europeans,'' 26.07.2019 [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/b4e46873-7528-11e9-9f05-01aa75ed71a1/language-en/format-PDF/source-126837758.

[2] Directive (EU) 2016/1148, OJ L 194, 19.7.2016, p. 1–30.

[3] T. Häring, A. Rosin, and H. Biechl, "Using common household thermal storages to support the PV- and battery system in nearly zero energy buildings in off-grid mode," Sustain. Energy Technol. Assessments, vol. 35, no. May, pp. 12–24, Oct. 2019, doi: 10.1016/j.seta.2019.05.014.

[4] T. Korotko, A. Rosin, and R. Ahmadiahangar, "Development of prosumer logical structure and object modeling," Apr. 2019, doi: 10.1109/CPE.2019.8862390.

[5] R. Ahmadiahangar, T. Häring, A. Rosin, T. Korõtko, and J. Martins, "Residential Load Forecasting for Flexibility Prediction Using Machine Learning-Based Regression Model," in Proceedings - 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe, EEEIC/I and CPS Europe 2019, Jun. 2019, pp. 1–4, doi: 10.1109/EEEIC.2019.8783634.

[6] S. Mishra, H. Koduvere, I. Palu, R. Kuhi-Thalfeldt, and A. Rosin, "Assessing demand side flexibility with renewable energy resources," in 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC), Jun. 2016, pp. 1–6, doi: 10.1109/EEEIC.2016.7555546.

[7] M. R. Asghar, G. Dán, D. Miorandi, and I. Chlamtac, "Smart meter data privacy: A survey," IEEE Commun. Surv. Tutorials, vol. 19, no. 4, pp. 2820–2835, Jun. 2017, doi: 10.1109/COMST.2017.2720195.

[8] H. He and J. Yan, "Cyber-physical attacks and defences in the smart grid: a survey," IET Cyber-Physical Syst. Theory Appl., vol. 1, no. 1, pp. 13–27, Dec. 2016, doi: 10.1049/iet-cps.2016.0019.

[9] P. Kumar, Y. Lin, G. Bai, A. Paverd, J. S. Dong, and A. Martin, "Smart Grid Metering Networks: A Survey on Security, Privacy and Open Research Issues," IEEE Commun. Surv. Tutorials, vol. 21, no. 3, pp. 2886–2927, Jul. 2019, doi: 10.1109/COMST.2019.2899354.

[10] M. A. Mengistu, A. A. Girmay, C. Camarda, A. Acquaviva, and E. Patti, "A Cloud-Based On-Line Disaggregation Algorithm for Home Appliance Loads," IEEE Trans. Smart Grid, vol. 10, no. 3, pp. 3430–3439, May 2019, doi: 10.1109/TSG.2018.2826844.

[11] S. Hosseini, N. Henao, S. Kelouwani, K. Agbossou, and A. Cardenas, "A Study on Markovian and Deep Learning Based Architectures for Household Appliance-level Load Modeling and Recognition," in 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE), Jun. 2019, vol. 2019-June, pp. 35–40, doi: 10.1109/ISIE.2019.8781186.

[12] G. Tang, Z. Ling, F. Li, D. Tang, and J. Tang, "Occupancy-aided energy disaggregation," Comput. Networks, vol. 117, pp. 42–51, Apr. 2017, doi: 10.1016/j.comnet.2016.11.019.

[13] Smart Grid Task Force, ''Expert Group 2: Regulatory Recommendations for Privacy, Data Protection and Cyber-Security in the Smart Grid Environment – Data Protection Impact Assessment Template for Smart Grid and Smart Metering systems,'' v.2 of 13 September 2018. [Online]. Available at: https://ec.europa.eu/energy/sites/ener/files/documents/dpia_for_publication_2018.pdf.

[14] Smart Grid Task Force, ''Expert Group 2: Recommendations to the European Commission for the Implementation of Sector-Specific Rules for Cybersecurity Aspects of Cross-Border Electricity Flows, on Common Minimum Requirements, Planning, Monitoring, Reporting and Crisis Management,'' June 2019. [Online]. Available at: https://ec.europa.eu/energy/sites/ener/files/sgtf_eg2_report_final_report_2019.pdf.

[15] Directive (EU) 2019/944, OJ L 158, 14.6.2019, p. 125–199.

[16] F. Al-Turjman and M. Abujubbeh, "IoT-enabled smart grid via SM: An overview," Futur. Gener. Comput. Syst., vol. 96, pp. 579–590, Jul. 2019, doi: 10.1016/j.future.2019.02.012.

[17] M. Wigan, "User issues for smart meter technology," IEEE Technol. Soc. Mag., vol. 33, no. 1, pp. 49–53, Mar. 2014, doi: 10.1109/MTS.2014.2301856.

[18] D. L. S. Mendes, R. A. L. Rabelo, A. F. S. Veloso, J. J. P. C. Rodrigues, and J. V. dos Reis Junior, "An adaptive data compression mechanism for smart meters considering a demand side management scenario," J. Clean. Prod., vol. 255, May 2020, doi: 10.1016/j.jclepro.2020.120190.

[19] J. E. Rubio, C. Alcaraz, and J. Lopez, "Recommender system for privacy-preserving solutions in smart metering," Pervasive Mob. Comput., vol. 41, pp. 205–218, Oct. 2017, doi: 10.1016/j.pmcj.2017.03.008.

[20] L. Wei, L. P. Rondon, A. Moghadasi, and A. I. Sarwat, "Review of Cyber-Physical Attacks and Counter Defense Mechanisms for Advanced Metering Infrastructure in Smart Grid," Proc. IEEE Power

Eng. Soc. Transm. Distrib. Conf., vol. 2018-April, May 2018, Accessed: Apr. 22, 2020. [Online]. Available: http://arxiv.org/abs/1805.07422.

[21] CEN-CENELEC-ETSI, “Functional architecture for communications in smart metering systems,” pp. 1–70, 2011, [Online]. Available: ftp://ftp.cen.eu/cen/Sectors/List/Measurement/Smartmeters/CENCLCETSI_TR50572.pdf.

[22] D. Jacobson and L. Dickerman, “Distributed intelligence: A critical piece of the microgrid puzzle,” Electr. J., vol. 32, no. 5, pp. 10–13, Jun. 2019, doi: 10.1016/j.tej.2019.05.001.

[23] S. Tonyali, K. Akkaya, N. Saputro, A. S. Uluagac, and M. Nojoumian, “Privacy-preserving protocols for secure and reliable data aggregation in IoT-enabled Smart Metering systems,” Futur. Gener. Comput. Syst., vol. 78, pp. 547–557, Jan. 2018, doi: 10.1016/j.future.2017.04.031.

[24] J. Foreman and D. Gurugubelli, “Cyber Attack Surface Analysis of Advanced Metering Infrastructure,” 2016.

[25] M. Bae, K. Kim, and H. Kim, “Preserving privacy and efficiency in data communication and aggregation for AMI network,” J. Netw. Comput. Appl., vol. 59, pp. 333–344, Jan. 2016, doi: 10.1016/j.jnca.2015.07.005.

[26] N. Fadhel, F. Lombardi, L. Aniello, A. Margheri, and V. Sassone, “Towards a semantic modelling for threat analysis of IoT applications: A case study on transactive energy,” in IET Conference Publications, 2019, vol. 2019, no. CP756, doi: 10.1049/cp.2019.0147.

[27] G. Giaconi, D. Gunduz, and H. V. Poor, “Smart Meter Privacy with Renewable Energy and an Energy Storage Device,” in IEEE Transactions on Information Forensics and Security, Jan. 2018, vol. 13, no. 1, pp. 129–142, doi: 10.1109/TIFS.2017.2744601.

[28] M. S. Piscitelli, S. Brandi, and A. Capozzoli, “Recognition and classification of typical load profiles in buildings with non-intrusive learning approach,” Appl. Energy, vol. 255, p. 113727, Dec. 2019, doi: 10.1016/j.apenergy.2019.113727.

[29] M. Wigan, “User issues for smart meter technology,” IEEE Technol. Soc. Mag., vol. 33, no. 1, pp. 49–53, Mar. 2014, doi: 10.1109/MTS.2014.2301856.

[30] S. Yussof, M. E. Rusli, Y. Yusoff, R. Ismail, and A. A. Ghapar, “Financial impacts of smart meter security and privacy breach,” Conf. Proc. - 6th Int. Conf. Inf. Technol. Multimed. UNITEN Cultiv. Creat. Enabling Technol. Through Internet Things, ICIMU 2014, pp. 11–14, 2015, doi: 10.1109/ICIMU.2014.7066595.

[31] L. Pereira and N. Nunes, “Performance evaluation in non‐intrusive load monitoring: Datasets, metrics, and tools—A review,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 6, Nov. 2018, doi: 10.1002/widm.1265.

[32] A. Miyasawa, Y. Fujimoto, and Y. Hayashi, “Energy disaggregation based on smart metering data via semi-binary nonnegative matrix factorization,” Energy Build., vol. 183, pp. 547–558, Jan. 2019, doi: 10.1016/j.enbuild.2018.10.030.

[33] G. A. Raiker, B. Subba Reddy, L. Umanand, A. Yadav, and M. M. Shaikh, “Approach to Non-Intrusive Load Monitoring using Factorial Hidden Markov Model,” in 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), Dec. 2018, pp. 381–386, doi: 10.1109/ICIINFS.2018.8721436.

[34] T. Bernard, M. Verbunt, G. Vom Bogel, and T. Wellmann, “Non-Intrusive Load Monitoring (NILM): Unsupervised Machine Learning and Feature Fusion : Energy Management for Private and Industrial Applications,” in 2018 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE), May 2018, pp. 174–180, doi: 10.1109/ICSGCE.2018.8556735.

[35] M. Devlin and B. P. Hayes, “Non-Intrusive Load Monitoring using Electricity Smart Meter Data: A Deep Learning Approach,” in 2019 IEEE Power & Energy Society General Meeting (PESGM), Aug. 2019, vol. 2019-Augus, pp. 1–5, doi: 10.1109/PESGM40551.2019.8973732.

[36] M. Aiad and P. H. Lee, “Unsupervised approach for load disaggregation with devices interactions,” Energy Build., vol. 116, pp. 96–103, Mar. 2016, doi: 10.1016/j.enbuild.2015.12.043.

[37] A. Ruano, A. Hernandez, J. Ureña, M. Ruano, and J. Garcia, “NILM techniques for intelligent home energy management and ambient assisted living: A review,” Energies, vol. 12, no. 11. MDPI AG, Jun. 10, 2019, doi: 10.3390/en12112203.

[38] J. Cho, Z. Hu, and M. Sartipi, “Non-Intrusive A/C Load Disaggregation Using Deep Learning,” in 2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), Apr. 2018, vol. 2018-April, pp. 1–5, doi: 10.1109/TDC.2018.8440358.

[39] A. U. Rehman, T. Tjing Lie, B. Valles, and S. R. Tito, “Low Complexity Non-Intrusive Load Disaggregation of Air Conditioning Unit and Electric Vehicle Charging,” in 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), May 2019, pp. 2607–2612, doi: 10.1109/ISGT-Asia.2019.8881113.

[40] S. Singh and A. Majumdar, “Deep Sparse Coding for Non–Intrusive Load Monitoring,” IEEE Trans. Smart Grid, vol. 9, no. 5, pp. 4669–4678, Sep. 2018, doi: 10.1109/TSG.2017.2666220.

[41] T.-T.-H. Le, J. Kim, and H. Kim, “Classification performance using gated recurrent unit recurrent neural network on energy disaggregation,” in 2016 International Conference on Machine Learning and Cybernetics (ICMLC), Jul. 2016, vol. 1, pp. 105–110, doi: 10.1109/ICMLC.2016.7860885.

[42] P. A. Schirmer, I. Mporas, and M. Paraskevas, “Evaluation of Regression Algorithms and Features on the Energy Disaggregation Task,” in 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Jul. 2019, pp. 1–4, doi: 10.1109/IISA.2019.8900695.

[43] A. Boulemtafes, A. Derhab, and Y. Challal, “A review of privacy-preserving techniques for deep learning,” Neurocomputing, vol. 384, pp. 21–45, Apr. 2020, doi: 10.1016/j.neucom.2019.11.041.

[44] R. G. Rajasekaran, S. Manikandaraj, and R. Kamaleshwar, “Implementation of Machine Learning Algorithm for predicting user behavior and smart energy management,” in 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), Feb. 2017, pp. 24–30, doi: 10.1109/ICDMAI.2017.8073480.

[45] Regulation (EU) 2016/679, OJ L 119, 4.5.2016, p. 1–88.

[46] Regulation (EU) 2019/881, OJ L 151, 7.6.2019, p. 15–69.

[47] Intersoft Consulting. ''Recital 32: Conditions for Consent.'' gdpr-info.eu. https://gdpr-info.eu/recitals/no-32/ (accessed January 14, 2020).

[48] A. Agarwal and A. Agarwal, “The Security Risks Associated with Cloud Computing,” INT'L J. Comput. Appl. Eng. SCI, pp. 257–258, 2011, Accessed: May 06, 2020. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.207.9119.

[49] A. Kasper and A. Antonov, “Towards Conceptualizing EU Cybersecurity Law,” in ZEI Discussion Paper Series, C 253 (2019), Accessed: May 14, 2020. [Online]. Available: http://www.zei.uni-bonn.de/aktuelles/2019/zei-discussion-paper-c-253-2019

[50] A. Antonov and T. Kerikmäe, “Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU,” in The EU in the 21st Century, Springer International Publishing, 2020, pp. 135–154.

[51] European Commission, ''White Paper on Artificial Intelligence: A European approach to excellence and trust,'' EU, Brussels, Belgium, Rep. COM(2020) 65 final, 19.2.2020. Accessed: 14 May 2020. [Online]. Available: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[52] D. L. S. Mendes, R. A. L. Rabelo, A. F. S. Veloso, J. J. P. C. Rodrigues, and J. V. dos Reis Junior, “An adaptive data compression mechanism for smart meters considering a demand side management scenario,” J. Clean. Prod., vol. 255, May 2020, doi: 10.1016/j.jclepro.2020.120190.

Publication III

1.1 Antonov, A. (2022). Managing Complexity: The EU's Contribution to Artificial Intelligence Governance. Revista CIDOB d'Afers Internacionals, (131), 41–65. https://doi.org/10.24241/rcai.2022.131.2.41/en

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).

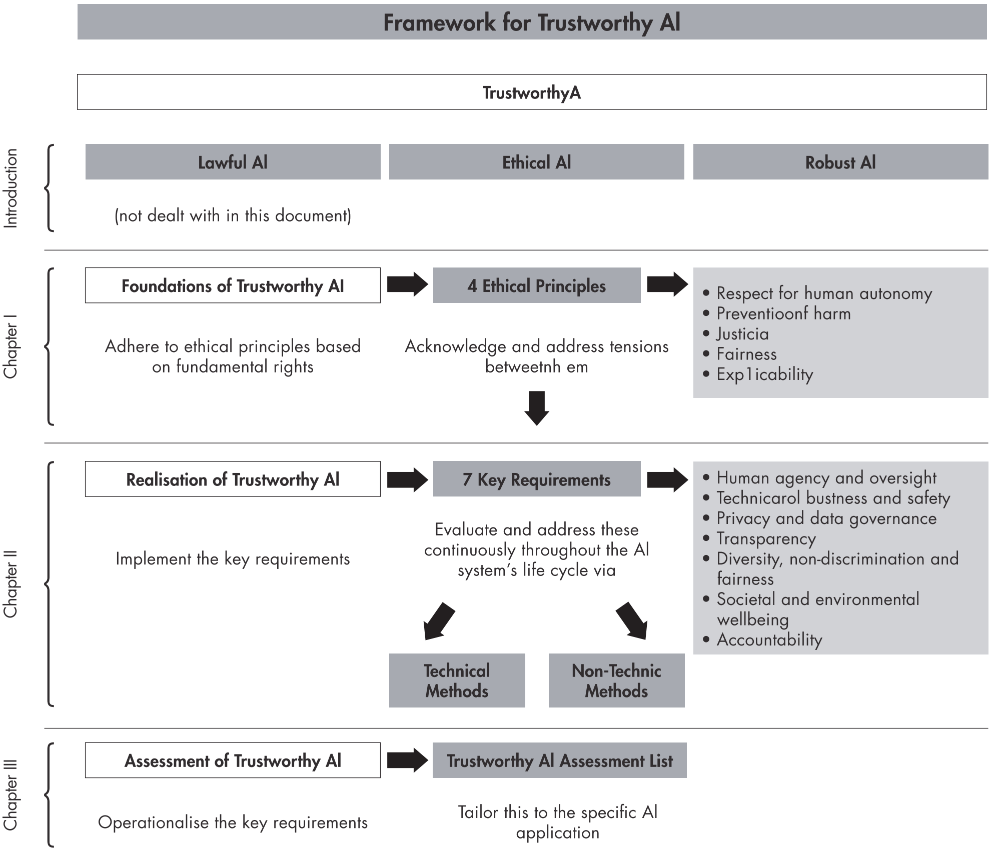
Ethics guidelines on AI, laying the groundwork for the concept of human-centric, Trustworthy AI, and setting a global standard; (ii) *White Paper on AI*, setting out the vision for an AI ecosystem based on trust and excellence; and (iii) *Legislative Proposals*, in particular the Digital Services Act, 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).

