



RESEARCH
ARTICLE



OPEN
ACCESS



PEER
REVIEWED

Balancing public interest, fundamental rights, and innovation: The EU's governance model for non-high-risk AI systems

Michael Gille *Hamburg University of Applied Sciences*

Marina Tropmann-Frick *Hamburg University of Applied Sciences*

Thorben Schomacker *Hamburg University of Applied Sciences*

DOI: <https://doi.org/10.14763/2024.3.1797>

Published: 30 September 2024

Received: 18 February 2024 **Accepted:** 3 April 2024

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Licence: This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 License (Germany) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. <https://creativecommons.org/licenses/by/3.0/de/deed.en>
Copyright remains with the author(s).

Citation: Gille, M., Tropmann-Frick, M., & Schomacker, T. (2024). Balancing public interest, fundamental rights, and innovation: The EU's governance model for non-high-risk AI systems. *Internet Policy Review*, 13(3). <https://doi.org/10.14763/2024.3.1797>

Keywords: Artificial intelligence, Co-regulation, Self-regulation, Codes of conduct, AI Act

Abstract: The question of the concrete design of a fair and efficient governance framework to ensure responsible technology development and implementation concerns not only high-risk artificial intelligence systems. Everyday applications with a limited ability to inflict harm are also addressed. This article examines the European Union's approach to regulating these non-high-risk systems. We focus on the governance model for these systems established by the Artificial Intelligence Act. Based on a doctrinal legal reconstruction of the rules for codes of conduct and considering the European Union's stated goal of achieving a market-oriented balance between innovation, fundamental rights, and public interest, we explore our topic from three different perspectives: an analysis of specific regulatory components of the governance mechanism is followed by a reflection on ethics and trustworthiness implications of the EU's approach and concluded by an analysis of a case study from an NLP-based, language-simplifying artificial intelligence application for assistive purposes.

This paper is part of AI systems for the public interest, a special issue of *Internet Policy Review* guest-edited by Theresa Züger and Hadi Asghari.

1. Introduction

With its Artificial Intelligence Act (AI Act), the EU seeks to foster innovation while ensuring the risk-conscious protection of public interest and fundamental rights (Art. 1 (1) AI Act, Recitals 1 seq., 6 seq.). As Artificial Intelligence (AI) enters the mainstream, the need for effective and equitable governance mechanisms to safeguard responsible technology development and deployment becomes increasingly pressing. This is true not only for high-risk AI systems, such as biometric identification and emotion recognition, but also for less conspicuous *everyday applications* with lower risk profiles due to their limited potential to cause harm. This article scrutinises the European Union's approach to regulating these non-high-risk AI systems, focusing on the AI Act's (AI Act)¹ governance model for systems classified in this category.

By introducing voluntary compliance requirements with codes-of-conduct for non-high-risk systems, the EU aims to flexibly align its risk-oriented approach to key values and objectives without hampering the technological progress of these non-high-risk AI systems. Starting from the European Union's declared intention to strike a market-oriented balance between public interest, fundamental rights, and innovation (AI Act, Recitals 1, 2, 5), we tackle our subject from three angles: a governance perspective on the policy approach, an AI4SG-informed (AI for Social Good) viewpoint, and we analyse a case study from the field of AI for assistive applications. We argue that the initial criticism in the context of an analysis of the first draft that this class of AI systems is under-regulated (Stuurman & Lachaud, 2022) should, in consideration of the final version of the AI Act, give way to a more differentiated view. We also show that the approach in the AI Act, despite an at first glance binary categorisation of risks (high risk or not high risk) to fundamental rights (Kusche, 2024), is likely to create some room for risk-proportional voluntary co-regulation.

Following a brief doctrinal legal reconstruction of the rules for non-high-risk AI models (Section 2), we set out to examine the EU's governance approach for AI systems not posing high risks, with codes of conduct playing a central role (Section

1. This paper is based on the Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. See entry under Regulation 2024/1689 in references for further details.

3). Starting with key aspects of AI4SG conceptualisations manifest in the AI Act, we apply these concepts to the governance model for non-high-risk AI systems by exploring selected functional elements of its mechanism and, to this end, focus on key performance indicators (Section 4). An analysis of an NLP-based, language-simplifying AI application for assistive purposes enables us to further concretise how an operationalisation of the governance approach relates to central points discussed in the previous sections of this paper and some of the EU's policy objectives at large (Section 5). Finally, we summarise the key findings and provide an outlook (Section 6).

2. Governance of non-high-risk AI systems

The AI Act is part of the EU digital single market strategy and aims at a properly functioning market pursuant to Art. 114 of the Treaty on the Functioning of the European Union (TFEU). In support of a level playing field for the cross-border circulation of AI-related services and goods, national legislation by the member states would have led to a legal fragmentation. The EU legislature thwarts such a development by introducing comprehensive harmonised rules. Against this backdrop, the EU Commission intends to create “a legal framework for trustworthy AI” and a human-centric “ecosystem of trust” that is “a force for good in society with the ultimate aim of increasing human well-being” (European Commission, 2021). In line with one of the pathways set out in the EU's “White paper on artificial intelligence” (European Commission, 2020), the Commission has opted for an approach that provides for an EU-wide harmonised risk-oriented regulation of AI systems. For non-high-risk systems this implies a non-mandatory application of codes of conduct by all providers of such AI systems (European Commission, 2021). The AI Act stipulates a framework for the drawing-up of these codes of conduct. The central benchmark for all risk classifications is the threat to fundamental rights, whereby the EU Commission considers to be essential:

The rights to freedom of expression, freedom of assembly, human dignity, non-discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation, as applicable in certain domains, protection of personal data and private life, or the right to an effective judicial remedy and a fair trial, as well as consumer protection. (European Commission, 2020)

The scope of the provision for non-high-risk systems is defined by its negation and comprises all lawful AI systems not classified as high-risk (Art. 95 AI Act). Within this catch-all provision, the AI Act makes no further categorical risk-based distinc-

tion. As non-high-risk the AI Act also classifies systems that fall into one of the eight high-risk areas (Annex III AI Act) but “do not pose a significant risk of harm to the health, safety of fundamental rights of natural persons” (Art. 6 (3) AI Act), though this “reduced risk” comes with additional documentation requirements (Art. 6 (4) AI Act). The AI Office, a task-related designation for the EU Commission, assumes a central governance role on EU level, including contributions to the “implementation, monitoring and supervision of AI systems” (Art. 3 (47) AI Act). By virtue of this function, the Commission contributes to developing new standards and the implementation of common rules in all EU member states. The AI Board (Art. 65 AI Act) can issue recommendations and opinions regarding the development and use of codes of conduct as well as the specification of the requirements for high-risk AI systems (Art. 66 (e) (i) AI Act) that are also relevant for non-high-risk AI systems.

The EU AI Office and the EU member states are mandated to “facilitate and encourage” the development of “codes of conduct and related governance mechanisms” (Art. 95 (1) AI Act). These governance mechanisms are optional. “Related” governance tools are likely to include practices such as voluntary commitments, compliance frameworks for (internal) auditing, guidelines, or measures recognising compliance with standards or principles (certification, labelling, “trustmarks”). The providers and deployers of non-high-risk AI systems who develop and implement such codes are encouraged to translate a number of aspects on the basis of clear objectives that can be translated into measurable key performance indicators including: 1) European ethics guidelines for trustworthy AI; 2) environmental sustainability; 3) promotion of AI literacy; 4) inclusive and diverse AI system design and participatory, inclusive, and diverse development; and 5) assessment and prevention of negative impact on vulnerable persons, accessibility, gender equality.

The AI Act designates “individual providers or deployers of AI systems” (either themselves or delegated to organisations representing them) as the parties responsible for drawing-up codes of conduct (Art. 95 (3)). A “provider” is:

A natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge. (Art. 3 (3) AI Act)

The term “deployer” denotes “any natural or legal person, public authority, agency

or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity” (Art. 3 (4) AI Act).

Both providers and deployers are dubbed as “operators” (Art. 3 (8) AI Act). A hierarchisation through the distinction between the rule-makers (providers and deployers of AI systems) and the possibly involved further parties (stakeholders, representative organisations, civil society organisations, academia pursuant to Art. 95 (3) AI Act) is obvious and harbours potential for conflict. This point will be subject to further analysis in the following sections.