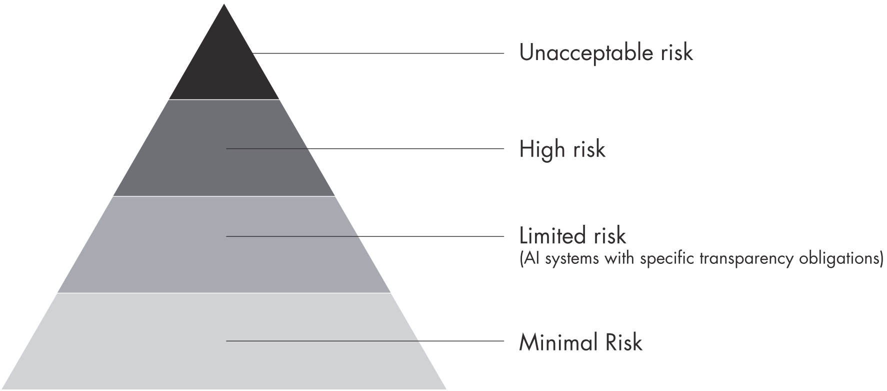
Figure 1. Framework for trustworthy AI of the High-Level Expert Group on AI (AI HLEG)


Framework for Trustworthy AI
TrustworthyA
Introduction
Lawful AI
Ethical AI
Robust AI
(not dealt with in this document)
Chapter I
Foundations of Trustworthy AI
4 Ethical Principles
• Respect for human autonomy
• Preventioonf harm
• Justicia
• Fairness
• Exp1icability
Adhere to ethical principles based on fundamental rights
Acknowledge and address tensions betweetnh em
Chapter II
Realisation of Trustworthy AI
7 Key Requirements
• Human agency and oversight
• Technicarol bustness and safety
• Privacy and data governance
• Transparency
• Diversity, non-discrimination and fairness
• Societal and environmental wellbeing
• Accountability
Implement the key requirements
Evaluate and address these continuously throughout the AI system's life cycle via
Technical Methods
Non-Technic Methods
Chapter III
Assessment of Trustworthy AI
Trustworthy AI Assessment List
Operationalise the key requirements
Tailor this to the specific AI application

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.

Publication IV

2.5 Kerikmäe, T., Nyman Metcalf, K., Hoffmann, T., Minn, M., Liiv, I., Taveter, K., Shumilo, O., Solarte, C. & **Antonov, A.** (2019). 1st Report on Legal Framework and Analysis Related to Autonomous Intelligent Technologies. In Riigikantselei (pp. 1−11).

TAL
TECH

1st REPORT ON LEGAL FRAMEWORK AND ANALYSIS RELATED TO AUTONOMOUS INTELLIGENT TECHNOLOGIES[1]

By TalTech expert group: Tanel Kerikmäe, Katrin Nyman Metcalf, Thomas Hoffmann, Mari Minn, Innar Liiv, Kuldar Taveter, (including doctoral students Olga Shumilo, Maria Claudia Solarte and Alexander Antonov).

According to the agreement, the preliminary analysis has been conducted, directions and perspectives have been established and relevant literature/material collected and archived. The outcome is the conclusion to be discussed with the Client/Government Office on 08.01.2019.

NB. As the annexes are rather voluminous, they will be provided, if necessary, to a platform (dropbox) available for Government Office.

NB.2 As agreed, the current Report will be presented in English due to several English speaking contributors and terminology. The next reports will be presented in Estonian language and the current report (without annexes) will be summarized in Estonian as well.

SYNOPSIS

[1] This report is composed with financial support of ETAG funded project: PUT1628 (PUT1628) "Eesti e-valitsemise teenuste riikliku süsteemi analüüs EL-i digitaalse ühtse turu õigusliku raamistiku kontekstis (1.01.2017−31.12.2018)", Tanel Kerikmäe, Tallinna Tehnikaülikool, Majandusteaduskond, Õiguse instituut . and by the Government Office Project: Lep18096 "Autonoomsete intelligentsete tehnoloogiate õigusliku reguleerimise ettepanekute väljatöötamine (15.12.2018−30.04.2019)", Tanel Kerikmäe, Tallinna Tehnikaülikool, Majandusteaduskond, Õiguse instituut .

Main:

- One central aim of regulation should be monitor systems and react against risks and filter the AI usage through data protection (especially 'automatic data processing');
- fragmentation within the EU through divergent national rules would be replaced by purpose-oriented regulation, not with specific rules.

Liability issues:

- the concept of strict liability is widely applied in many legal systems, incl the Estonian legal system (liability for damage caused by major source of danger, product liability).
- Approaches: non-contractual and contractual liability / allocating the risk between the designer, manufacturer and user of the respective autonomous system which caused the harmful act/did not perform correctly any other obligations/ essential question whether humans must be able to (manually) regulate and control autonomous intelligent technologies;
- Aspect of delegation of powers: private entities will have responsibility for increased numbers of public tasks, including usage of AI.

Proposals to be considered:

- to establish a registry of robots operating in the physical world;
- new competences of Technical Regulatory Authority (Tehnilise Järelevalve Amet);
- establishing a national insurance fund for autonomous intelligent technologies;
- impact assessment model/standard should be made for AI (benefits and risks risk);
- 6 categories of Estonian legal acts should be analysed taking account the best practice of EU Member States and leading Global actors in the field. Special attention: machine readability of legal acts and legal decision making + granularity of access control and responsibility.

I THEORETICAL POINTS OF VIEW

1. Current research on legal and regulatory governance of artificial intelligence (AI) systems focuses on **whether existing legal systems are able to handle the consequences of further implementation of AI systems** (Annex 1: list of sources composed by Kuldar Taveter and MC Solarte). The main issue here is achieving a better understanding of the distribution of responsibilities and liabilities between humans and their use of AI systems, considering the plethora of interacting technology (e.g. chatbots) developers, cloud providers, hardware manufacturers, telco and network operators, and application and platform service providers.

2. One of the prevailing approaches so far has been the **"hard-wiring" of norms and values into socio-technical systems** (Yeung, 2015). This includes but is not limited to compliance, security, data protection and privacy by design, which are concerned with the implementation of AI techniques for differential privacy, and with algorithmic governance through formal analysis and representation of regulatory principles, allocating rights, distributing liability, and managing legal identity. Others have pointed out that law and privacy cannot be fully hardcoded (Koops & Leenes, 2014), and may produce "legal by design" (LbD) (Lippe, Katz & Jackson, 2015) rather than "legal protection by design" (LPbD) (Hildebrandt 2017). According to Pagallo (2013), one can image a strict liability set-up being applicable for putting AI systems onto the market that cause damage or harm, with the possibility to refute liability in case of a certification or evidence of security or safety by design. Consequently, LbD can, if adopted, simply overrule human agency by unwarranted implementation of legal decision systems that force people into compliance. On the other hand, LPbD highlights the need to build transparency and contestability into AI infrastructures, as required, for instance, by the new EU General Data Protection Regulation (GDPR)[2]. LPbD is consistent with the core data protection principles of fairness, transparency, accountability and responsibility required by the GDPR. LpbD is challenged by the principle that conformity with normative values can always be better addressed through "legal technology" applications that filter out human agency, such as blockchain technology, smart contracts, agreement technologies, and cryptocurrencies. It is also in the nature of algorithms that they "find their own way" and thus are not transparent. The more intelligent artificial intelligence gets, the more this will be noticeable. Therefore, as it has been recently pointed out by Brownsword (2016), Yeung (2018), and Hildebrandt (2017), **further research is required into the relationship and interaction between management of liabilities by "legal technology" and LPbD.**

3. In the Estonian context, it seems that the legislator is about to rush into adopting the "legal by design" approach by introducing legal personhood of software agents and robots.[3] However, this requires further thinking and analysis and discussions between lawyers and computer scientists, if the step at all is decided to be a viable option (on serious doubts about that see III). There is a danger with approving certain technologies as in accordance with legal norms in that this may lead to such specific technologies being favoured, even when new technologies might have been created

[2] European Commission (2018). 2018 reform of EU data protection rules. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en, last accessed on 20.12.2018.

[3] Triniti (2017). Analüüs SAE tase 4 ja 5 sõidukite kasutusele võtmiseks koos seaduseelnõu väljatöötamiskavatsuse kirjeldustega. http://www.ituudised.ee/uudised/2017/09/25/isejuhtivate-soidukite-oigusanaluus-tostatab-robootikaseaduse-loomise-vajaduse, last accessed 20.12.2018.

that could be better. Focusing on the protection provided rather than the exact technology (a purpose-oriented approach) is preferable, even if it is admitted that this may be more complicated. **One central aim of regulation should be on introducing the responsibility to "police" and monitor systems and react against risks. Other legal areas that require further research, including for Estonia, concern data protection, non-discrimination, presumption of innocence and privacy, requiring meaningful transparency and explanation regarding AI applications** (Hildebrandt, 2011). An important venue for such discussions will be the Workshop on Responsible Artificial Intelligence Agent[4], which has been accepted to take place in conjunction with the 2019 International Conference on Autonomous Agents and Multi-agent Systems.

II TECHNOLOGICAL POINTS OF VIEW

4. The Client explicitly specified in the 5 December 2018 meeting that **any topics of ethics[5] should not be the core of this work**, as the Government Office plans to focus on ethics in cooperation with 'international organizations' and the EU initiative (High Level Expert Group on AI/Ethical guidelines 18.12.2018), which needs to be discussed on 22 January 2019 by Government Office representative first;

5. One should clarify beforehand whether only 'autonomous intelligent technologies' **having a physical form in an unrestricted physical world is the scope of this research** or whether all algorithms should be taken into consideration. So far, this research follows the first approach, as otherwise it could just as well be discussed how e.g. Google's search engine as 'autonomous intelligent technology' should be regulated when operating in our jurisdiction.

6. It can also be considered (although there are several thoughts by members of the expert group) to **establish a registry of robots operating in the physical world**, whereas the option of granting robots a full scale legal personality which could 'take responsibility'[6] itself has not found much support in the group.[7]

[4] AAMAS 2019, Accepted Workshops, http://aamas2019.encs.concordia.ca/accws.html, last accessed 20.12.2018.

[5] Probably one of the most influential thinkers in this field currently – Luciano Floridi - https://digitalethicslab.oii.ox.ac.uk/luciano-floridi/

[6] Gurney, J. K. (2013). Sue my car not me: Products liability and accidents involving autonomous vehicles. *U. Ill. JL Tech. & Pol'y*, 247.

[7] Douma, F., & Palodichuk, S. A. (2012). Criminal liability issues created by autonomous vehicles. *Santa Clara L. Rev.*, *52*, 1157.

Potentially the registry of robots should include:

a. The manufacturer of the robot;
b. The current legal owner of the robot;
c. The current user of the robot;
d. The original firmware/software installed in that robot by the manufacturer;
e. The current firmware/software running in that robot;
f. The original configuration (permissions, rules (~ "values")) set by the manufacturer;
g. The current configuration (permission, rules (~ "values")) in that robot, and
h. The main location of the robot (centre of main acitivities).

7. Potentially, **the black box of robots** should include in minimum:
    a. Log allowing the replay of granular aspects of context (up to some interval – a day, week etc),
    b. input (audio, video, others),
    c. technical decisions and output (actuators).

8. Micropayments should be included in the regulation as a separate kind of transaction in lieu driver/owner/user.

9. The question is, how such requirements – or any requirements – can be enforced. The easiest is to reply on existing regulatory structures (or a modification of these) rather than creating something entirely new. This would mean the **Technical Regulatory Authority** (Tehnilise Järelevalve Amet). The benefit of using an existing body is not just to save time, money and work with creating something new, but also as both they and the companies they regulate have experience of what it means when special demands are made on private firms by a specific authority – which would also be the case for such requirements as mentioned here. They also have experience of keeping records when firms have an obligation of registering (like in the communications field).

10. **A national <u>insurance</u> fund for autonomous intelligent technologies**[8] can be considered with the primary function of coordinating, aggregating and storing statistics of negative events, which allows **<u>private insurance companies</u>** to step in, scale up and develop relevant and agile products.

11. An essential decision will have to be taken **whether humans must be able to (manually) regulate and control autonomous intelligent technologies** or whether a

[8] Schroll, C. (2014). Splitting the bill: creating a national car insurance fund to pay for accidents in autonomous vehicles. *Nw. UL Rev.*, *109*, 803.

'regulatory technology' (RegTech) or 'supervisory technology' (SupTech)[9][10] may be necessary for this task. For instance, regulators in financial supervision are seriously discussing algorithms constantly monitoring the compliance (of other algorithms). In practice, this would require manufacturers invest into algorithms checking the compliance of their products (before releasing them to unrestricted urban space).

12. References (or established alignments) should be made with present regulations about 'automatic decision' (credit risk) and other 'automatic data processing' [of personal data]. Machine readability of legal acts and legal decision making is important and a recurring theme from many aspects.[11]

13. The question to be answered: What should be the anticipated future interplay between standards and legal acts? For better framing the output, **perhaps SAE J3016 'Automation Levels' should be used and regulated differently?**

III LEGAL POINTS OF VIEW

14. Current AI Policies in the EU (ECOSOC), the Council of Europe and selected EU Member States have been analyzed (**Finland, Sweden, Germany, Denmark, France** (Annex 2, 3), just as well as recent legal acts and legal theory in **United States of America** (Annex 4)[12]. The analysis includes more than hundred pages of best practices that will be taken into account in elaborating suggestions to Estonia including common elements deriving from comparative study such as established ecosystems, conducted research, governmental initiatives, funding, main challenges and competitive sectors.

15. **One possible analogy to consider, even if it does not answer all questions is with the legal situation of domestic animals.** Animals are seen in the legislation of many countries as something more than just “things” – inanimate objects. They do not however have legal personality of their own and the responsibility for their actions is held by someone else with a supervisory duty – a human with a certain relationship to the animal (owner, handler). In recent years there have globally been several cases where it has been suggested to give animals legal standing and/or more rights than

[9] Toronto Center (2017) FinTech, Regtech and SupTech: What They Mean for Financial Supervision

[10] Dias, D & Staschen, S (2017) Regtech and Digital Finance Supervision: A Leap into the Future

[11] The best research framing the area is: Daniel Ben-Ari, Yael Frish, Adam Lazovski, Uriel Eldan, & Dov Greenbaum, "Danger, Will Robinson"? Artificial Intelligence in the Practice Of Law: An Analysis and Proof of Concept Experiment, 23 Rich. J.L. & Tech. See also: Kerikmäe, T.; Hoffmann, T.; Chochia, A. (2018). Legal Technology for Law Firms: Determining Roadmaps for Innovation. Croatian International Relations Review, 24 (81), 91−112; Kerikmäe, T; Särav, S. (2017). Paradigms for Automatization of Logic and Legal Reasoning. In: Krimphove, D.; Lentner, G. M. (Ed.). Law and Logic: Contemporary Issues (205−222). Duncker & Humblot.

[12] See also conclusive source on AI strategies in the world: https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

just rights not to be abused under animal protection legislation. Although most such cases have reached the conclusion that such rights cannot be given, the debate is ongoing, and commentators have suggested similarities with robots: beings that act to a certain extent independently but are not capable of possessing full rights and legal personality.

16. Although it is in the nature of law to focus on threats and risks, in order to find ways to avoid these, it is at the same time important when discussing a legal framework for AI to not overly focus on the threats. The starting point should be on the opportunity: what benefit does the new technology offer? When this is well understood, it can be determined what the risks are and how they can be avoided. This approach reduces two different risks. First, that so many threats are seen or imagined, that development of the new technology will not be allowed to proceed. Secondly, that technologies are developed that do not have any practical use or benefit. **There is long experience in rule-making of environmental (and to some extent social) impact assessments. Such impact assessments should be made for AI and it should be an *impact* assessment, meaning benefits as well as risk seen as a package.** Existing legislation on use of robots like health and safety legislation relating to industrial robots should not be forgotten as a source of inspiration, as in many instances introduction of AI is a change in degree rather than fundamentally (robots get gradually more intelligent but they are already among us).

17. One central ethical and philosophical point to be kept in mind when legislating or creating regulatory systems is that the question of motivation has to be seen in a completely different way. AI does not act because of a specific reason that may help determine its actions and thus also help to predict future actions, provide mitigating circumstances, show what measures may be successful in preventing future harmful actions and so on. **Any legislation/regulation of the AI as such (rather than the persons controlling it) thus cannot contain any allusion to motives and reasons.**

18. **Another legal specificity with AI (potentially) is that there may be more cases of companies and organisations outsourcing features of their process.** Thus, important decisions may be taken somewhere else than in the organisation (including public body) that carries out the actual task of the organisation. Private entities will have responsibility for increased numbers of public tasks. Inspiration may be found from liberalisation and privatisation of telecommunications starting from the 1980s, when private firms took over more and more central elements for carrying out public power.

19. **The relationship between Estonian law making and EU law making** is challenging but it is very important to find a correct balance. On the one hand, the EU moves slowly and will most likely not create comprehensive regulation but more likely a regulatory

framework (analogous to the situation for telecommunications), but on the other hand, **fragmentation within the EU through divergent national rules would be negative. This dilemma can be dealt with if regulation is purpose-oriented rather than too specific** (which also has benefits from other perspectives, as mentioned elsewhere).

20. To protect against risks of hacking, which can cause enormous damage for internet-of-things (autonomous technologies) **granularity of access control can be used, meaning that there are many different levels at which access is given and responsibility exists for e.g. encryption**. This should be reflected also in the legal responsibility.

21. The expert group is of the opinion that **the current and potential future uses of autonomous intelligent systems do not mandate a need to change the fundamentals of legal regulation**, which include the following ideas and principles:

a) only humans are capable *to act* within the meaning of legal realm;
b) only *human activity* can incur legal liability;
c) consequently, the *subjects* of any legal obligation can ultimately be only humans.

22. The **employment of an autonomous intelligent system is different from other human activities only in the *way* that the act occurs** (and the way that the act conveys the intent of the human individual performing the action). Basically, all autonomous intelligent systems can be described as *tools* within the context of regulation.

23. **Complex and powerful tools have a greater potential for harm**, including harm that is unintentional, i.e. harm caused by situations in which the person wielding the tool does not wish to cause harm, or even actively tries to prevent any harm from happening. Because of the inherent *risk* of using such tools, the concept of strict liability is widely applied in many legal systems, incl the Estonian legal system (**liability for damage caused by major source of danger**[13], **product liability**[14]). However, autonomous systems with learning algorithms are increasingly operating in many fields where so far actions have been based exclusively on human decisions and – if the respective act turned out to be wrongful or constitutes the breach of a contractual duty– on human fault. Even though legal systems worldwide differ considerably in terms of fault in its technical sense as necessary criteria for liability (strict liability as core concept in common law, fault-based liability in most civil law systems), fault in its more extensive meaning as human error is a key cause for liability. For instance, while at present 90 percent of all traffic accidents are currently due to human error,[15] the more autonomous systems will be used in traffic, the more liability cases will be caused rather by system errors where the autonomous system did not calculate correctly the actions triggered, usually due to an imperfect design of its algorithms and/or technical failure of the hardware

VÕS § 1056
VÕS § 1061
[15] https://ec.europa.eu/transport/themes/its/road_en

(even though the overall number of accidents will be only a fraction of today's figures). In contractual liability, the non-performance will be caused by such reasons alike, depriving the human on behalf the system acts of any proper contribution in the causation chain beyond the fact that he initiated the use of the system for these purposes in the first place.

Approaches to handle liability legally:

a) As in both **non-contractual and contractual liability scenarios** autonomous systems replace human legally relevant action comprehensively, it has been argued that this novel scope of autonomous acts should be legally recognized as legal acts of their own nature as well, to be more specific by a recognition of the (partial) legal capacity and thus a personal liability of the artificial intelligence.

   Still, few supporters of this approach seem to be aware how essentially alien such a solution would be within our existing legal order – and how little need there in fact is for such a ground-breaking extension. At present, legal capacity is only granted to natural persons, legal persons and certain associations of individuals, all of them represented by natural persons. The recognition of (partial) legal capacity of autonomous systems could hardly dogmatically or practically be implemented into our system of private law: For instance, autonomous systems would have to be provided with proper assets to make them able to cover eventual damages, which raises various questions starting from the concept of a "start-up capital", the amount of such assets, its origin, liability in case that the capital does not suffice, and – obviously – the procedural rights and its practice of sued autonomous systems as well as enforcement. Besides – just to think this thought to the end from a procedural approach as well - from a legal point of view an autonomous system could always refer to contributory negligence or even intent by their respective creators who failed to design them sufficiently or to match the tasks the claimant claims them to have failed at (or of the user who ordered them to perform certain tasks towards the claimant), which would simply finally re-allocate liability at these natural persons again. A (partial) legal capacity of autonomous systems thus does not solve any arising problems, but in contrast creates a multitude of new ones.

b) A second approach tries to **allocate the risk between the designer, manufacturer and user of the respective autonomous system which caused the harmful act/did not perform correctly any other obligations.** This is also how todays legal systems handle liability already. Allocation may, anyhow, become more demanding if a direct causation of the respective human's contribution cannot be exactly reconstructed or calculated any more once the damage has occurred, or – at least in fault-based liability legal systems – if all human parties involved succeed to provide evidence that the damage occurred was not predicable or preventable from their perspective and that they are thus exempt from proper liability.

   For a very similar situation private law has already developed earlier well-established and efficient mechanisms to guarantee the compensation of damages of victims of devices which are too complex to be entirely controlled by humans and too dangerous to let the victim bear the risk that the defendant does not have the funds to compensate these damages: Operators of such "permitted hazards", which range from every-day tools as motor vehicles to very specific devices as nuclear power plants, are subjects to strict liability and compulsory insurance, based on the concept that those who profit from a certain advanced technology shall also bear the risks of running it, not regarding any proper fault or negligence – and that they should additionally be insured in case that any damage substantiates. Also the Estonian legal system has established these mechanisms (**liability for damage caused by major source of danger**[16], **product liability**[17]). Similar mechanisms could be established for tools whose hazards do not derive from their direct physical potential danger to others, but from the autonomous nature of its decision-making procedures – as autonomous systems. Liable in that context would categorically be that person on whose account the system is operated, although the user/operator may generally be entitled to compensate this liability from programme designers/manufacturers, depending on their eventual contributory breaches of duties. This recourse anyway would take place according to the same rules and regulations already at hand, e.g. for product liability, although some amendment providing certain clarifications in terms standard characteristics of autonomous systems etc. may be advisable.

24. For the area of autonomous intelligent systems, **the research group has not identified any acute and major gaps** in current legislation which would either leave some scenarios without adequate legal protections or cause legal uncertainty to such a degree that it would constitute a significant and unjustified burden on innovation and economic development.

25. Also taking into account the fact that the Estonian legal system has as its integral part a large body of international norms, of which the various pieces of EU law are of particular importance (incl, in many instances, direct effect), does not lead to different interim results Given the fact that Estonia is a part of EU single market, and that EU has aspirations of creating a single digital area, introducing purely local regulations for a subject matter that is not inherently subject to territoriality would be highly problematic. One should thus adopt and, where necessary, further develop the principles enshrined for example in GDPR[18] and other relevant EU instruments, instead of "inventing the wheel" in terms of national regulations.

26. Whereas there are no fundamental changes required, **there nevertheless are important amendments that should be considered and, if necessary, introduced in the existing legislation in order to properly take into account the increasing use of autonomous intelligent systems both in public and private sectors**. The legislation we must look at include the following:
    a. A. Statutes governing the exercise of public authority and public (state) liability (state liability act[19], procedural codes for courts of law, administrative procedure

[16] VÕS § 1056

[17] VÕS § 1061

[18] Regulation (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 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)

[19] RT I, 17.12.2015, 76

act[20], law enforcement act[21], public information act[22], taxation act[23], general part of the social code act[24], various acts governing public services and support, etc)

b. B. Statutes governing civil law and liability (general principles of the civil code act[25], law of obligations act[26], environmental liability act[27])
c. C. Statutes governing criminal liability (penal code[28])
d. D. Statutes governing fair business practices and consumer protection (competition act[29], consumer protection act[30], general part of the economic activities code act[31])
e. E. Statutes protecting individuals from discrimination and arbitrary decision-making (equal treatment act[32], gender equality act[33])
f. Certain sector-specific statutes (health services organisation act[34], traffic act[35], information society services act[36], acts governing financial services, etc)

[20] RT I, 28.12.2017, 21
[21] RT I, 12.12.2018, 46
[22] RT I, 07.12.2018, 9
[23] RT I, 07.12.2018, 5
[24] RT I, 30.12.2015, 3
[25] RT I, 30.01.2018, 6
[26] RT I, 22.03.2018, 4
[27] RT I, 10.11.2016, 8
[28] RT I, 07.12.2018, 21
[29] RT I, 07.12.2018, 22
[30] RT I, 12.12.2018, 65
[31] RT I, 29.06.2018, 31
[32] RT I, 26.04.2017, 9
[33] RT I, 07.07.2015, 11
[34] RT I, 21.12.2018, 6
[35] RT I, 06.07.2018, 14
[36] RT I, 12.12.2018, 39

Publication V
2.1 Kasper, A., & **Antonov, A.** (2019). Towards Conceptualizing EU Cybersecurity Law. In *ZEI Discussion Paper*, C253. Bonn: Center for European Integration Studies [Zentrum für Europäische Integrationsforschung]. https://hdl.handle.net/20.500.11811/9849

Zentrum für Europäische Integrationsforschung
Center for European Integration Studies
Rheinische Friedrich-Wilhelms Universität Bonn

Discussion Paper

Agnes Kasper / Alexander Antonov

Towards Conceptualizing EU Cybersecurity Law

C253
2019

Dr. Agnes Kasper is a Senior Lecturer of Law and Technology at the Tallinn University of Tallinn, Department of Law. She has been teaching legal aspects of cybersecurity to law students, as well as to IT students in the cybersecurity program at TalTech since 2012. Dr Kasper holds diplomas in international business, law and management. She has received additional formal trainings on technical aspects of cybersecurity and digital evidence. Dr Kasper served at embassies, human rights organizations and she was leading the legal department in an IT consultancy and development company. She has also acted in advisory capacity in consultations with governments on issues relating to cybersecurity. Her research focuses on regulatory aspects of cybersecurity; in particular, she is interested in emerging technologies. She is a frequent speaker in events, seminars, conferences focusing on aspects of law, technology and security. In 2015 the Estonian Ministry of Defence awarded Dr Kasper with the first prize for her doctoral thesis „Multi-Level Analytical Frameworks for Supporting Cyber Security Legal Decision Making".

Alexander Antonov is a doctoral student at the Department of Law at TalTech University, Estonia. He holds a Master Degree in ''M.Sc. International Security and Law'' from the University of Southern Denmark. His main research areas are Public International Law and International Humanitarian Law. In his Ph.D. project he scrutinizes the tools International Law provides for the regulation of the use of emerging technologies in warfare.

Agnes Kasper / Alexander Antonov

Towards Conceptualizing EU Cybersecurity Law

1. Introduction

The European Union has a wide spectrum of legal instruments addressing various aspects of cybersecurity, ranging from electronic communication laws, data protection regulations through network and information security legislation to instruments dealing with cybercrime and recommendations on coordinated response to large scale cyber incidents – all this without having a commonly accepted definition of cybersecurity.

The 2013 Cybersecurity Strategy describes cybersecurity in general terms in a footnote as the "safeguards and actions that can be used to protect the cyber domain, both in the civilian and military fields, from those threats that are associated with or that may harm its interdependent networks and information infrastructure". [1] The proposed Cybersecurity Act purports to define cybersecurity as it "comprises all activities necessary to protect network and information

[1] European Commission, "Joint Communication to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace," 7 February, 2013.

systems, their users, and affected persons from cyber threats”,[2] however the definition is not explained in available preparatory documents, although the word cybersecurity is used 462 times in the impact assessment.[3] According to these existing wordings, which are overly broad, cybersecurity is a process or activity. Other instruments, such as the 2017 Communication on Resilience, Deterrence and Defence: Building strong cybersecurity for the EU[4], also refer to cybersecurity as it was an attribute or a desired state to be achieved. The lack of clarity about this core concept raises questions about coherence and consistency of already adopted and newly proposed legislative acts in the field of cybersecurity. Precisely what harms EU cybersecurity-related laws seek to prevent? Understanding the harms is essential to prioritizing goals, limits and scope of the relevant legal framework.

Therefore, we propose to take a step back and examine the subjects, methods and reasons behind relevant EU regulatory acts in order to determine the scope and goals of EU laws that aim to promote cybersecurity. It is also expected that “EU cybersecurity law” as a legal framework is constrained by the competences of the EU, as well as by the principles of subsidiarity and proportionality, hence will necessarily differ from that of a federal state or that of a Member State. Conceptualizing EU cybersecurity law will also allow to examine how lawmakers can improve the legal framework for

[2] COM (2017) 477: *Proposal for a Regulation of the European Parliament and of the Council* on ENISA, the “EU Cybersecurity Agency”, and repealing Regulation (EU) 526/2013, and on Information and Communication Technology cybersecurity certification (”Cybersecurity Act”).

[3] See *Commission Staff Working Document Impact Assessment*. Accompanying the document *Proposal for a Regulation of the European Parliament and of the Council* on ENISA, the “EU Cybersecurity Agency”, and repealing Regulation (EU) 526/2013, and on Information and Communication Technology cybersecurity certification (”Cybersecurit. Act”), SWD/2017/0500 final – 2017/0225 (COD); opinion of the Regulatory Scrutiny Board, SEC/2017/0389 final. Online at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=pi_com:SEC(2017)389).

[4] Joint Communication to the European Parliament and the Council, Resilience, Deterrence and Defence: Building strong cybersecurity for the EU, JOIN/2017/0450 final.

cybersecurity and contribute to the stated need (by ENISA, 2012) to define common cybersecurity goals across the EU. In order to illustrate the challenges, we examine a high-profile cyber-attack (i.e. Wannacry ransomware 2017) to gain a fuller picture of the harms caused in or to Europe.

2. Wannacry crisis in the EU

2.1 The attack

Digital transformation, which is brought about by the rapid pace in technological change, challenges the regulatory framework of EU Member States' institutions, their private businesses and the EU as a whole.[5] Prior to forming a broad concept of "EU cybersecurity law", it is of utmost importance to scrutinize the severe impact a malicious cyber-attack can cause on different stakeholders.

To this end, we choose to study the large-scale cyber-attack "Wannacry", which "brought the issue of cyber resilience into the mainstream of public and political discourse", and we use it to shed some light upon what EU cybersecurity laws are about.[6]

On 13th May, 2017, the last business day of the week, a message reading *"Oops, your files have been encrypted"* appeared on more than 200.000 computer screens throughout the world demanding a ransom of between $ 300 and $ 600 being paid in Bitcoin in exchange

[5] Maria Solarte-Vasquez and Katrin Nyman Metcalf, "Smart Contracting: A Multidisciplinary and Proactive Approach for the EU Digital Single Market", *Baltic Journal of European Studies,* vol. 7, no. 2 (2017), p. 218.

[6] Julian King, "Commissioner King's keynote speech at the, 'WannaCry again? Making our businesses digitally great and cyberproof' conference", 15 February, 2018. Online at: https://ec.europa.eu/commission/commissioners/2014-2019/king/announcements/commissioner-kings-keynote-speech-wannacry-again-making-our-businesses-digitally-great-and_en. "Last year, the WannaCry malware did not just cause computers to freeze, but hospitals to close. It brought the issue of cyber resilience into the mainstream of public and political discourse."

for decrypting files stored on compromised devices.[7] Various major businesses in the European Union as the French carmaker Renault, the German transport company DB, or Spain's telecommunications operator Telefónica felt victim to the ransomware attack, which these companies could have avoided had they followed Microsoft's advise in March to close a vulnerable loophole in the Windows operating system by updating their computer software.[8] One of the gravest consequences of the disruptive attack was witnessed by the British National Health Service (NHS), where 80, or one third of all NHS trusts and 595 general practises were forced to cancel almost 19000 appointments, hundreds of surgeries and even cancer referrals.[9] Wannacry did not hold back from spreading to devices in critical infrastructure, disrupting information systems, which store laboratory data and radiographs.[10]

The malware had two components. The first, called EternalBlue, a tool exploiting a vulnerability in Windows operating systems enabling the worm to reach other computers without the end user's

[7] Russell Goldman, "What We Know and Don't Know About the International Cyberattack," The New York Times, 12 May, 2017. Online at: https://www.nytimes.com/2017/05/12/world/europe/international-cyberattack-ransomware.html; see also: Chris Graham, "NHS cyber attack: Everything you need to know about 'biggest ransomware' offensive in history," *The Telegraph*, 20 May 2017. Online at: https://www.telegraph.co.uk/news/2017/05/13/nhs-cyber-attack-everything-need-know-biggest-ransomware-offensive/.