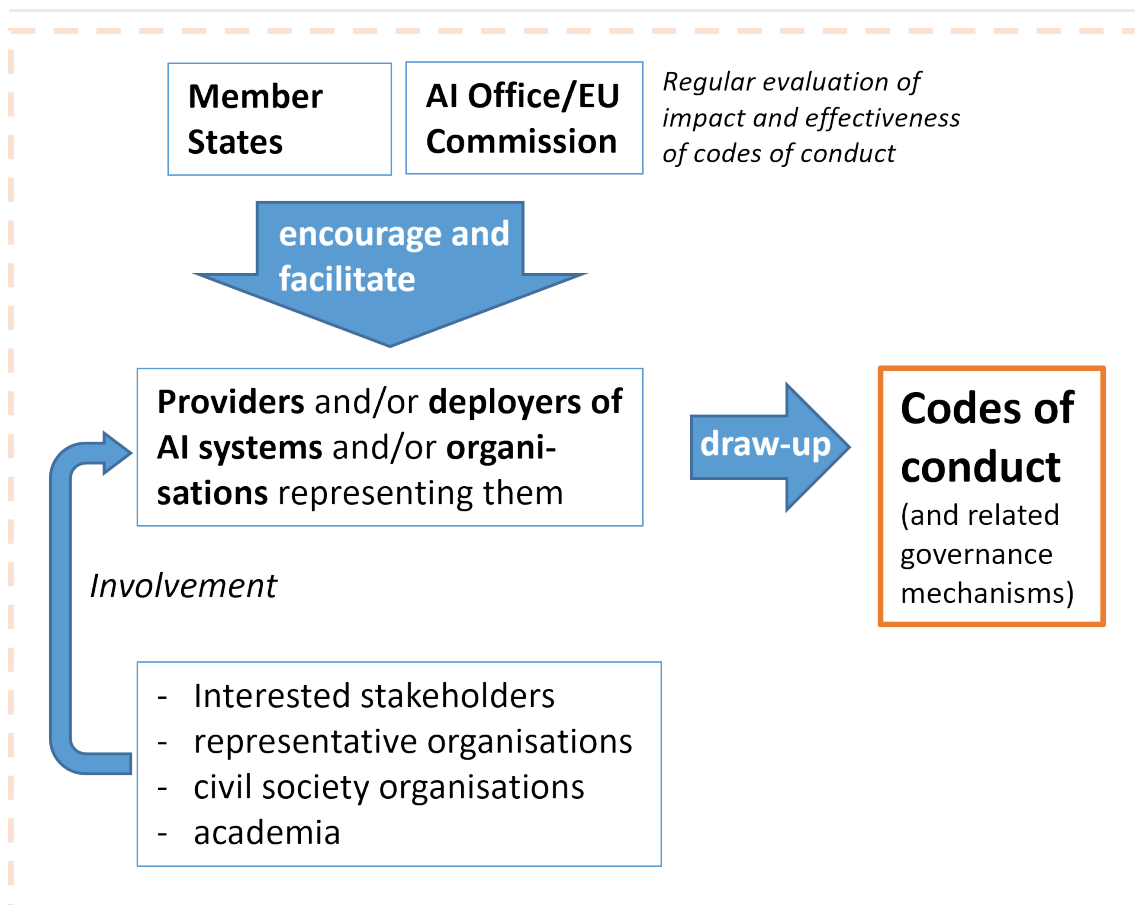


FIGURE 1: Governance for non-high-risk AI systems.

The transparency obligation (Art. 50 AI Act), being the only compulsory requirement, must be complied with by informing users about their interaction with an AI system, especially when content is created or manipulated (Hoffmeister, 2024). Beyond that, the AI Act specifies comprehensive and, quite in contrast to its preponderantly non-binding nature, detailed substantive aspects by defining concrete orientation points for the formulation of codes of conduct. These legally non-binding “specific requirements to all AI Systems” (Art. 95 (2) AI Act) encompass all or selected requirements mandatory for high-risk AI systems pursuant to Art. 8-15 AI

Act (e.g. risk management, data governance, documentation and transparency, human oversight) and without explicitly referencing a “state-of-the-art”, existing “technical solutions” and “industry best practices” (e.g. model datasheets / cards) for the application of these requirements. According to these provisions, AI systems must not pose risks to safety or fundamental rights and must provide for a lawful and responsible management and protection of data the AI system draws upon (AI Act, Recital 27).

Specific requirements that, based on a non-exhaustive list, are to be translated by providers and deployers of AI systems into “clear objectives” (Art. 95 (2) AI Act) and corresponding key performance indicators are designed to gauge whether these objectives have been attained. The operationalisation of objectives at the management level through the definition and measurement of key performance indicators ascertains a follow-up of what has been achieved. The conceptual proximity of the two elements, “clear objectives” and “key performance indicators”, with standardised management processes and audits is conspicuous. The non-binding “specific requirements” (Art. 95 (2) AI Act) are related to elements from ethics guidelines for trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019), environmental sustainability impact assessment, and management and the promotion of AI literacy. Further elements that should be taken into account when drawing-up codes of conduct are a diverse, gender-sensitive, and inclusive AI system design and development, also in view of accessibility and vulnerable persons. Although it is acknowledged in principle, it is largely unclear how and when stakeholder participation should occur (see Section 4 and 5).

General “hard” legal requirements outside the AI Act will also play a role when drawing-up codes of conduct. The functioning of the EU digital common market requires a uniform regulatory environment not only for general AI regulation but also for codes of conduct. This includes, but is not limited to, the protection of fundamental rights and democracy (AI Act, Recital 1). An example for other areas of regulation is the risk of anti-competitive results of codes of conduct. These would be suspicious from a competition law point of view, e.g. if competitors use governance arrangements to restrain competition by way of standardisation agreements (EU Commission, 2023).

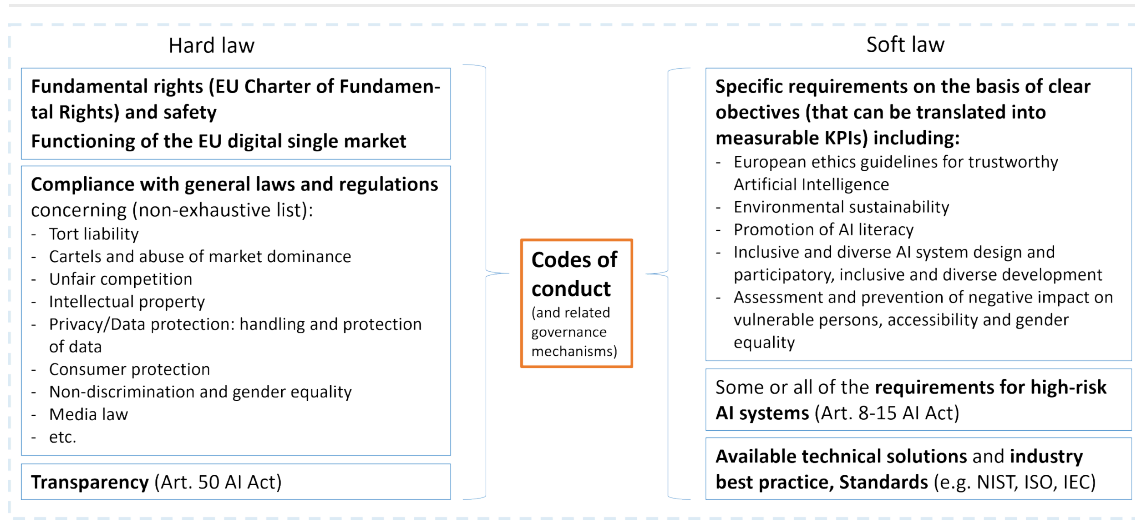


FIGURE 2: Hard law and soft law dimensions of codes of conduct for non-high-risk AI systems.

Organisations can, but do not have to, draw up their own codes of conduct; they can adopt recommendations by the European Artificial Intelligence Board or announce compliance with other standards. By leaving the decision entirely to the organisation's discretion, the flexibility allows the adoption of domain- or industry-specific standards (Laux et al., 2024). Thus, beyond the transparency obligations (Art. 50 AI Act) there is no specific "hard" restriction or obligation originating in the AI Act for non-high-risk AI systems.

All in all, fundamental uncertainties remain, starting with what exactly the EU Commission understands by "facilitate" and "encourage" and how voluntary the adoption of or compliance with codes of conduct will effectively be. The AI Board can issue recommendations and opinions on how codes of conduct are to be developed and applied (Art. 66 (e) (i) AI Act). It is not (yet) foreseeable to what extent the "elements" (requirements) referred to in the AI Act need to be implemented. Further uncertainty factors are the concrete implementation of the components "clear objectives" and "key performance indicators", as well as the fact that the statements referred to (High-Level Expert Group on Artificial Intelligence, 2019) will be subject to change and the reference to these rules is to be understood as a dynamic reference. Moreover, the legal liability framework depends on the duties of care, which have yet to be concretised and will in turn depend on soft law in accordance with Art. 95 of the AI Act (Hacker, 2023; Gille et al., 2023).

3. The EU's voluntary co-regulatory approach

Governance mechanisms for AI must balance and respond to the needs of innovation, the market, and fundamental rights (European Commission, 2020). Premised

on the assumption that soft governance instruments are the key to counteracting "too much" regulation as well as uninformed and inflexible legislation, the EU Commission saw no alternative to the introduction of a non-binding industry-led rulemaking (European Commission, 2020). In any case, with international competition gaining momentum and the concrete technological development being uncertain, preventive and detailed regulation was not deemed an option.

Technological advancement brings about risks that call for effective, yet flexible regulatory approaches (von Grafenstein, 2022). Traditional command-and-control frameworks with prescriptive and enforceable rules are criticised as inadequate in view of the tumultuousness of technological developments (Pagallo et al., 2019; Senden, 2005; OECD, 2015). Outpacing rigid rules, technology calls for other solutions, with "civil regulation" being the most auspicious one (Pagallo et al., 2019; Keller, 2008). A flexible and non-compulsory governance approach is associated with the necessary room for manoeuvre to respond to rapid technological developments (Mökander et al., 2022, p. 256).

The possible benefits of a "soft" governance approach include responsiveness to new difficulties and technical changes as well as greater flexibility and efficiency in adapting to technical disruptions due to expertise and practical grasp of the technology, no lengthy political procedures, and no strict framework that inhibits invention which might lead to greater legitimacy and is easier to harmonise internationally (Black, 2001; Hepburn, 2006). Among the drawbacks mentioned in the literature are aspects like inadequate oversight, profits being put before the interests of the public, insufficient protection and a general lack of responsibility, anti-competitive behaviour paired with a lack of consumer protection, and inadequate risk-shifting together with big players able to more easily afford compliance expenditures than their smaller competitors (Brown & Marsden, 2013; Senden, 2005). To assess the relevance of (some) of these benefits and downsides one needs to turn to the concrete design of the EU policy approach.

Notwithstanding their role as a central governance pillar for non-high-risk AI systems, the AI Act does not define codes of conduct. Generally, codes of conduct "aim to define standards and principles that ought to guide the behaviour of the addressee in a particular way" (Keller, 2008). In contrast to codes of ethics that usually consist of general principles abstractly expressing moral values and intending to guide professional decision-making, codes of conduct are somewhat more concrete and operative in that they often define (approved and disapproved) behaviour and procedures (Antonucci & Scocchi, 2018). Further, codes of conduct need to be distinguished from standards. Standardisation sets voluntary technical specifica-

tions or quality requirements for goods or services (Laux et al., 2024).

In principle, codes of conduct can be categorised on a continuum that ranges from “pure” self-regulation to co-regulatory approaches with stronger top-down state influence (Marsden, 2011). Self-regulation can generally be understood as rule-making by private parties other than public legislators, i.e. without involving national parliaments, international or supranational legislative institutions, or international treaties; in a narrower sense, it refers to “pure” corporate self-regulation and organisation, i.e. the autonomous rulemaking by companies for themselves (Muchlinski, 2021, p. 623). In a more common understanding, it refers to industry-, or more broadly, civil-society-driven private regulation efforts. Hence, the distance from state-centred regulation, though in varying shapes, is a prominent feature. Self-regulation can take different forms, e.g.

- codes of conduct outlining and concretising ethical principles, (industry) best practices, and standards;
- industry standards in the narrower sense (e.g. technical specifications, interoperability requirements and the like);
- certification schemes that support compliance with pre-defined criteria and transparency;
- compliance frameworks that establish (internal) auditing and enforcement mechanisms (Black, 2001).

By contrast, the umbrella-term *co-regulation* comprises forms of hybrid rulemaking and encompasses a variety of notions within a continuum between hard, top-down legislation and soft, bottom-up self-regulation. The term denotes regulatory phenomena that combine general legislation with the exercise of (partial) governance autonomy by one or more self-regulatory entities (Marsden, 2011, p. 46). The degree to which regulation influences the government can vary, but at least some sort of government involvement is required. In its most reduced form, the term requires a shared regulatory role between the government and stakeholders, including a legal link between self-regulation and the legislator (Black, 2001; Pagallo et al., 2019). Co-regulation, in another definition, can broadly be concretised as “non state law’ [supported by] some government involvement” (van Schooten & Verschuuren, 2008). As the EU’s soft governance for non-high-risk systems falls within this continuum, the following paragraphs will examine its specific conceptualisation and design.