[8] Sam Jones, "Timeline: How the WannaCry cyber attack spread," *FT*, 14 May, 2017. Online at: https://www.ft.com/content/82b01aca-38b7-11e7-821a-6027b8a20f23; consider also: Handelsblatt, "Cyberangriff legt 450 Bahn-Computer lahm,", 16 May 2017. Online at: https://www.handelsblatt.com/unternehmen/handel-konsumgueter/wanna-cry-cyberangriff-legt-450-bahn-computer-lahm/19809190.html?ticket=ST-2221470-N9RWTH0YgdtJ5A3foRbK-ap2); see further: Michael Schilliger, "Elf Antworten zur Cyberattacke 'WannaCry'," *NZZ*, 13 May, 2017. Online at: https://www.nzz.ch/digital/globaler-cyberangriff-sieben-antworten-zur-cyberattacke-wanacrypt-20-ld.1292982).

[9] National Audit Office, *Investigation: WannaCry cyber attack and the NHS*, 25 April, 2018. Online at: https://www.nao.org.uk/wp-content/uploads/2017/10/Investigation-WannaCry-cyber-attack-and-the-NHS.pdf; see further: Graham, *supra* note 7; see also: BBC, "NHS 'could have prevented' WannaCry ransomware attack," 27 October, 2017. Online at: https://www.bbc.com/news/technology-41753022.

[10] Schilliger, *supra* note 8.

permission through channels created to transmit and share data.[11] As soon as a recipient opened an enclosed file in an email, which contained the malicious programme, the malware started spreading at an unprecedented speed to other Windows systems linked to the infected computer.[12] The second element pertains to the encryption of the files stored on the computer, locking down data and systems. A message box popped up on the screen demanding the user to pay in cryptocurrency to restore the accessibility of one's data.[13]

It is worthwhile mentioning that the disruptive component of Wannacry, EternalBlue, was initially written by the N.S.A. to take advantage of Windows's vulnerability for spying activities on companies and foreign intelligence services.[14] One month prior to the

[11] Qian Chen & Robert Bridges, "Automated Behavioral Analysis of Malware: A Case Study of WannaCry Ransomware," Conference Paper (2017), at 2: "The dropper of the malware carries two components. One uses the "EternalBlue" exploit against a vulnerability of Windows' Server Message Block (SMB) protocol to propagate, and the other is a WannaCry ransomware encryption component."; see further: Liliy Hay Newan, "The Ransomware Meltdown Experts Warned About Is Here," *Wired*, 5 December, 2017. Online at: https://www.wired.com/2017/05/ransomware-meltdown-experts-warned/ "Once WannaCry enters a network, it can spread around to other computers on that same network, a typical trait of ransomware that maximizes the damage to companies and institutions.".

[12] Nicole Perlroth and David E. Sanger, "Hackers Hit Dozens of Countries Exploiting Stolen N.S.A. Tool," *The New York Times*, 12 May, 2017. Online at: https://www.nytimes.com/2017/05/12/world/europe/uk-national-health-service-cyberattack.html?hp&action=click&pgtype=Homepage&clickSource=story-heading&module=first-column-region&region=top-news&WT.nav=top-news: "The malware was circulated by email. Targets were sent an encrypted, compressed file that, once loaded, allowed the ransomware to infiltrate its targets. The fact that the files were encrypted ensured that the ransomware would not be detected by security systems until employees opened them, inadvertently allowing the ransomware to replicate across their employers' networks."; see also: Graham, *supra* note 7: "Hackers have been spreading "ransomware" called WannaCry, also known as WanaCrypt0r 2.0, WannaCry and WCry. It is often delivered via emails which trick the recipient into opening attachments and releasing malware onto their system in a technique known as phishing".

[13] See e.g.: Goldman, *supra* note 7.

[14] Schillinger, *supra* note 8; see also: The International Institute for Strategic Studies, "The WannaCry ransomware attack," *Strategic Comments*, vol. 23, no. 4 (2017), at vii-viii.

attack, this crucial element of the code turned out to have fallen into the hands of a cyber criminal group, known as “Shadow Brokers“ who leaked it to the public on their webpage in April.[15] Various actors, there under Microsoft, heavily criticised the N.S.A. and some even claimed that it should incur responsibility for the cyber-attack.[16]

2.2 Response and impact

Amid the outbreak of the virus, Microsoft provided an emergency patch to Windows XP, Windows 2003 and Windows 8 users that helped prevent the malware from spreading further.[17] Additionally,

[15] The International Institute for Strategic Studies, *supra* note 14; consider also: Andy Greenberg, “Hold North Korea Accountable for Wannacry – and the NSA, too,” *Wired*, 19 December, 2017. Online at: https://www.wired.com/story/korea-accountable-wannacry-nsa-eternal-blue/: “WannaCry's origins stretch back to April, when a group of mysterious hackers calling themselves the Shadow Brokers publicly released a trove of stolen NSA code. The tools included an until-then-secret hacking technique known as EternalBlue, which exploits flaws in a Windows protocol known as Server Message Block to remotely take over any vulnerable computer”.

[16] Brad Smith, “The need for urgent collective action to keep people safe online: Lessons from last week’s cyberattack,” The Al Blog, 14 May, 2017. Online at: https://blogs.microsoft.com/on-the-issues/2017/05/14/need-urgent-collective-action -keep-people-safe-online-lessons-last-weeks-cyberattack/sm.001p0mwmqc3 ld351 07z1pj4ntjs26: “{E}xploits in the hands of governments have leaked into the public domain and caused widespread damage. An equivalent scenario with conventional weapons would be the U.S. military having some of its Tomahawk missiles stolen. And this most recent attack represents a completely unintended but disconcerting link between the two most serious forms of cybersecurity threats in the world today – nation-state action and organized criminal action.”; see also: Greenberg, supra note 15; see further: Ellen Nakashima and Craig Timberg, “NSA officials worried about the day its potent hacking tool would get loose. Then it did.,” The Washington Post, 16 May, 2017. Online at: https://www.washingtonpost.com/ business /technology/nsa-officials-worried-about-the-day-its-potent-hacking-tool-would-get-loose-then-it-did/2017/05/16/50670b16-3978-11e7-a058-ddbb23c75d82_story. html? noredirect=on&utm_term=.ececf4d96f19.

[17] Mark Scott and Nick Wingfield, “Hacking Attack Has Security Experts Scrambling to Contain Fallout,” *The New York Times*, 13 May, 2017. Online at https://www.nytimes.com/2017/05/13/world/asia/cyberattacks-online-security- .html: “Microsoft took the unusual step of releasing free security patches for older versions of Windows, including Windows XP, that it no longer routinely updates. It said the patches could help protect users from attacks, which have not targeted Windows 10, the latest edition of the software.” Greenberg, *supra* note 15.

by coincidence a security analyst from the UK found a ‘’kill switch“ in the code, which he activated by purchasing a web address the ransomware inquired.[18] The attack subsided significantly after a few days, but the vulnerability in the systems remained for those computers that had still not been updated since the hackers could easily rewrite the code and infect other systems without a kill-switch implanted. It was also for this reason the European Cybercrime Centre (EC3), Europol, distributed awareness materials on social media platforms and created an information webpage outlining key strategies on how to protect private data from malware attacks.[19] In addition, it referred to the NoMoreRansom initiative, which primarily informs and dissuades consumers affected by ransomware from financing cybercrime activities.[20] The majority of large corporations did not give in to the demands of the cyber criminals and spend most resources on either rebuilding or restoring data from backups.[21]

The cyber-assault has been attributed to the State sponsored North Korean cybercrime group called “Lazarus” and affected thousands of

[18] Jones, *supra* note 8: “Security analysts stress it could have been worse but for the actions of an anonymous British security researcher. After lunch on Friday, a 22-year-old cyber analyst, who writes online under the pseudonym MalwareTech, returned to his desk and spotted something crucial in WannaCry’s code — the first stage of its infection process. The obscure web address the ransomware was querying, he noticed, was unregistered and inactive. So he bought it for $11 and activated it. It turned out to be a form of “kill switch” baked into WannaCry by its creators. Activating the address told the ransomware, upon each new infection, not to proceed any further. Once he had control of it, WannaCry was stopped in its tracks”.

[19] Europol, “How does the WannaCry ransomware work?,” 4 December, 2018. Online at: https://www.europol.europa.eu/wannacry-ransomware); see also: General Secretariat of the Council of the European Union, Cybersecurity – Information from the Commission, 9621/17, 31 May, 2017, at 2: “In the context of the public response to the WannaCry attack, Europol (via its European Cybercrime Centre [EC3]) created a dedicated information page 3 and disseminated flyers and awareness materials via Europol social media channels”.

[20] General Secretariat of the Council of the European Union, *supra* note 19, at 2.

[21] Jonathan Beer, “WannaCry” ransomware attack losses could reach $4 billion,” *CBSNews*, 16 May, 2017. Online at: https://www.cbsnews.com/news/wannacry-ransomware-attacks-wannacry-virus-losses/: “Most of the organizations won’t pay {…} “They will rebuild and recover from their backups or other sources.”

companies and public services worldwide.[22] In an interview with the German news service "Tagesscha" the head of Europol, Steven Wilson, described the events as the "largest cyber-attack the world witnessed so far" taking a great toll on the economy.[23] In the same vein, leading IT experts as Mikko Hyppönen spoke of the "largest ransomware-epidemic in history".[24] Ransomware attacks were not a new phenomenon in 2017. The magnitude of Wannacry, however, was "unprecedented" with over 230.000 computers in 150 countries being targeted in total.[25] It was not without reason why also the director of the European Union Agency for Law Enforcement Cooperation, Rob Wainright, classified the virus as a novel type of malicious attack.[26]

Considering the EU's efforts on strengthening stability of cyberspace through international cooperation, one month after Wannacry unfolded, the Council of the European Union approved the "Draft Council Conclusions on a Framework for a Joint EU Diplomatic

[22] BBC, "Cyber-attack: US and UK blame North Korea for WannaCry," 19 December, 2017. Online at: https://www.bbc.com/news/world-us-canada-42407488; see also: Reuters, "Britain believes North Korea was behind 'WannaCry' NHS cyber attack," 27 October 2017. Online at: https://uk.reuters.com/article/us-britain-security-northkorea/britain-believes-north-korea-was-behind-wannacry-nhs-cyber-attack-idUKKBN1CW153.

[23] Tagesschau, "Europol zu WannaCry: Das ist der größte Cyberangriff bisher," 17 May, 2017. Online at: https://www.tagesschau.de/ausland/europol-wannacry-101.html.

[24] Spiegel Online, "WannaCry" – Attacke – Fakten zum globalen Cyberangriff," 13 May, 2017. Online at: http://www.spiegel.de/netzwelt/web/wannacry-attacke-fakten-zum-globalen-cyber-angriff-a-1147523.html.

[25] Europol, *supra* note 19: "The recent attack is at an unprecedented level and requires a complex international investigation to respond effectively and identify the culprits." Consider also: Julian King, "Commissioner King's speech at the EU Cybersecurity *Conference Digital Single Market, Common Digital Security 2017*," 15 September, 2017. Online at: https://ec.europa.eu/commission/commissioners/2014-2019/king/announcements/commissioner-kings-speech-eu-cybersecurity-conference-digital-single-market-common-digital-security_en.

[26] CBS, *supra* note 21: "There is no precedent for a ransomware attack of this kind of scale," {…}. This is the first one that we have seen … that has been able to attack computers directly with this kind of success."

Response to Malicious Cyber Activities", the so-called "Cyber Diplomacy Toolbox".[27] With this initiative, the EU member states reiterated that cyber-attacks do not occur in a legal vacuum and agreed that the EU will respond with restrictive measures against individuals affiliated with cybercriminal gangs or even against states which promote such malicious activities by providing either sanctuary for them or hire them for political purposes.[28]

As stated by the General Secretariat of the Council of the EU, the Wannacry ransomware attack triggered cooperation between Member States within the framework of the NIS directive.[29] For the first time since its adoption, the affected EU countries exchanged intelligence on a cyber-attack on this legal basis.[30] In the State of the Union Address in 2017, the president of the EU Commission, Jean-Claude Juncker, mentioned cyber security as the EU's fourth policy priority of the subsequent year.[31] In summer 2018, the Council of the EU

[27] Council of the European Union, *Draft Council Conclusions on a Framework for a Joint EU Diplomatic Response to Malicious Cyber Activities ("Cyber Diplomacy Toolbox") – Adoption*, 7923/2/17 REV 2, 7 June 2017.

[28] *Ibid*: "The EU affirms that malicious cyber activities might constitute wrongful acts under international law and emphasises that States should not conduct or knowingly support ICT activities contrary to their obligations under international law, and should not knowingly allow their territory to be used for internationally wrongful acts using ICTs, as it is stated in the 2015 report of the United Nations Groups of Governmental Experts (UN GGE). {...}. The EU affirms that measures within the Common Foreign and Security Policy, including, if necessary, restrictive measures, adopted under the relevant provisions of the Treaties, are suitable for a Framework for a joint EU diplomatic response to malicious cyber activities and should encourage cooperation, facilitate mitigation of immediate and long-term threats, and influence the behavior of potential aggressors in a long term."

[29] General Secretariat of the Council of the European Union, *supra* note 19: "The recent WannaCry cyberattack where a wave of ransomware attacks impacted organizations and citizens across the globe was the first time where Member States exchanged information on cybersecurity incident within the mechanism for operational cooperation under the NIS Directive, the so-called Computer Security Incident Response Teams network. This is yet another real-life example that proves how important cooperation in the area of cybersecurity is."

[30] *Ibid*.

[31] Jean-Claude Juncker, "Fourth priority for the year ahead: I want us to better protect Europeans in the digital age." Online at: http://europa/eu/rapid/press-release_SPEECH-17-3165_en.htm.

recalled the Commission's 2017 recommendation on creating a "Coordinated Response to Large-scale Cybersecurity Incidents and Crises" and underlined, *inter alia*, that EU Member States "need to make use of the existing crisis management mechanisms, processes and procedures at national and European level".[32]

Debating malicious cyber activities in the EU, eleven months after the attack, the Foreign Affairs Council of the EU "condemn{ed} the malicious use of information and communications technologies (ICT), including in Wannacr" and "stresse{" that cyber-attacks "undermin" the EU's "stability, security and the benefits provided by the internet and the use of ICT".[33]

Considering the harms caused by Wannacry, even though none was injured or killed nor data had been stolen in the attack, (1) the economic damage was significant.[34] Whereas Cyence Risk Analytics estimated the costs at $ 4 billion, others predicted a loss of hundreds of millions of dollars.[35] (2) Not only did the assault temporarily hamper the companies' productivity, (3) but it also worsened their business reputation. Looking at the case of the NHS, the British public was seriously concerned about its national health service and questioned its failure to keep up with modern cybersecurity standard.[36] The image of the NHS suffered further when the UK

[32] General Secretariat of the Council, *supra* note 19, at 2-3.

[33] Council of the European Union, *Council conclusions on malicious cyber activities – approval*, 7517/18, 16 April 2018: "The EU firmly condemns the malicious use of information and communications technologies (ICTs), including in Wannacry and NotPetya, which have caused significant damage and economic loss in the EU and beyond. Such incidents are destabilizing cyberspace as well as the physical world as they can be easily misperceived and could trigger cascading events. The EU stresses that the use of ICTs for malicious purposes is unacceptable as it undermines our stability, security and the benefits provided by the Internet and the use of ICTs."

[34] Suzanne Barlyn, "Global cyber attack could spur $53 billion in losses: Lloyd's of London," *Reuters*, 17 July 2017. Online at: https://www.reuters.com/article/us-cyber-lloyds-report-idUSKBN1A20AB.