The hybrid strategy of regulated self-regulation adopted by the AI Act for non-high-risk AI systems combines bottom-up soft self-regulation and “pure” top-down hard control. However, a more than just tentative subsumption of the governance

framework under current classification methodologies (e.g. Black, 2001; Marsden, 2011; Brown & Marsden, 2013) for non-high-risk AI systems appears to be unfeasible. There are too many factors that depend on the actual development of soft law and the process of official implementation of the AI Act to be able to categorise the regulatory approach exhaustively in a concrete theoretical framework at this stage. There still is too much room for the design of roles, interpretation and implementation of rules and principles, and future development of the governance framework that cannot be estimated yet (see Section 2). Rather than simply placing the AI Act's approach into already-existing, all-encompassing theoretical frameworks, no matter how elaborated they may be, we will instead turn to a general analytical characterisation of the AI Act's decentred governance pattern based on concepts put forward in the literature.

The EU's code-of-conduct-approach for non-high-risk AI systems is suspended between elements of a non-compulsory regulatory framework. There is no obligation to draw-up codes of conduct nor have sanctions or other coercive measures for non-compliant operators been arranged for. The codes of conduct can be drawn-up and implemented unilaterally by single providers, or they can be developed and used bilaterally or multilaterally involving deployers and organisations, presenting providers and / or deployers (see Figure 1). Moreover, stakeholder involvement is optional. It does not even have to be a code of conduct. Related instruments are also suitable. Counterposed to this hands-off approach with no direct enforcement and almost maximum flexibility, also regarding context, domain, and industry, there is a less *laissez-faire* side to the coin. The AI Office and the EU member states have considerable leverage potential in how they will "encourage" and "facilitate" the drawing-up of codes of conduct, taking into account the EU Commission's authority to monitor (Art. 112 (7) AI Act) and publish guidelines with an impact on codes of conduct (Art. 6 (5), 66 (e) (i), 80 (1), 96 AI Act). But even an extensive interpretation of this authorisation is unlikely to effectively lead to direct control.

Though providers and deployers of AI systems do not have to introduce codes of conduct, once they take this path, the AI Act's provision for the development of objectives and key performance indicators referred to provide detailed input. One can speak of *content-curated voluntary self-regulation*, i.e. self-regulation that is framed and content is suggested to a degree that the term co-regulation is not far off for this reason alone. The same holds for the implementation-process with its central pillars, clear objectives, and key performance indicators. Another argument against a governance classification as purely self-regulatory is that sanctionable infringements (Art. 99 AI Act) are linked to guidelines pursuant to Art. 96 AI Act.

Such guidelines, which have yet to be developed, can address codes of conduct directly (Art. 96 (1) AI Act) and indirectly (Art. 96 (1) in conjunction with Art. 8-15, 50 AI Act).

The co-regulatory approach is “smart” regulation in that it drives development towards the realisation of social preferences, is flexible to allow for susceptibility to unforeseen developments, and provides the agility to increase regulatory granularity with expanding expertise (Pagallo et al., 2019, p. 17). The “decentred” and “heterarchical” (Black, 2001) configuration with relatively concrete input regarding both content and procedure of implementation is in the case of the AI Act’s code of conduct not accompanied by a mandatory audit scheme or a third-party assessment. However, it is arranged for a regular evaluation and review of the “impact and effectiveness” of voluntary codes of conduct (Art. 112 (7) AI Act), i.e. the development is monitored at defined intervals by the EU Commission. Currently, although only a vague possibility, it is not entirely far-fetched that the EU will tighten strings in the future and turn to hard regulation.

From a broader legal perspective (Figure 2), codes of conduct are not completely toothless, even if they are not mandatory. The conceptual backbone of the EU AI Act is the risk-centred categorisation. The EU’s entire forthcoming liability framework for AI systems is based on risk assessment, a non-legal concept (Li et al., 2022) with a strong normative connotation (Kaminski, 2023). Liability law sends the signal that a duty of care is owed by the party in control of the risk source and thereby has a preventive effect (Gille et al., 2023). Any interpretation of AI Act concepts such as “public interest” and “fundamental rights” must consider this risk dimension and the underpinning risk allocation patterns in the proposed legislation. Depending on the risk classification, the regulatory intensity rises proportionally, allowing flexible and targeted interventions. The AI Act’s tendency to view AI industry development through a risk lens is not just a question of risk management but rather a broader question of risk governance with fundamental rights serving as the benchmark.

Risks, their classification and allocation are the medium through which the central objectives of the AI Act are turned into reality. Defined as “the combination of the probability of an occurrence of harm and the severity of that harm” (Art. 3 (2) AI Act), risks are, from an institutional theory point of view, transformed uncertainties that allow for actuarial management (North, 1990). Liability law introduces disincentives by introducing liability risks that AI system operators will want to avoid or minimise. Although the AI Act is often reflected in isolation and with focus on its pre-emptive effects, there is also a “hard” remedial side to it. The purported en-

forcement gap (Wörsdörfer, 2023) is at least indirectly filled by general liability provisions and insurance implications, as well as liability rules under (unfair) competition and data protection law. At least in the medium to long term, standards of due diligence under liability law will develop along the main lines of codes of conduct. Not abiding by these standards of care can lead to a liability in tort (Marchant, 2019, p. 13).

4. “Specific requirements” for human centric and trustworthy artificial intelligence

The EU explicitly connects the development of non-high-risk AI systems with the expectation of “a larger uptake of ethical and trustworthy AI” (AI Act, Recital 165). Against the background of this legislative intention, the question arises as to how codes of conduct be implemented on the basis of “clear objectives and key performance indicators” (Art. 95 (2) AI Act). Umbrella terms for ethically good AI changed iteratively over time, and many concepts are derived from information philosophy (Baum, 2017). Terminological ingenuity seems to have experienced a particular upswing during the emergence of Transformer (Vaswani et al., 2017) models around the end of 2017. One strand of ethics-driven conceptualisations, using terms such as “Good AI Society” (Cowls & Floridi, 2018), “AI4People” (Floridi et al., 2018), and “AI4SG” (AI for social good) (Taddeo & Floridi, 2018), is the basis for the following analysis. A working definition provided by Cowls and colleagues (2021) invokes ethical principles, beneficence, and non-maleficence and specifies AI4SG as the:

Design, development and deployment of AI systems in ways that help to (i) prevent, mitigate and/or resolve problems adversely affecting human life and/or the wellbeing of the natural world, and/or (ii) enable socially preferable or environmentally sustainable developments, while (iii) not introducing new forms of harm and/or amplifying existing disparities and inequities.

The process of defining continues, but it has not left the recent policy discussion completely untouched. Concepts reflected in the AI4SG-debate, although broader, appear to have played some role in shaping provisions of the AI Act. The AI Act bundles it under: “human centric and trustworthy artificial intelligence” (Art. 1 (1) AI Act), which, as we reconstructed in Section 2, culminates in detailed specific requirements for codes of conduct.

The concept of “trustworthiness” is used by the EU Commission to support acceptance and dissemination of AI in Europe (Beckert, 2021). It seems to be emerging

as a core element for non-high-risk AI systems and is premised on the rather tentative equation that trust is established when risks are absent. However, beyond that the term trustworthiness can only be found in the recitals, in Art. 1 and Art. 95 of the AI Act. It is covered extensively by the "European ethics guidelines for trustworthy AI" (High-Level Expert Group on Artificial Intelligence, 2019). The guidelines are likely to be further developed in parallel with the implementation process for the requirements of the AI Act. The current version from 2019 (High-Level Expert Group on Artificial Intelligence, 2019) provides initial input for the practical implementation of the Trustworthy Guidelines through the "Trustworthy AI Assessment List" (European AI Alliance, n.d.; High-Level Expert Group on Artificial Intelligence, 2019). By providing concrete requirements for trustworthy AI, the Guidelines contribute to the implementation of the co-regulatory approach (Art. 95 (2) AI Act): (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination, and fairness; (6) environmental and societal well-being; and (7) accountability. The list should be seen as non-exhaustive and without imposing any hierarchy.

Risks cannot be objectively assessed and measured (Kusche, 2024). Nonetheless, the success of the governance for non-high-risk AI systems depends, among other things, on the availability of reliable benchmarks for measuring the legislative objectives. This provision not only specifies requirements but goes further and explicitly calls for "clear objectives" that constitute a basis for assessing whether the prerequisites have been met. This verification needs a system for measuring, i.e. one or more metrics (Cambridge University Press, 2024b). These metrics need to be based on goals to serve their purpose as key performance indicators (Mökander et al., 2022). Key performance indicators show whether defined criteria have been satisfied (Cambridge University Press, 2024a).

Table 1 lists all requirements for codes of conduct for non-high-risk systems that can be extracted from the AI Act (Art. 95 (2) AI Act) and gives examples of objectives combined with examples of pertinent metrics. These metrics were selected via a bibliometric indicator of their scientific impact in the form of the number of citations on Google Scholar, prioritising on standardised metrics provided by established agencies. This form of literature selection makes no claim to completeness. It is also potentially prone to biases and is controversial (Goldenfein & Griffin, 2022). Nevertheless, it fulfils the purpose of showing, in principle, how such a selection procedure could be carried out. Selecting key performance indicators involves the identification and definition of objectives, involvement of multiple stakeholders, and should be put into effect with due regard for the situation and

the task.

TABLE 1: Requirements of human centric and trustworthy artificial intelligence with exemplary objectives and metrics