[35] Beer, *supra* note 21: "Cyber risk modeling firm Cyence estimates the potential costs from the hack at $4 billion, while other groups predict losses would be in the hundreds of millions."

[36] Graham, *supra* note 7; see also: BBC, *supra* note 22.

Department of Health and Social Care made public that Wannacry resulted in a loss of £ 92 million in British taxpayers money.[37] (4) Decreased public confidence into e-services, which many EU-citizens rely on in their everyday-life[38], and into the security of computer systems in general, that store vast amount of sensible private data of millions of clients and patients, constituted additional harms. (5) Taking a broader view on the effects of the attack, it can be said that cyberspace and the physical world in general was destabilized. (6) Critical infrastructures were affected in the EU, which is concern for the sovereignty and territorial integrity of the Member States.

Despite the Commission's multidimensional approach in improving the EU member states' cyber resilience, there is no commonly accepted definition of cybersecurity in the EU, leaving each of the EU governments room for different interpretation of this increasingly important legal area. Juncker's statement that cyber threats could destabilize the economy of democracies more effectively than ''guns and tanks" given the speed and virulence malware spread with, serves as further proof for the need to formulate the idea of EU cybersecurity law.[39] With European cybersecurity being challenged every day, the EU's goal to harmonize national law systems of member states in regard to cyber security and therefore increase the EU's resilience against cyber-attacks can be better attained if the affected states identified the multifarious harmful effects on their economy and society. With six main harms caused by Wannacry being established, the subsequent chapters set out the core elements of EU cybersecurity law.

[37] Matthew Field, "WannaCry cyber attack cost the NHS £92m as 19.000 appointments cancelled," *The Telegraph*, 11 October 2018. Online at: https://www.telegraph.co.uk/technology/2018/10/11/wannacry-cyber-attack-cost-nhs-92m-19000-appointments-cancelled/.

[38] Tanel Kerikmäe (ed.), *Regulating eTechnologies in the European Union: Normative Realities and Trends*, 2014, p. 1.

[39] Jean-Claude Juncker, *supra* note 31: "Cyber-attacks can be more dangerous to the stability of democracies and economies than guns and tanks{…} Cyber attacks know no borders and no one is immune".

3. Cybersecurity: lost in translation?

3.1 Lack of consistent terminology

The cybersecurity field in general uses many concepts from neighbouring domains, but it has been infiltrated with terms from political science as well.[40] Cybersecurity is not synonymous with security of network and information systems, although for the last few years there has been some confusion for a good reason, which was also pointed out in a recommendation by the European Network and Information Security Agency (ENISA): Member States should "[a]gree on a commonly accepted working definition of cyber security that is precise enough to support the definition of common goals across the EU".[41] Cybersecurity remains a field where different perceptions and narratives determine its content for the respective actor, in particular that EU Member States emphasize certain aspects of cybersecurity in their strategic and policy documents, while downplaying others.[42] Terminology used in international forums, such as the UN, where discussion is held about 'information security' (although certainly deals with issues above the micro-level), reflects on the lack of coherent conceptual framework in this field.[43]

[40] For example it is customary to label some hacker groups as 'Advanced Persistent Threat' or APT, in addition to giving them descriptive fantasy names, such as APT29 or Cozy Bear – a Russian hacker group believed to be associated with Russian intelligence.

[41] ENISA, National Cyber Security Strategies – Setting the course for national efforts to strengthen security in cyberspace, 2012. p. 12. Online at: https://www.enisa.europa.eu/publications/cyber-security-strategies-paper)

[42] See the different national concepts in the cybersecurity strategies of EU Member States, collected at ENISA website. Online at: https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map.

[43] The UK in its 2017 Response to General Assembly resolution 71/28 "Developments in the field of information and telecommunications in the context of international security" stated that "The United Kingdom uses its preferred terminology of 'cybersecurity' and related concepts throughout its response, to avoid confusion given the different interpretations of the term 'information security' in this context." Online at: https://www.un.org/disarmament/topics/informationsecurity/.

The difference between data security and network and information security[44] also needs to be emphasized, since although data security is a vital component of cybersecurity, for instance the Wannacry attack compromised more than just the availability of data and affected European critical infrastructure operators in the health, energy, transport, finance and telecom sectors, manufacturers and service providers throughout Europe.[45] Data and information is held in systems and transmitted through networks, which are increasingly relied on for everyday services, in particular when put into the context of Internet of Things era, where billions of appliances are connected to the internet. Focusing on information and data security, as well as systems and network security ensures that threats to cyber-physical systems, such as smart grids, autonomous automobiles, medical monitoring, industrial control systems, robotic surgery systems, etc. are also addressed. In turn, this enables regulators to link security compromises of systems and networks to their consequences, such as potential physical injuries or property damages.

A working definition of cybersecurity has been used in the 2013 Cybersecurity Strategy of the European Union, which in footnote no. 4 states that "Cyber-security commonly refers to the safeguards and actions that can be used to protect the cyber domain, both in the civilian and military fields, from those threats that are associated with or that may harm its interdependent networks and information infrastructure. Cybersecurity strives to preserve the availability and

[44] The ISO/IEC 27000: 2017 standard defines information security as the 'preservation of confidentiality, integrity and availability of information'. ISO/IEC 27032:2018 refers to network security as it 'is concerned with the design, implementation and operation of networks for achieving the purposes of information security on networks within organizations, between organizations, and between organizations and users'. ISO/IEC 27032:2018 defines cyberspace security as 'Preservation of confidentiality, integrity and availability of information in Cyberspace', and it emphasizes that cybersecurity is not synonymous with information, network, internet security or critical information infrastructure protection.

[45] ENISA, WannaCry Ransomware: First ever case of cyber cooperation at EU level, 15 May, 2017. Online at: https://www.enisa.europa.eu/news/enisa-news/wannacry-ransomware-first-ever-case-of-cyber-cooperation-at-eu-level.

integrity of the networks and infrastructure and the confidentiality of the information contained therein."[46] The High Level Scientific Advisors on cybersecurity in the European digital single market has also added the same definition to their glossary, but felt that this needs to be complemented by a reference to "prevention and law enforcement measures to fight cybercrime".[47]

These approaches made little distinction between the technically oriented concepts, such as network and information security, and the emerging understanding seems to be that cybersecurity addresses concerns beyond the micro level of organizations and businesses. ENISA has also concluded that "[c]ybersecurity is an enveloping term and it is not possible to make a definition to cover the extent of the things Cybersecurity covers", however contextual definitions are already in use.[48] Therefore, we do not aim to define cybersecurity in this paper, but we work with existing understandings, in order to put cybersecurity into context for the legal community.

3.2 Cyberspace elements - what needs to be secured?

In order to unlock the concept of cybersecurity law, we need to find the constitutive elements of cyberspace that needs to be secured. We adopt the definition by Ottis and Lorents, who stated that "cyberspace is a time-dependent set of interconnected information systems and

[46] European Commission, "Joint Communication to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace," 7 February, 2013.

[47] SAM High Level Scientific Advisors, Scientific Opinion, no. 2/2017, Cybersecurity in the European Digital Single Market, 27 March 2017, p. 97. Online at: https://ec.europa.eu/research/sam/pdf/sam_cybersecurity_report.pdf#view=fit&pagemode=none.

[48] ENISA, "Definition of Cybersecurity – Gaps and overlaps in standardisation", December 2015. Online at: https://www.enisa.europa.eu/publications/definition-of-cybersecurity.

human users that interact with these systems”.[49] It is thus revealed that two elements of the system (cyberspace) are information systems and human users, and the properties of these elements are interconnectedness and interaction with information systems respectively. Cybersecurity laws can relate to either of these elements, i.e. addressing the state of information systems or conduct of human users. Norms expressed in regulatory instruments aim to influence these elements, by stating that “something ought to or may or must not be or be done”.[50]

As to the first element, information systems, we can find that concepts of network- and information security and relating industry standards have already elaborated on how to approach the task of securing interconnected information systems (which necessarily include infrastructure, networks, data and information). [51] Cybersecurity professionals commonly refer to three security requirements, confidentiality, integrity and availability, known as the “CIA Triad”[52], which can relate not only to data and information in systems and networks, but also to systems and networks themselves.[53]

As to the second element of cyberspace, the human user, however, it also becomes clear that the technically-oriented approach to cybersecurity, when nearly-equated with network and information security, might lose sight of a constitutive element of the system: the human user that interact with information systems.

[49] Ottis, R., Lorents, P., Cyberspace: Definition and Implications. In Proceedings of the 5th International Conference on Information Warfare and Security, Dayton, OH, USA, 8/9 April, 2010. Reading: Academic Publishing Limited, pp. 267-270.

[50] G. H. v. Wright, Norm and Action, 1963.

[51] See a reference material for relevant standards in ENISA, Definition of Cybersecurity, Gaps and overlaps in standardization, 2015. Online at: https://www.enisa.europa.eu/publications/definition-of-cybersecurity.

[52] According to ISO/IEC 27000/2017. Confidentiality refers to a property that information is not made available or disclosed to unauthorized individuals, entities or processes; Integrity is the property of accuracy and completeness; and Availability is the property of being accessible and usable upon demand by an authorized entity.

[53] See also this approach in Jeff Kosseff, Defining Cybersecurity Law, Iowa Law Review, vol. 103: 985, 2018, pp. 985-1031.

Solms and Niekerk held that while information security refers to the human users' role in the security process, in cybersecurity humans become targets or inadvertent participants of cyber-attacks, hence there are threats that fall outside the scope of information security.[54] Examples include cyber bullying, which does not (necessarily) constitute loss of confidentiality, integrity and availability of data, systems or networks, but causes a direct harm to the person being bullied.[55] Another case in point would be interference with automated home appliances, such as a security system, which can be remotely turned off in order to burgle the home, where again it can be argued that there is no impact on confidentiality, integrity and availability of information assets and system of the victim.[56] Affected are other assets of the person. Accordingly, cybersecurity is more than the mere protection of networks and information systems, it also covers the protection of functions and assets that rely on or can be reached via cyberspace.[57]

Therefore the process of cybersecurity should have aims and objectives that goes beyond the mere protection of confidentiality, integrity and availability of information, systems and networks themselves, and address the harms that may result as a consequence of degradation of functioning of computer systems, or due to interference with some interactions between information systems and their users. Yet, we should be more focused on aggregate interactions, from the perspective of the society. In the cyber-enabled society, where information's importance is equivalent to that of money, energy, etc. and computerized systems are used to govern the society, in the center of focus are threats, risks, incidents, unlike in approaches

[54] Rossouw von Solms, Johan van Niekerk, From information security to cyber security, Computers & Security, 38, 2013, pp. 97-102.

[55] Ibid. 99.

[56] Ibid.

[57] Ibid. 102.

to information society, e-society or IT society etc.[58] In other words the main point of concern for cybersecurity is the functioning of societies that - to any degree - depend on computerized systems to the extent that severe degradation in the functioning of these computerized systems can pose an existential threat to that society.[59] But interference with interactions between the society and computerized systems can also have similar impact.

Examples can include the degradation of the functioning of the information systems in the financial sector as a whole, in a society, where 98% of all financial transactions are completed via electronic means. The consequences of such events in 2007 in Estonia were felt not only on the level of the individual financial institutions, such as the interruption of their operations and unavailability of internet banking interfaces for customers, etc. but it affected the financial sector as a whole. Similarly, the Wannacry attack bore significant influence on individual companies and institutions, but the scale of disruption also affected the normal existence of the society in the UK, 80 out of 236 hospital trusts' services were impacted, and 8% of General Practitioners practices felt victim to the attack.[60]

However, degradation of the functioning of computer systems may not always be involved, where we can still detect interference with interactions between society and information systems, in particular taking into account the recent years technological developments in the field of artificial intelligence. For example in case using troll armies (automated, or potentially artificial intelligence based) in social media networks to polarize audiences on social and political issues, do not necessarily degrade the functioning of information systems and

[58] Lorents P., Ottis R., Rikk R., Cyber Society and Cooperative Cyber Defence, in: Aykin N. (eds) Internationalization, Design and Global Development, IDGD, 2009. Lecture Notes in Computer Science, vol. 5623, Springer, Berlin/Heidelberg.

[59] Ibid, p. 180.

[60] UK, NHS Report, "Lessons learned review of the WannaCry Ransomware Cyber Attack", 2018. Online at: https://www.england.nhs.uk/wp-content/uploads/2018/02/lessons-learned-review-wannacry-ransomware-cyber-attack-cio-review.pdf .

networks, but aims to influence the interactions between the systems and users. A recent media report in 2018 stated that Russian troll factories have been used to discredit life-saving vaccines.[61] Shortly before this, the World Health Organization also published worrisome statistics indicating record high measle cases, including at least 37 fatal infections in Europe in 2018, although vaccination provides effective protection against the disease.[62] We are not able, nor have the intention to show a causal link between the troll's action and the measles outbreak in this particular case, nevertheless it suggests the magnitude of impact of a potentially effective campaign by trolls to manipulate the population into self-harming behaviour, or as we see it interfering with the interactions between the society and computerized systems, without degrading the functioning of these systems.

3.3 Towards a consequences-based approach to cybersecurity in the EU

The EU's cybersecurity efforts as a whole reflect a comprehensive understanding and approach, however it has been characterized by commentators as fragmented, and patchwork.[63] The EU has recently reached a political agreement on the Cybersecurity Act that signifies a global landmark in cybersecurity legislation.[64] Article 2 (1) of the (still) draft defines cybersecurity for the purposes of the regulation as "all activities necessary to protect network and information systems,

[61] Harry de Quetteville, "How Russian troll factories used Twitter to discredit life-saving vaccines", *The Telegraph*, 13.10.2018. Online at: https://www.telegraph.co.uk/news/0/inside-story-russian-troll-factories-using-twitter-discredit/.

[62] World Health Organization Regional Office for Europe, "Measles cases hit record high in the European Region", 20.08.2018. Online at: http://www.euro.who.int/en/media-centre/sections/press-releases/2018/measles-cases-hit-record-high-in-the-european-region.

[63] Maria Garzia Porcedda, "Patching the Pathchwork: appraising the EU regulatory framework on cybersecurity breaches", *Computer Law and Security Review,* 34, 2018, pp.1077-1098.

[64] European Commission, "EU negotiators agree on strengthening Europe's cybersecurity", 11.12.2018. Online at: https://ec.europa.eu/commission/news/cybersecurity-act-2018-dec-11_en.

their users, and affected persons from cyber threats".[65] This definition departs from the previous ones in a very significant way, since in addition to networks and information systems, it views the human user as the constitutive element of the system to be secured. It also implies a two-way of interaction[66] between human users and information systems, and it recognizes that information and interaction with information systems can influence events and human behaviour and society outside cyberspace. Therefore, the definition encompasses both the user's effect on information systems and the information systems' effects on users, however it would be plausible to think that the main concern is not about isolated cases.

The Wannacry incident's scale and immediate consequences resulted in significant disruption of a service as a whole in the healthcare system in the UK. Therefore, due to the reliance on computerized systems in the provisions of healthcare services the interaction between users and respective information systems was compromised – some due to infection by the Wannacry cyptoworm, but others due to turning off systems and devices as a precaution.[67] In particular in the cases of turning off the systems as a precautionary measure in order to avoid infection, we can argue that the availability of information is not compromised (the computers and devices can be turned back on and usage may continue), yet the service that is underlied by these systems is hampered.

In 2017 the Estonian ID card crisis also demonstrated that concern about *potential* authenticity and integrity breaches can lead to

[65] Interinstitutional File: 2017/0225(COD), Final version of the text on Proposal for *Regulation of the European Parliament and of the Council* on ENISA, the "EU Cybersecurity Agency", and repealing Regulation (EU) 526/2013, and on Information and Communication Technology cybersecurity certification ("Cybersecurity Act"). Online at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:ST_15786_2018_INIT&from=EN.

[66] Oxford dictionary defines interaction as reciprocal action or influence.

[67] National Audit Office, Investigation: WannaCry cyber attack and the NHS, 25 April, 2018. Online at: https://www.nao.org.uk/wp-content/uploads/2017/10/Investigation-WannaCry-cyber-attack-and-the-NHS.pdf.

significant disruptions in the delivery of e-services, although there are no reports about actual misuses.[68] Also in this case the interaction between society and the Estonian information systems was significantly disrupted, raising additional questions about trust in the systems, although the integrity and authenticity of the services and data was not actually compromised, and systems could perform their functions just as before the discovery of the vulnerability. Again, as a precautionary measure Estonian authorities blocked digital certificates of 760 000 ID cards, and started to update those persons' certificates first, who need their ID cards for their work, such as doctors, justice officials, civil servants, etc.[69] The Estonian lessons learned show that a non-incident can create a significant crisis, comparable to that of an incident.