REQUIREMENTS FROM ART. 95 (2) AI ACT	EXAMPLE OBJECTIVE	EXAMPLE METRIC
1. European ethics guidelines for trustworthy AI, (current version: High-Level Expert Group on Artificial Intelligence, 2019)		
1.a Human agency and oversight	AI human rights impact assessments (European Union Agency for Fundamental Rights., 2020)	HRESIA (Mantelero, 2022)
1.b Technical robustness and safety	Adversarial robustness (Göllner et al., 2023)	Randomised Smoothing (Cohen et al., 2019)
1.c Privacy and data governance	Privacy leakage (Göllner et al., 2023)	Membership Inference Attacks (Shokri et al., 2017)
1.d Transparency	Faithfulness, feature importance scores, SHAP values (Göllner et al., 2023)	Debiased-CAM (Zhang et al., 2022)
1.e Diversity, non-discrimination and fairness	<i>please see (4) Inclusive and diverse AI system design and participatory, inclusive and diverse development</i>	
1.f Environmental and societal well-being	Human well-being	IEEE 7010-2020 (IEEE Standards Association., 2020)
1.g Accountability	Objective- accountability (Krafft et al., 2022)	PACT (Bondi et al., 2021)
2 Environmental sustainability	GHG emissions from production (OECD, 2022, p. 24)	tons of CO2e (OECD, 2022, p. 24) ISO/IEC 30134 (ISO/ IEC JTC 1/SC 39)
3 Promotion of AI literacy	Non-experts AI literacy (Laupichler et al., 2023)	SNAIL (Laupichler et al., 2023)
4 Inclusive and diverse AI system design and participatory, inclusive, and diverse development	Diversity attributes (Shams et al., 2023)	Distribution of gender (Budescu & Budescu, 2012)
5 Assessment and prevention of negative impact on vulnerable persons, accessibility, gender equality	Fairness (Göllner et al., 2023)	Bayesian Fairness (Dimitrakakis et al., 2019)

From a technical and engineering perspective, metrics such as accuracy, precision, recall, and F1 score can also be useful in evaluating the performance and reliability of AI algorithms. These metrics quantify the system's ability to achieve its intended goals accurately and consistently. Furthermore, metrics related to model interpretability, such as feature importance scores or SHAP values, are crucial for understanding how AI systems achieve their outcomes and for identifying poten-

tial biases or discrimination patterns. SHapley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017) quantify each feature's importance by measuring the change in the expected model prediction when conditioned on that feature. In this way, impactful features can be identified and evaluated for potential bias patterns.

However, it is important to recognise the limitations of such technical metrics when it comes to capturing the multi-layered nature of trustworthiness and human-centricity of AI systems (Göllner et al., 2023). Since the AI Act centres on the risks of AI systems for fundamental rights, it is doubtful that quantifiable metrics are sufficient (Kaminski, 2023). Soft metrics such as trust and perceived fairness are likely to provide a better understanding of the subjective experiences and perceptions of end users. From a metrology point of view, these metrics are often described as “measurement techniques and models which enable the objective quantification of properties which are determined by human perception” (Pointer, 2003). While these metrics are more difficult to quantify, they provide feedback on the usability, acceptability, ethical implications, and especially of AI systems’ risks to fundamental rights. In addition, qualitative assessments, including stakeholder consultations, ethical impact assessments, and scenario-based evaluations, provide a view of the broader societal impacts of AI technologies. Thus, while technical metrics provide essential quantitative measures, it is crucial to complement them with soft metrics and qualitative assessments to ensure a holistic understanding of the trustworthiness and human-centeredness of AI systems. We are therefore in favour of a broad interpretation of the term “key performance indicator”, which also includes qualitative indicators.

The following metrics collectively provide a multidimensional framework for assessing the trustworthiness and human-centricity of AI systems, covering both technical performance and broader societal impact. However, it is important to recognise the inherent challenges in quantifying some aspects, particularly those related to subjective user perceptions and societal expectations.

TABLE 2: Measurable metrics (non-exhaustive list)

MEASURABLE METRICS	EXAMPLES
Performance under diverse conditions	accuracy, precision, recall, f1-score, robustness
Model interpretability and explainability	feature importance scores, SHAP values, complexity of explanations, comprehensibility
Bias detection and mitigation metrics	disparate impact analysis, fairness metrics
Scalability and efficiency metrics	inference time, resource utilisation
Security metrics	vulnerability assessment, adversarial robustness

MEASURABLE METRICS	EXAMPLES
Compliance with regulatory standards and guidelines	GDPR, defined ethical principles, privacy leakage

TABLE 3: Soft metrics (non-exhaustive list)

SOFT METRICS	EXAMPLES
User	satisfaction, trust, empowerment, perception, and acceptance
Model	perceived usefulness, transparency, and openness
Ethical considerations	fairness, privacy, accountability, ethical impact
Societal impact	inclusivity, accessibility, stakeholder engagement, and collaboration
Alignment with organisational or societal goals and values	
Alignment with human values and preferences	

While these metrics contribute to a possible framework for assessing AI systems, quantifying subjective user experiences and societal perceptions poses inherent challenges. Hence, achieving a robust governance requires a balanced approach, integrating quantitative metrics, qualitative assessments, and stakeholder engagement. Yet, bridging the gap between technical performance, risks for fundamental rights, and societal impact remains a significant challenge, demanding continuous refinement and critical evaluation in policy formulation for AI, whereby the most fundamental parameter remains the risk assessment with regard to the violation of fundamental rights. This challenge, which is closely linked to the code of conduct conceptualisation (Art. 95 AI Act), will be further analysed by way of a case study analysis provided in the next chapter.

5. Case study on key performance indicators for an assistive AI-application

To further concretise the challenges of operationalising the AI Act’s governance approach for non-high-risk systems and to show how these relate to central points discussed in the previous sections of this paper, we build on a case study analysis. The case is not only an example of the role that soft law, developed in a co-design process, can play for the development of AI applications. The analysis also illustrates how difficult it is to delineate high-risk AI systems from non-high-risk systems. The legal domain, in which the case takes place, is particularly risk-prone regarding the impairment of fundamental rights, i.e. a risk-proportional practice is a

challenge, especially against the background of the vulnerable addressees of the assistive communication application. Moreover, the case is also a good example of the process set out in the AI Act of defining objectives and identifying pertinent key performance indicators (Section 4). At the same time, it demonstrates the need for soft metrics, especially with regard to AI systems that aim at non-average users (Kaminski, 2023).