The definition of cybersecurity in the draft Cybersecurity Act resonates with the service-oriented approach of Solms and Niekerk.[70] It covers technical and non-technical activities, however in the absence of a clear definition of cybersecurity it is difficult to devise legal tests for the purposes of determining precisely what activities would fall into the above category. While functions of and services that networks and information systems should perform can relatively easily be identified in technical terms, what can be considered as adverse effect on users and other persons is more challenging to identify given the endless ways cyberspace can be used. The analysis of the Wannacry case has already pointed towards some harms that may be considered, therefore protective measures and activities should address, *inter alia*, the potential and actual economic damages, decrease in productivity, reputational damages, decrease of trust in computer systems, destabilization of physical world, and potential losses in sovereignty.

[68] Tallinna Tehnikaülikool, ID-kaardi kaasuse õppetunnid, 2018. Online at: https://www.ria.ee/sites/default/files/content-editors/EID/id-kaardi_oppetunnid.pdf.

[69] Ibid.

[70] See 54.

We claim that what is to be secured by EU cybersecurity regulation are *interconnected information system*s, including data, information systems and networks, and aggregate *interactions* between human users and these information systems. In our view, what distinguishes network and information security regulation from cybersecurity regulation is that cybersecurity regulation aims to protect not only confidentiality, integrity and availability of data, information systems and networks[71], but also certain *interactions* between these and the society involving two or more Member States.

However, this line of thought and the proposed definition of cybersecurity by the EU Cybersecurity Act also opens a Pandora's box. What exactly is considered as a threat that can affect information systems' users and persons so that it becomes a concern for the EU? Which regulatory measures are best suited to address this issue? In which areas of cybersecurity management (i.e. prevention, detection, response, recovery) the EU is best placed to regulate? What oversight, supervision and enforcement measures ensure achievement of the objectives of the cybersecurity policy of the EU and respect the rule of law and fundamental human rights at the same time? The next part of this paper looks for some answers to these questions in the existing EU framework.

4. *Cybersecurity laws*

General legal frameworks and challenges

Gercke proposed a catalogue of "mandatory" and "optional" cybersecurity laws: the former category comprises of definitions, cybercrime laws and data protection legislation; while the latter optional areas include network and critical infrastructure protection, reporting obligations, international cooperation, electronic evidence,

[71] This is a simplified view from us in respect of security requirements that can also include authenticity, non-repudiation, accountability, reliability, etc. depending on the precise standard, context and needs.

electronic transactions, digital signatures, child online protection, liability of internet service providers and potential restrictions on the use of certain technology.[72]

Gercke offered a comprehensive view on cybersecurity legal framework and also noted that cybersecurity was often conflated with cybercrime, however not all cybersecurity incidents are criminal acts.[73] Wannacry used a known vulnerability for which Microsoft had issued a security patch in March 2017 for supported Windows versions[74], and spread to devices that have not applied the update. Not applying this patch, or other similars, generally does not constitute a criminal act, but may give rise to disciplinary or negligence claims, or non-compliance with data protection regulations, etc. However, precisely the unpatched vulnerabilities in systems were exploited by the creators of the Wannacry cryptovirus, which can already be described in the terms of the Cybercrime Convention. Fight against and preventing cybercrime is but one component of cybersecurity.[75]

Cybersecurity is still often seen as a purely technical or awareness problem, not a legal one. Available reports on the reactions and lessons learned from Wannacry did not address legal issues at the affected organizations' level.[76,77] Nevertheless, there are significant information gaps, often framed as problems in cybersecurity information sharing among private sector players, between private and public sector and between countries. These issues reach beyond

[72] Marco Gercke, Content of a Comprehensive Cybersecurity Legal Framework, Cri, 2/2014.

[73] Marco Gercke, Content of a Comprehensive Cybersecurity Legal Framnework, Cri, 2/2014, p. 34.

[74] See online at: https://support.microsoft.com/en-us/help/4013389/title.

[75] Marco Gercke, Content of a Comprehensive Cybersecurity Legal Framework, Cri, 2/2014, p. 34.

[76] See UK NHS Report, "Lessons learned review of the WannaCry Ransomware Cyber Attack", 2018. Online at: https://www.england.nhs.uk/wp-content/uploads/2018/02/lessons-learned-review-wannacry-ransomware-cyber-attack-cio-review.pdf .

[77] See Deutsche Bahn Interim Report, January-June 2017. Online at: https://www1.deutschebahn.com/resource/blob/1047480/1f573efc5d5d1f119dba29a882272eea/zb2017_dbkonzern_en-data.pdf.

technology, and concern exceptions in the data protection regulation, breach notification obligations of operators (private or public) and information exchange on potentially national security-related questions between EU Member States when collectively planning prevention, detecting, responding to or recovering from cyber incidents and events.

In EU context it also needs to be clarified which issues fall within the competence of EU law and what aspects remain within the competence of Member States, how the two levels interact, respecting the main principles of subsidiarity and proportionality. This involves mapping of cross-border interdependencies of cyber societies, since while an availability crisis can hit across sectors, the Estonian ID-card (chip vulnerability) crisis appears to be more contextual in the absence of pan-European information systems for the support of relevant societal functions.

It would be expected that the EU's primary concerns are rather the generic and strongly interlinked services, however local cybersecurity management should also remain a high priority. In the light of the EU's own modest operational capabilities in this regard (such as ENISA still has only very limited resources and performs advisory, training and support functions, although there are plans to increase EU level capabilities[78]), the EU's role in securing cyber societies will probably remain mainly complementary and supportive to that of Member States, including coordination, providing platforms for information exchange and cooperation, harmonization, mediating capacity building, research and development, etc. The more intensive role will be confined to areas, were the EU has exclusive competence or shares competences with Member States, most prominently concerning the Digital Single Market. In the following chapters we outline the main existing and proposed EU documents and legislation

[78] European Commission, "EU negotiators agree on strengthening Europe's cybersecurity", 11.12.2018. Online at: https://ec.europa.eu/commission/news/cybersecurity-act-2018-dec-11_en.

pertaining to cybersecurity, analyze what harms they aim to address and how, and point out pertinent issues legislators would have to devote further scrutiny on.

5. Conceptual shifts in EU cybersecurity policy

5.1 Initial place of cybersecurity concerns in EU legislation

The EU has demonstrated intensifying legislative activity in the field of network and information security since the early 2000's.[79] It was emphasized from the beginning that "security is becoming a key priority because communication and information have become a key factor in economic and societal development"[80] and many of the currently binding EU laws have their non-binding predecessors from 10-15 years ago addressed in the third pillar[81] of the EU[82].

Generally the provisions dealing with security in networks and information systems in early EU regulations had two main considerations: protection of privacy and personal data[83], and harmonizing requirements for the sake of completing the single

[79] The first instrument with specific focus on security was the Commission's, 26.1.2001, Communication (COM(2000) 890 final), 'Creating a Safer Information Society by Improving the Security of Information Infrastructures and Combating Computer-related Crime'.

[80] Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions – Network and Information Security: Proposal for A European Policy Approach /COM/2001/0298 final/.

[81] From 1993 until 2009 in the EU's 'three pillar system' the first pillar referred to economic, social and environmental policies; the second pillar stood for Common Foreign and Security Policy; and the third pillar consisted of Police and Judicial Cooperation in Criminal Matters.

[82] See for example in the field of fighting cybercrime Council Framework Decision 2005/222/JHA of 24 February 2005 on attacks against information systems, which was replaced by Directive 2013/40/EU of the European Parliament and of the Council of 12 August, 2013 on attacks against information systems.

[83] See for example *Directive 97/66/Ec of the European Parliament and of the Council* of 15.12.1997, concerning the processing of personal data and the protection of privacy in the telecommunications sector.

market. However, the establishment of ENISA sparked a debate on the conceptual framework of network and information security in the EU, which was considered by the EU's court,[84] and it held that these measures also form "part of a normative context circumscribed by the Framework Directive and the specific directives and directed at completing the internal market in the area of electronic communications". [85] Therefore it can be claimed that the EU's primary concern was data security, and the broader network and information or cybersecurity aspects were rather incidental in special legal regimes[86], having to do more with the completion of the internal market, than with the potential harms that can result from misuses or degradation of functioning of computer systems. These provisions set

[84] The legal basis for EU action in the 'first pillar' in the areas of network and information security has been addressed in case C-217/04 UK vs. EU Parliament and Council. More precisely, the establishment of ENISA by Reg. No 460/2004, its objectives and the tasks assigned to it by Regulation EC No. 460/2004 were regarded as measures for approximation in the meaning of Art. 114 of TFEU (ex Article 95 TEC).

[85] C-217/04 United Kingdom vs. European Parliament and Council, paras. 59-60.

[86] Several legal provisions were listed in the judgment that "express concern of the Community legislature in relation to network and information security". These included Article 8 (4) (c) and (f), framework dir. 2002/21/EC, which state the need for high level of protection of personal and privacy, as well for maintaining the integrity and security of public communications networks. The Authorization Directive 2002/20/EC briefly refers to security and personal data protection as part of those maximum conditions that may be attached to general authorization to provide electronic communication networks and services, and Article 23 of the Universal Service Directive 2002/22/EC refers to integrity and availability of public telephone services, in particular emergency services in cases of catastrophic events. More detailed provisions can be found in the e-Privacy Directive 2002/58/EC, which in Article 4 and 5 deals with network security and confidentiality of communications. Noteworthy in Article 4 that it requires service providers to take technical and organizational measures having regard to the state of the art, costs, appropriateness of measures and risks present, a language that reflect focus on prevention and will appear more prominently later and outside the narrow field of electronic communications. In addition to these, the Personal Data Protection Directive and the e-Signatures Directive also touched upon security issues within their specific contexts, in Article 17 and 3 (4) respectively. Certain other security aspects of digital assets, protection of intellectual property in the information society, are addressed by the EU's specialized regulatory regimes on copyrights, patents, database protection, etc.

out some vague and overall requirements for information and network security, but their scope was limited to the telecommunication sector, personal data protection and e-signatures. Therefore many information society services as they emerged fall outside the scope of these laws, such as most cloud services, search engines, e-marketplaces, internet telephony services, unless they were in the specific signal transmission business, which qualifies as electronic communications service for the purposes of the telecom regulations[87], or processed personal data and relevant data protection rules (eventually) came into play[88].

However, the legislative landscape has significantly changed since the first elements of cybersecurity-related provisions were put in place and whereas network and information security used to be understood as merely complementary to the electronic communications field, today the picture is more complex, in particular that cybersecurity is a broader concept than network and information security. Virtually the entire legal framework has already been revised and updated, yet a significant EU reform in cybersecurity has just begun. Today there are numerous legal instruments of the EU having a bearing on cybersecurity and several proposals are pending.

[87] Article 2 (c) of the Framework Directive defines that "electronic communications service" means a service normally provided for remuneration which consists wholly or mainly in the conveyance of signals on electronic communications networks, including telecommunications services and transmission services in networks used for broadcasting, but exclude services providing, or exercising editorial control over, content transmitted using electronic communications networks and services; it does not include information society services, as defined in Article 1 of Directive 98/34/EC, which do not consist wholly or mainly in the conveyance of signals on electronic communications networks.

[88] Although some provisions of the Personal Data Protection Directive needed clarifications by the courts, for example in the Google vs Spain case (Case C-131/12), popularly known as addressing the ‘right to be forgotten’.

5.2 Cybersecurity becomes a priority

In the context of the second pillar of common foreign and security policy the 2007 cyber-attacks against Estonia have led to a turning point, and cybersecurity was identified as a security issue in the report on the implementation of the European Security Strategy (ESS) submitted by SG/HR Javier Solana to the European Council in December 2008.[89] The term "cybersecurity" turned into a policy buzzword after the adoption of the EU's 2013 Cybersecurity Strategy[90] and cybersecurity is now an integral part of EU policies. The document addressed cybersecurity in a comprehensive fashion and foresaw that proposed activities would operate within different legal frameworks, notably network and information security, law enforcement and defence, and on two levels, the national and EU level.[91] It established five priorities: achieving cyber resilience; drastically reducing cybercrime; developing cyber-defence policy and capabilities related to the Common Security and Defence Policy (CSDP); develop the industrial and technological resources for cybersecurity; and establish a coherent international cyberspace policy for the European Union and promote core EU values. The 2013 strategy is centered mostly on the importance of cybersecurity for economic reasons, but also mentions some particular concerns, thereby implying what harms are considered: economic losses both in terms of damages and decreased productivity, decreased confidence of citizens to use e-services, physical and impalpable harms to citizens, and the loss of autonomy for citizens outside the EU.

[89] EEAS, "Report on the implementation of the European Security Strategy – Providing Security in a Changing World", 11.12.2018. Online at: https://europa.eu/globalstrategy/en/report-implementation-european-security-strategy-providing-security-changing-world.

[90] *Joint Communication To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions* Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace, JOIN/2013/01 final.

[91] Ibid. p. 17.

5.3 Raising the stakes: EU's new cybersecurity strategy

The overall strategy is currently formulated in the European Commission's Joint communication on Resilience, Deterrence and Defence: Building strong cybersecurity for the EU, which updated the 2013 strategy document.[92] The new vision is moving from the comprehensive approach towards a more integrated one, where economic, political and strategic threats enjoy equal attention, and cybersecurity can be seen as a horizontal policy issue, or a common societal challenge, having elements in multiple layers of government, economy and society. Therefore, the updated strategy goes beyond the previously stated areas of network and information security, cybercrime, cyber defence and external relations, and proposes measures in product liability, consumer protection, labour market, financial services, education, trade and investment fields as well. Emphasis is on building resilience and deliver better EU response to cyber-attacks, signifying a shift from a reactive to a proactive approach.

The threats outlined in the introduction part of the Communication imply that the EU is ready to address potential harms by different measures. The concern about negative economic impact of misuses and degradation in the functioning of computer systems is still central, however the issue has grown in magnitude and worries are expressed about potential economic destabilization, decreased political autonomy, disrespect for territorial integrity, physical harms, decrease in consumer trust and the decreased ability of states to provide order in the society by enforcing their laws.

In the next section we identify legal measures that are either already available or are proposed and, if and when adopted, can be used in the future to address the potential and actual harms identified so far.

[92] JOIN (2017) 450.

6. EU Cyberssceurity Laws

6.1 Information society laws and cyber resilience

6.1.1 Electronic Communications

In 2009 several provisions requiring operators of electronic communications networks and services to implement security measures were incorporated into the EU's Telecom regulatory framework. The "Better Regulation Directive" established a regime for undertakings providing public communications networks or publicly available electronic communications services imposing requirements to implement risk-based security management practices, state-of-the-art technical and organizational measures, as well as to notify national authorities of a breach of security or loss of integrity incidents with significant impact.[93]

The newly adopted European Electronic Communications Code (EECC) kept the underlying structure of the security regime, however now it clearly includes security of networks and services and end-user benefits[94], whereas the EECC also extends to services that fall outside scope of the previous framework[95], adds definitions of "security of networks and service" and "security inciden", and clarifies a number of important points on the breach notification obligations, roles and powers of authorities and relevant institutions. This brings a significant expansion of the EU's oversight on the electronic

[93] Articles 13a and 13b of *Directive 2009/140/Ec Of The European Parliament And Of The Council* of 25 November 2009, amending Directives 2002/21/EC on a common regulatory framework for electronic communications networks and services, 2002/19/EC on access to, and interconnection of, electronic communications networks and associated facilities, and 2002/20/EC on the authorization of electronic communications networks and services.

[94] Art. 1 (2) (a) of the EECC, Proposal for a *Directive of the European Parliament and of the Council, E*stablishing the European Electronic Communications Code, COM/2016/0590 final – 2016/0288 (COD).

[95] It was unclear whether i.e. if or to what extent internet telephony services or electronic processing services (for email service) fall under the regime. See for example Case C-142/18, Case C-193/18.

communications field, which will cover not only those service providers that operate the core communication infrastructures, but also those that built up new business models relying on the core infrastructures for the provision of their services, but not engaging in the "signal conveyance" business, hence did not fit the definition in Article 2 (c) of the Framework Directive[96]. The EECC redefines the meaning of "electronic communication service", now expressly including internet access service, interpersonal communication services (both number-based and number independent ones), as well as traditional signal conveyance.[97] This will result in higher-level security requirements imposed on a new layer of service providers in the field of communications in the EU, filling another gap in cybersecurity-related legislation. The EECC is to be implemented by the end of 2020.

6.1.2 Electronic Signatures and Trust Services

Significant piece of the cybersecurity puzzle lays with the eIDas Regulation[98] that replaced the 1999 e-signatures directive. It is hard to overestimate the role of the eIDas Regulation, since it lays down the foundations for mutual recognition and assessment of electronic identification or eID means, and it also defines assurance levels, i.e. criteria for assigning a degree of confidence for claimed or asserted

[96] According to Directive 2002/21/EC, the Framework Directive, "electronic communications service" means a service normally provided for remuneration which consists wholly or mainly in the conveyance of signals on electronic communications networks, including telecommunications services and transmission services in networks used for broadcasting, but exclude services providing, or exercising editorial control over, content transmitted using electronic communications networks and services; it does not include information society services, as defined in Article 1 of Directive 98/34/EC, which do not consist wholly or mainly in the conveyance of signals on electronic communications networks.

[97] Article 2 (4) of the EECC.

[98] Regulation (EU) No. 910/2014 of the European Parliament and of the Council of 23 July, 2014, on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC.

identity of persons by electronic identification means.[99] The second part of the eIDAS Regulation details the conditions and requirements for providing various trust services.[100] These trust services serve as points of reference for digital security, which include creation and verification of electronic signatures, website authentication, guaranteeing the origin and integrity of electronic seals, electronic time-stamps, etc. The Regulation establishes security requirements for trust service providers, referring to technical and organizational measures and risk-based approach[101], as well as breach notification obligations similarly to the EECC and NIS Directive. Our societies need reliable authentication and e-identification just as much as anonymity in cyberspace.

6.1.3 ISP liability

The e-commerce directive provides another pillar in cybersecurity-related legislation, more precisely it exempts intermediary service providers from liability for information transmitted, based on their neutral role[102]. Furthermore ISP's are not obliged to monitor their services and seek for illegal activity therein. The limits of this framework have been elaborated on in a series of court cases[103] and additional self-regulatory arrangements were established by concerned service providers in order to bridge the disconnect between illegal content online and enforcement mechanisms. However, the current regime is increasingly difficult to sustain, as these services can

[99] Chapter II of *Regulation (Eu) No. 910/2014 of the European Parliament and of the Council* of 23 July, 2014, on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC.

[100] Chapter III of the *Regulation (Eu) No 910/2014 of the European Parliament and of the Council* of 23 July, 2014, on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC.

[101] Article 19.

[102] Articles Directive 2000/31/EC of the European Parliament and of the Council of 8 June, 2000, on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market.

[103] See for example ECJ cases Google vs Louis Vuitton and the others Joined Cases C-2366/08 and C-238/08; L'Oreal vs eBay Case C-324/09; Delfi vs Estonia at ECHR.

be abused by third parties and presence of illegal content online has serious consequences for users, potentially for societies.

There is abundance of illegal material online, and media frequently reports on one or another ISP failing to remove such content.[104] Illegal material can come in different forms and shapes, can range from copyright-infringing audiovisual media, hate-speech and information relating to terrorism, and child exploitative content. Recent EU legislation qualifies the liability exemption regime and accommodates the particularities of different illegal content online. Specific responses were designed in this respect, for example amending the Audiovisual Media Services Directive[105] and in Chapter XIa of setting forth rules particularly addressed to video-sharing platform services to protect the public from harmful material online, practically imposing an obligation on these providers to apply proactive measures to identify illegal activity and content online, albeit also encouraging co- and self-regulation. In addition the Commission has issued a recommendation to support this policy.[106] However the EU is drawing some red-lines in this field, since clear-cut rules were proposed in 2018 for cases when service providers have been informed about illegal activity, including the obligation of hosting service providers to remove terrorist content or disable access to it within one hour from receipt of a removal order issued by a competent authority.[107] These examples demonstrate the ongoing

[104] See for example BBC report on Facebook failing to remove child exploitation images https://www.bbc.com/news/technology-39187929 or Business insider report on You tube's slow reaction to notifications about illegal content https://www.businessinsider.com/youtube-purges-over-400-channels-millions-of-videos-to-address-child-exploitation-concerns-2019-2.

[105] Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November, 2018, amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities PE/33/2018/REV/1.

[106] Commission Recommendation of 1 March, 2018, on measures to effectively tackle illegal content online (C (2018) 1177 final).

[107] Proposal for a *Regulation of the European Parliament and of the Council* on preventing the dissemination of terrorist content online COM (2018) 640 final.

policy shift, where the liability exemptions of the neutral gatekeepers are curtailed – short of a better solution to address the proliferation of illegal online content.

6.1.4 Consumer protection

Generally the EU's consumer protection framework does not address problems of cybersecurity in specific terms, however some sporadic provisions already require that certain products are constructed so as to ensure the protection of personal data and privacy of users and subscribers. Article 3 and 4 of the Radio Equipment Directive contain broad requirements for data security for connected consumer products, such as smart watches, connected toys[108], drones, etc., however its operational range is still unclear.[109] Issues of basic encryption, software updates, weak or lack of authentication in connected consumer products, and product liability remain highly-debated open questions despite initiatives in this field.[110]

6.1.5 Payment Services

Among the sectoral measures the Second Payment Services Directive (PSD2)[111] should also be mentioned as contributing to building strong cybersecurity in Europe. In the PSD2 an additional element, strong customer authentication is emphasized[112], besides the risk-based management and incident reporting obligations imposed on payment

[108] See for example a recent security alert for childrens' smart watches. Online at: https://ec.europa.eu/consumers/consumers_safety/safety_products/rapex/alerts/?event=viewProduct&reference=A12/0157/19&lng=en.

[109] European Commission, *Report from the Commission to the European Parliament and the Council* on the operation of the Radio Equipment Directive, 2014/53/EU, COM(2018), 740 final.

[110] See for example Proposal for a *Directive of the European Parliament and of the Council* on certain aspects concerning contracts for the supply of digital content, and the Commission is also reviewing the Product Liability Directive (Directive 85/374/EEC)

[111] *Directive (Eu) 2015/2366 of the European Parliament and of the Council* of 25 November, 2015, on payment services in the internal market, amending Directives 2002/65/EC, 2009/110/EC and 2013/36/EU and Regulation (EU) No. 1093/2010, and repealing Directive 2007/64/EC.

[112] Articles 97 and 98.

service providers. The European Banking Authority is currently working on a draft for regulatory technical standards on strong customer authentication and common and secure communication under Directive 2015/2366 (PSD2).[113]

6.1.6 Personal Data Protection

In several EU regulatory instruments preventive measures are dominant, paying less attention to incident response, recovery and business continuity aspects. The General Data Protection Regulation[114] can be seen as a cybersecurity instrument, essentially aiming to prevent misuses of personal data by imposing heavy limitations on their processing in the first place. Additionally the GDPR dedicates Article 32-34 to security of personal data, setting forth technical requirements and a breach notification regime. In this context, the Police Directive[115] applies a similar approach, prescribes security measures and notification obligations, however the scope is different [116] and it is complementary to the GDPR, within the competences of the EU. The above instruments, however say little about responding to security incidents and recovery from them. These aspects are apparently left for the particular organizations implicated and to standards to be applied.

[113] See online at: https://eba.europa.eu/documents/10180/1761863/Final+draft+ RTS +on+SCA+and+CSC+under+PSD2+%28EBA-RTS-2017-02%29.pdf.

[114] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April, 2016, 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.

[115] *Directive (Eu) 2016/680 of the European Parliament and of the Council* of 27 April, 2016, 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.

[116] The Directive applies 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, including the safeguarding against and the prevention of threats to public security.

6.1.7 High Common Level Network and Information Security

Slightly more concern is given to incident response in the EU's first cybersecurity law, the Network and Information Security Directive[117], which obliges Member States to adopt national strategies for network and information security, aims to establish appropriate structures for national level management and cooperation among these in the EU, as well as it imposes important security requirements for operators of essential services and digital service providers. Although the NIS Directive signifies an important effort for harmonization in the field of cybersecurity, its effects are expected to be far weaker than it was intended in the original proposal put forth by the Commission, since public sector information systems as well as a portion of providers of information society services have been excluded from its scope and cooperation measures, including information sharing mechanisms, were reduced to the very minimum based on voluntary action by member states.[118]

This regulatory framework leaves the question of response and recovery aspects mainly open, however the European Commission has issued a Recommendation that serves as a blueprint for action in case of cyber incidents with EU-wide effects.[119] This plan was tested during the Wannacry incident first time, with reportedly positive results[120] and the case pointed out how important cooperation in the area of cybersecurity is. Yet cooperation in incident response is just one piece of the puzzle, as the 'non-incident' of the ROCA

[117] *Directive (Eu) 2016/1148 of the European Parliament and of the Council* of 6 July, 2016, concerning measures for a high common level of security of network and information systems across the Union.

[118] Compare the current NIS Directive to the Commission Proposal for a *Directive of the European Parliament and of the Council Co*ncerning measures to ensure a high common level of network and information security across the Union/* COM/2013/048 final – 2013/0027 (COD) *.

[119] Commission Recommendation (EU) 2017/1584 of 13 September, 2017, on co-ordinated response to large-scale cybersecurity incidents and crises, C/2017/6100.

[120] Online at: https://www.bna.com/wannacry-provided-first-n73014451505/.

vulnerability discovery caused a crisis situation in Estonia.[121] The Estonian experience with the naturally constrained flow of research and scientific information also makes the case for the establishment of European level network in this area.[122]

6.1.8 Cybersecurity Act

However, the European plans are more ambitious and cybersecurity is elevated to a significant policy issue, which requires appropriate coordination and enforcement mechanisms. The Commission has proposed the Cybersecurity Act, establishing a permanent mandate for the EU Cybersecurity Agency and a framework for cybersecurity certification.[123] The Explanatory Memorandum of the Cybersecurity Act mentions a number of policy areas, sectors and refers to legal acts, where the EU's Cybersecurity Agency (currently ENISA) will have assigned tasks.

These include, naturally the policy area of network- and information security, but also sectors with "cybersecurity element", such as

[121] In 2017 the discovery of a vulnerability in the chips used in the Estonian ID-card led to serious concerns about the security of the infrastructure underlying the Estonian digital state. Although no security breaches or misuses were identified, the case pointed out some shortcomings in preparedness and unknown societal dependencies on current technologies. To mention a few points, the concentration of critical competences into a small number of experts and the unexpected dependency of the public sector and critical infrastructures on the ID-card for the performance of their tasks were brought to light. For an overview see online at: https://www.ria.ee/sites/default/files/content-editors/kuberturve/roca-vulnerability-and-eid-lessons-learned.pdf or the more detailed Estonian version online at: https://www.ria.ee/sites/default/files/content-editors/EID/id-kaardi_oppetunnid.pdf.

[122] Proposal for a *Regulation of the European Parliament and of the Council* establishing the European Cybersecurity Industrial, Technology and Research Competence Centre and the Network of National Coordination Centres. A contribution from the European Commission to the Leaders' meeting in Salzburg on 19-20 September, 2018, COM/2018/630 final.

[123] Proposal for a *Regulation of the European Parliament and of the Counci*l on ENISA, the "EU Cybersecurity Agency", and repealing Regulation (EU) 526/2013, and on Information and Communication Technology cybersecurity certification ("Cybersecurity Act").

finance, transport, energy.[124] It is also foreseen that the Agency will support policy and law in electronic communications, electronic identity and trust services. The Network and Information Security (NIS) Directive has already been expressly tied to ENISA, entrusting it the coordination of CyberEurope cycle of exercises with Member States, assisting the Member States and the Commission with expertise, advice, guidelines and facilitating the exchange of best practices.[125] ENISA also has significant role in assisting in the implementation of legal and regulatory requirements of network and information security arising from the NIS Directive or any other legal act,[126] as well as it will report on the implementation of the EU legal framework. ENISA will be tasked to prepare a candidate European cybersecurity certification scheme. This process should result in establishing points of reference for the ‘’duty of care“ principle and lead to the application of ‘’security by design and default“ approach by producers of connected devices.

6.2 EU legal acts and cyber deterrence and defence

Since the adoption of the 2013 EU’s Cybersecurity Strategy the legal framework tackling cybercrime has improved across the EU, whereas the substantive part of the Council of Europe Cybercrime Convention was practically implemented via the “Botnet Directiv”.[127] The Directive does not address questions of self-defense and remedies for

[124] Proposal for a *Regulation of the European Parliament and of the Council* on ENISA, the “EU Cybersecurity Agency”, and repealing Regulation (EU) 526/2013, and on Information and Communication Technology cybersecurity certification (“Cybersecurity Act”) COM(2017) 477 final, 13 September, 2017, p. 7.

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

[126] Article 2 (3) of *Regulation (EU) No 526/2013 of the European Parliament and of the Council* of 21 May, 2013, concerning the European Union Agency for Network and Information Security (ENISA) and repealing Regulation (EC) No 460/2004.

[127] Directive 2013/40/EU of the European Parliament and of the Council of 12 August, 2013, on attacks against information systems and replacing Council Framework Decision 2005/222/JHA.

victims. The Botnet Directive is also complemented by another directive on combating sexual abuse and exploitation of children and child pornography.[128] Although there are still some open questions on implementation of the above Directives[129], the procedural and cooperation aspects of fighting cybercrime proved to be more controversial.

One of the major failures of EU legislators has been the Data Retention Directive[130], which was cancelled by the European Court of Justice due to its disproportionate measures obliging service providers to collect data on electronic communications. Since investigation, detection and prosecution of serious crimes is rather difficult when electronic communications data is unavailable or erased, imposing data retention obligations in the electronic communications sector appeared a reasonable step. However, in the Digital Rights Ireland case the Court pronounced that

> "[a]s regards the necessity for the retention of data required by Directive 2006/24, it must be held that the fight against serious crime, in particular against organised crime and terrorism, is indeed of the utmost importance in order to ensure public security and its effectiveness may depend to a great extent on the use of modern investigation techniques. However, such an objective of general interest, however fundamental it may be, does not, in itself, justify a retention measure such as that established by Directive 2006/24 being considered to be necessary for the purpose of that fight".[131]

[128] Directive 2011/92/EU *of the European Parliament and of the Council* of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA

[129] National transposition measures communicated by the Member States concerning: Directive 2013/40/EU of the European Parliament and of the Council of 12 August 2013 on attacks against information systems and replacing Council Framework Decision 2005/222/JHA. Online at: https://eur-lex.europa.eu/legal-content/EN/NIM/?uri=celex:32013L0040.

[130] Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC.

[131] Joined Cases C-293/12 and C-594/12, Digital Rights Ireland and Seitlinger and Others, para 51.

The judgement opened the door for EU-wide fragmentation of data retention regulations, some Member States keeping their relevant national rules, some cancelling them, which also led the European Court to provide further guidance in two consecutive cases addressing details of and conditions of data retention.[132] However, pending the proposal for the e-Privacy Regulation and discussions on data retention ongoing in EU institutions, coupled with the strong requirements of the GDPR, which has already proven to be an obstacle for information sharing with entities outside the EU[133], the fate of the EU's data retention regime appears to be still uncertain.

Yet, rules on collection of data in cyberspace for the purposes of investigations and evidence remained a central issue, including for the purposes of attributing cyber-attacks to perpetrators. Just after the adoption of the US CLOUD Act[134], which confers jurisdiction on the US authorities to request data held overseas from US companies, the EU has came up with its e-Evidence proposals to create a European Production Order and a European Preservation Order[135], including allegedly strong, but controversial safeguards[136], as well as to oblige service providers to designate a legal representative in the Union for the purposes of the legislation. In addition, the Commission has presented further proposals, including one addressing fraud and counterfeiting of non-cash means of payments, extending the scope

[132] Joined Cases C-203/15 and C-698/15, Tele2 Sverige and Case C-207/16 Ministerio Fiscal .

[133] See for example European Data Protection Board, Letter to ICANN, 05 July, 2018. Online at: https://edpb.europa.eu/news/news/2018/letter-icann_en.

[134] US, Clarifying Lawful Overseas Use of Data Act. Online at: https://docs.house.gov/billsthisweek/20180319/BILLS-115SAHR1625-RCP115-66.pdf#page=2201.

[135] Proposal for a *Regulation of the European Parliament and of the Council* on European Production and Preservation Orders for electronic evidence in criminal matters COM/2018/225 final – 2018/0108 (COD).

[136] For example Article 9 (5) of the proposal allows that private entities assess compliance with the Charter of Fundamental Rights of the European Union and object to cooperation on this ground.

of measures to virtual currencies.[137] The 2017 EU cybersecurity strategy addresses the question of deterrence as a mainly technical and capability issue, focussing on attribution, IPv6, forensic procedures and investigative capabilities of Member States' law enforcement authorities. In a recent initiative, the four EU cybersecurity organisations, ENISA, the European Defence Agency (EDA), the European Cybercrime Centre (EC3) and the Computer Emergency Response Team for the EU Institutions, Agencies and Bodies (CERT-EU) also signed a Memorandum of Understanding with a view to fostering cooperation and facilitating information exchange between the agencies.[138] In addition private-public cooperation is emphasized, but this overflows to the section dealing with cyber defence and external dimensions of cybersecurity – not without a point, since several global cases have already demonstrated the importance of cooperation between the private and public sectors.[139]

This leads us to the sphere of the EU, where coherence and common action is yet scarce: defence and international relations, the Common Security and Defence Policy. However, the EU has made significant steps in these areas approving the Cyber Diplomacy Toolbox[140], putting forward technology control proposals[141] and concerns for the

[137] Proposal for a *Directive of the European Parliament and of the Council* on combating fraud and counterfeiting of non-cash means of payment and replacing Council Framework Decision 2001/413/JHACOM/2017/0489 final – 2017/0226 (COD).

[138] General Secretariat of the Council, *EU Coordinated Response to Large-Scale Cybersecurity Incidents and Crises - Council conclusions*, 100086/18, 26 June, 2018, at 3.

[139] One of the first global cases include the spread of the Conficker worm, where the counter-action and clean-up initiatives were mainly rooted in the private sector.

[140] Council of the European Union, "Cyber Diplomacy Toolbox", 07 July, 2017, 7923/2/17 REV 2. Online at: http://data.consilium.europa.eu/doc/document/ST-9916-2017-INIT/en/pdf.

[141] European Commission, "Proposal for a *Regulation of the European Parliament and of the Council* setting up a Union regime for the control of exports, transfer, brokering, technical assistance and transit of dual-use items (recast)," COM (2016) 616 final, September 28, 2016. Online at: http://ec.europa.eu/transparency/regdoc/?fuseaction=list&coteId=1&year=2016&number=616&version=ALL&language=en.

origins of foreign direct investments[142]. Cybersecurity is also overlapping with other policy areas, such as countering hybrid threats[143] or development policy[144]. Although the EU has initiated cooperation and is engaged with international actors in discussing cybersecurity, significant legal measures currently adopted in this area are few.[145]

7. Conclusions

This paper has outlined some of the main cybersecurity legal challenges the EU is facing nowadays. Cybersecurity is an issue that will remain in the focus of the Member States and the EU, it will not be solved or go away miraculously. Yet, looking around ourselves, as users, members of organizations, people entrusted with carrying out societal functions, we should notice that we indeed depend on computer systems, which are not perfect and will never be. Yet, this dependency and inherent insecurity can be handled and managed, including by using legal tools, since cyberspace is human-created environment and serves human needs.

We reasoned that EU cybersecurity laws aims to protect not only confidentiality, integrity and availability of data, information systems and networks, but also certain interactions with these by the society. Although it is somewhat unclear what types of harms EU laws aim to prevent, hence it is difficult to assess what interactions should be in

[142] European Commission, "Proposal for a *Regulation of the European Parliament and of the Council* establishing a framework for screening of foreign direct investments into the European Union," COM (2017), 487 final, September 13, 2017. Online at: https://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-487-F1-EN-MAIN-PART-1.PDF

[143] *Joint Communication to the European Parliament and the Council* Joint Framework on countering hybrid threats a European Union response JOIN/2016/018 final.

[144] SWD (2017) 157 final, Commission Staff Working Document, Digital 4 Development: mainstreaming digital technologies and services into EU Development Policy.

[145] Rehrl, Jochen, European Security and Defense College, Federal Ministry of Defence of the Republic of Austria, "Handbook on Cyber Security", 2018. Online at: https://publications.europa.eu/en/publication-detail/-/publication/63138617-f133-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-81357173.

focus, we were able to observe that the policy framework developed from a protecting business interests and personal data to a more inclusive one eventually being concerned with harms to economic interests, individuals and national security. The potential harms include direct economic losses, decreased productivity, reputational damage, decreased consumer trust, physical and impalpable harm to citizens, but also economic destabilization, decreased ability to provide order in the society, decreased political autonomy, and losses in sovereignty.

Binding and stringent EU cybersecurity-related laws concern those private infrastructures that are at the core for the operation of cyberspace, in the electronic communications sector, as well as those that support the delivery of essential services for the society. Specific, cross-sectoral regulations regarding personal data protection also contribute to achieve cybersecurity aims in the EU and illegal or harmful content enjoys increasing attention from the EU regulators, generally raising the stakes for actors in the private sector in terms of liability. However, there are certain gaps and while implementation of security measures in the context of personal data processing extends to both private and public sectors, there are no EU level requirements to implement high-level network and information security measures in public administrations and for businesses other than the few listed in the NIS Directive. Social networks, app-stores, and most SME's, unless they are involved with the supply chain for those covered by the NIS Directive, fall outside the scope of the Directive. The EU also applies regulations that are coercive in nature in countering cybercrime as well as for establishing organizational structures in this field.

We can see from the regulatory choices that the EU does not impose strong authentication requirements easy-handedly and opts for alternative solutions, ultimately favouring user anonymity in other fields than payment services. The authors are inclined to attribute this choice to the fact that the functioning of the European society, as such, is less reliant on computerized systems for it basic functions, and it is rather some individual Member States and certain sectors[146] where deep dependencies exist, which can justify the dominantly soft touch approach from EU level. Although strong authentication in general would presumably contribute to building trust in e-services, by making the case for misuse harder (one can just imagine the impact of strong authentication for the use of social networks, for example), this neither would solve all the problems nor markets seem to be ready for such steps.

Soft and collaborative instruments, voluntary and alternative measures are chosen by the EU for supporting and facilitating cooperation and information exchange among Member States. However, some hard law instruments are used when it comes to information flowing from private sector to public authorities, i.e. incident reporting obligations. These obligations do not extend to “non-incident”, such as vulnerability discovery and disclosure, which are targeted by standardization efforts in the EU. Soft measures are applied for EU level coordination of responses in crisis situations, including large-scale cyber-attacks.

In the last few years the EU has made a great deal of progress in switching gears and moving from the reactive policy towards preventive and proactive approach in cybersecurity. In particular the adoption of the NIS Directive reflects this forward-looking nature of EU cybersecurity laws, which now oblige a range of actors to actually implement security measures, and do not leave room for alternative market-driven solutions (such as raising the prices of services/goods to compensate for the risks, or seeking insurance coverage, etc). However, there is little EU level guidance on private sector responses to cyber incidents, and recovery and business continuity aspects. The preventive approach is also visible in EU efforts to channel industry towards the adoption of “security by desig” practices and elaborating the content of “duty of car” principle. Yet, this way of thinking is not clearly identifiable when looking at the public sector and cooperation

[146] Such as the financial sector and Critical Infrastructure Protection.

among Member States. EU legal instruments dealing with Member States' own and common effort to address cybersecurity challenges remain dominantly backward-looking, focusing on coordination of crisis response, imposition of criminal penalties, as well as political responses to cyber-attacks.

Although most of the challenges are global, the EU appears to be internally focusing, emphasizing technological solutions. The EU's approach to cybersecurity is centered on technological solutions for a good reason, however more attention should be paid to social and human aspects, as well as to higher level commitment to common standards and joint action, in particular collective preventive action, keeping in mind the potential harms that cybersecurity laws should address.

Das **Zentrum für Europäische Integrationsforschung (ZEI)** ist ein interdisziplinäres Forschungs- und Weiterbildungsinstitut der Universität Bonn. *ZEI – Discussion Paper* richten sich mit ihren von Wissenschaftlern und politischen Akteuren verfassten Beiträgen an Wissenschaft, Politik und Publizistik. Sie geben die persönliche Meinung der Autoren wieder. Die Beiträge fassen häufig Ergebnisse aus laufenden Forschungsprojekten des ZEI zusammen.

The **Center for European Integration Studies (ZEI)** is an interdisciplinary research and further education institute at the University of Bonn. *ZEI – Discussion Paper* are intended to stimulate discussion among researchers, practitioners and policy makers on current and emerging issues of European integration and Europe´s global role. They express the personal opinion of the authors. The papers often reflect on-going research projects at ZEI.

Die neuesten ZEI Discussion Paper / Most recent ZEI Discussion Paper:

C 239 (2017) Michael Gehler
Revolutionäre Ereignisse und geoökonomisch-strategische Ergebnisse: Die EU- und NATO-„Osterweiterungen" 1989-2015 im Vergleich
C 240 (2017) Tapio Raunio/Matti Wiberg
The Impact of the European Union on National Legislation
C 241 (2017) Robert Stüwe
EU External Energy Policy in Natural Gas: A Case of Neofunctionalist Integration?
C 242 (2017) Ludger Kühnhardt
Weltfähig werden. Die Europäische Union nach dem Biedermeier
C 243 (2017) César Castilla
Perspectives on EU-Latin American Cooperation: Enhancing Governance, Human Mobility and Security Policies
C 244 (2017) Joe Borg
The Maltese Presidency of the European Union 2017
C 245 (2018) Ludger Kühnhardt
The New Silk Road: The European Union, China and Lessons Learned
C 246 (2018) Teodora Lađić
The Impact of European Integration on the Westphalian Concept of National Sovereignty
C 247 (2018) Wolfgang Reinhard
Die Expansivität Europas und ihre Folgen
C 248 (2018) Joseph M. Hughes
"Sleeping Beauty" Unleashed: Harmonizing a Consolidated European Security and Defence Union
C 249 (2018) Rahel Hutgens/Stephan Conermann
Macron's Idea of European Universities. From Vision to Reality
C 250 (2018) Javier González López
Bosnia and Herzegovina: a Case Study for the Unfinished EU Agenda in the Western Balkans
C 251 (2019) Günther H. Oettinger
Europäische Integration aus historischer Erfahrung. Ein Zeitzeugengespräch mit Michael Gehler
C 252 (2019) Chiara Ristuccia
Industry 4.0: SMEs Challenges and Opportunities in the Era of Digitalization
C 253 (2019) Agnes Kasper/Alexander Antonov
Towards Conceptualizing EU Cybersecurity Law

Die vollständige Liste seit 1998 und alle Discussion Paper zum Download finden Sie auf unserer Homepage: http://www.zei.de. For a complete list since 1998 and all Discussion Paper for download, see the center's homepage: http://www.zei.de.

Curriculum vitae

Personal data

Name:	Alexander Antonov
Date of birth:	29.07.1993
Place of birth:	Brest, Belarus
Citizenship:	German

Contact data

E-mail:	alexander.antonov@taltech.ee

Education

2018–2024	Tallinn University of Technology – PhD in Public Administration
2016–2018	University of Southern Denmark – MSc in International Security and Law – among the best legal Master's Theses
2013–2016	University of Southern Denmark – BSc in European Studies – *summa cum laude*
2004–2013	Heinrich-Heine-Gymnasium Heikendorf – Abitur (High School) – i.e. invited to the Enrichment Program of the Federal State of Schleswig-Holstein for gifted pupils (2007/2008)

Language competence

German	Native
English	Fluent
Danish	Intermediate (B2 certificate)
Estonian	Pre-Intermediate (A2 certificate)
Russian	Intermediate
Spanish	Basic

Professional employment and internships

2022–2023	Tallinn University of Technology – Project Assistant
2017–2017	German Embassy Tallinn – Internship
2017–2017	Estonian Atlantic Treaty Association – Internship

Publications (selected)

Antonov, A. (2022). Managing Complexity: The EU's Contribution to Artificial intelligence Governance. *Revista CIDOB d'Afers Internacionals*, (131), 41-65. https://doi.org/10.24241/rcai.2022.131.2.41/en

Antonov, A., Häring, T., Korõtko, T., Rosin, A., Kerikmäe, T., & Biechl, H. (2021). Pitfalls of Machine Learning Methods in Smart Grids: A Legal Perspective. In *Proceedings - 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, 12-14 November 2021, Rome (pp. 248-256). Institute of Electrical and Electronics Engineers (IEEE). Danvers. https://doi.org/10.1109/ISCSIC54682.2021.00053

Antonov, A., & Kerikmäe, T. (2020). Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU. In D. R. Troitiño, T. Kerikmäe, R. M. de la Guardia, & G. Á. P. Sánchez (Eds.), *The EU in the 21st Century: Challenges and Opportunities for the European Integration Process* (pp. 135–154). Springer. https://doi.org/10.1007/978-3-030-38399-2_9

Kasper, A., & **Antonov, A**. (2019). Towards Conceptualizing EU Cybersecurity Law. In *ZEI Discussion Paper*, C253. Bonn: Center for European Integration Studies [Zentrum für Europäische Integrationsforschung]. https://hdl.handle.net/20.500.11811/9849

Antonov, A. (2018). Russia's Aggression Against Ukraine: State Responsibility, Individual Responsibility and Accountability [Master's thesis, University of Southern Denmark]. JuSDUs https://www.sdu.dk/da/om-sdu/institutter-centre/juridisk-institut/jusdus/2018

Courses taught

2020–2022	Human Rights, Ethics and Technology (MOT5010)
2019–2023	Introduction to Criminal Law and Procedure (HOA6032)
2018–2023	Moot Court (HOX6040)

Elulookirjeldus

Isikuandmed

Nimi:	Alexander Antonov
Sünniaeg:	29.07.1993
Sünnikoht:	Brest, Valgevene
Kodakondsus:	Saksa

Kontaktandmed

E-post:	alexander.antonov@taltech.ee

Hariduskäik

2018–2024	Tallinna Tehnikaülikool – PhD (Avalik haldus)
2016–2018	Lõuna-Taani Ülikool – MSc Rahvusvahelise julgeoleku ja õiguse alal – parimate juriidiliste magistritööde hulgas
2013–2016	Lõuna-Taani Ülikool – BSc Euroopa õpingute – *summa cum laude*
2004–2013	Heikendorfi Heinrich-Heine-Gümnaasium – Keskharidus – kutsutud Schleswig-Holsteini liidumaa andekate õpilaste rikastamise programmi (2007/2008)

Keelteoskus

Saksa keel	emakeel
Inglise keel	kõrgtase
Taani keel	kõrgem kesktase (B2 tunnustus)
Eesti keel	kõrgem algtase (A2 tunnustus)
Vene keel	kesktase
Hispaania keel	algtase

Teenistuskäik

2022–2023	Tallinna Tehnikaülikool – Projekti assistent
2017–2017	Saksamaa Liitvabariigi Suursaatkond Eestis – Praktika
2017–2017	NATO Eesti Ühing – Praktika

Publikatsioonid (valitud)

Antonov, A. (2022). Managing Complexity: The EU's Contribution to Artificial intelligence Governance. *Revista CIDOB d'Afers Internacionals*, (131), 41-65. https://doi.org/10.24241/rcai.2022.131.2.41/en

Antonov, A., Häring, T., Korõtko, T., Rosin, A., Kerikmäe, T., & Biechl, H. (2021). Pitfalls of Machine Learning Methods in Smart Grids: A Legal Perspective. In *Proceedings - 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, 12-14 November 2021, Rome (pp. 248-256). Institute of Electrical and Electronics Engineers (IEEE). Danvers. https://doi.org/10.1109/ISCSIC54682.2021.00053

Antonov, A., & Kerikmäe, T. (2020). Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU. In D. R. Troitiño., T. Kerikmäe, R. M. de la Guardia, & G. Á. P. Sánchez (Eds.), *The EU in the 21st Century: Challenges and Opportunities for the European Integration Process* (pp. 135−154). Springer. https://doi.org/10.1007/978-3-030-38399-2_9

Kasper, A., & **Antonov, A**. (2019). Towards Conceptualizing EU Cybersecurity Law. In *ZEI Discussion Paper*, C253. Bonn: Center for European Integration Studies [Zentrum für Europäische Integrationsforschung]. https://hdl.handle.net/20.500.11811/9849

Antonov, A. (2018). Russia’s Aggression Against Ukraine: State Responsibility, Individual Responsibility and Accountability [Master’s thesis, University of Southern Denmark]. JuSDUs https://www.sdu.dk/da/om-sdu/institutter-centre/juridisk-institut/jusdus/2018

Õppetöö

2020–2022	Human Rights, Ethics and Technology (MOT5010)
2019–2023	Introduction to Criminal Law and Procedure (HOA6032)
2018–2023	Moot Court (HOX6040)

ISSN 2585-6901 (PDF)
ISBN 978-9916-80-203-8 (PDF)