The AI system research that, in the following, we take a closer look at, deals with an language-processing-based (NLP) automated text simplification for assistive use, converting German texts from standard or specialist language (administrative/public service communication from the legal domain) to accessible, comprehensibility-enhanced language (Schomacker et al., 2024). Comprehensibility-enhanced language is a variant of accessible communication, i.e. a “form of communication that is accessible and usable for people with different communication needs” (Hansen-Schirra et al., 2021). It is catered to persons with comprehension difficulties, which are among others: persons with intellectual disabilities, learning difficulties, dementia, aphasia, multiple disabilities, and language learners (Rink, 2018). The texts used for model training are in standard or specialist and in a corresponding easy language version. The data set focuses on the specifics of legal texts with everyday relevance for the vulnerable groups. Until now, no comparable data set exists in German. The research-output from this project includes a framework for the development and training of text simplification models for this domain.² The project team is interdisciplinary and consists of lawyers and computer scientists.

The simplified texts targeted at vulnerable persons with comprehension difficulties convey information about aspects of life relevant to fundamental rights (e.g. political information, social entitlement, health). One goal is to lay out the foundation for a suitable data set based on the soft law standard for German Easy Language (DIN SPEC 33429). Although it was not published to primarily function as such, this soft law standard is used for NLP-based automated text simplification, because it has been developed in a participatory co-design process (*DIN-Normenausschuss Ergonomie*, 2023). DIN SPEC 33429 only applies to German, but there are international endeavours to set standards for text simplification. For example, ISO 24495-1 (DIN 8581) for plain language. The soft-law-based development process of the text simplifying NLP followed a five-step methodology: 1) Identification of datasets and models used for generating German Easy Language (Schomacker et al., 2023), 2) Interviews with experts for Easy Language, 3)

2. See the OPEN-LS (n.d.) entry in reference list for further details.

Analysing current standards for Easy Language and deriving aspects of quality, 4) Combining the insights and identified gaps from previous steps into dataset criteria (Schomacker et al., 2024), and 5) Collecting data and annotating it in accordance with the criteria. The account of this research is integrated in this paper to exemplify the assessment of the suitability of an AI4SG-driven, risk-sensitive analysis of the AI Act's governance approach.

From an AI4SG informed viewpoint, several aspects of such an AI system for communication-assisting purposes indicate the need for a closer examination. A not (only) for-profit justification (Züger & Asghari, 2023) can be posited given the failure of the market to develop assistive applications and the assistive purpose as such being in the public interest. From a data perspective, the respect for human rights and privacy (Tomašev et al., 2020) as well as data ethics and process governance (Züger et al., 2022) are particularly conspicuous facets that require attention above and beyond the already high level of attention in view of the target group of assistive applications. Simplified language, in particular Easy Language, has a target group with comprehension difficulties. This prerequisite, makes explainability and interpretability, adaptability and user-friendliness (Akula & Garibay, 2021) highly relevant aspects when adjusting a text simplifying AI application to the special needs of these addressees. Furthermore, assessing and preventing negative impact on vulnerable persons and accessibility considerations are at the centre of a fundamental-rights oriented risk assessment in line with the AI Act and would need to be reflected transparently by the key performance indicators.

Important aspects of trustworthiness, flipside of an assessment sensitive to fundamental rights as required by the AI Act, are safety, effectiveness, and ethical soundness of the developed system. Trustworthy AI includes a whole spectrum of requirements and considerations, ranging from technical robustness to ethical considerations and societal impact (Göllner et al., 2023). A basic requirement for trustworthy AI is transparency, i.e. explaining clearly and adequately how the AI system works and makes its decisions and predictions. In the case of text simplification for people with cognitive disabilities, transparent AI is essential to ensure that users can understand and trust the results of the system. This includes not only explaining the simplification process but also disclosing any biases or limitations in the system's algorithms or training data. Information asymmetries exist when there is a lack of transparency between the regulated party and the regulator (Black, 2001).

In addition to transparency, promoting AI literacy among vulnerable populations is crucial for fostering trust and empowering users to make informed decisions about AI technologies, including whether and how to use them. This can be inferred from

autonomy-considerations (the power to decide) and even more basic, human dignity aspects (Floridi & Cowls, 2019). AI literacy encompasses understanding the capabilities and limitations of AI systems, as well as recognizing potential risks and biases (Long & Magerko, 2020). Tailored AI literacy programmes should focus on demystifying AI concepts, providing practical guidance on using AI applications effectively, and promoting digital literacy skills to mitigate potential risks associated with AI usage. Moreover, incorporating feedback mechanisms and user-friendly interfaces can further facilitate comprehension and engagement among vulnerable users, ultimately contributing to their autonomy and well-being.

The risk profile of the addressees of simplifying AI systems for assistive purposes (vulnerable persons with, e.g. cognitive impairment) and a high risk of bias (great heterogeneity within the group) lead to specific risk patterns (Gille et al., 2023) that require an adequate response in the key performance indicators employed for the task of text simplification. The diverse and heterogeneous nature of the target group therefore necessitates a thorough evaluation and validation of the system across different groups with users in real-world settings. The text simplification system must be resilient to errors, adversarial attacks, and unexpected inputs to ensure the safety and well-being of users (High-Level Expert Group on Artificial Intelligence, 2019). This necessitates rigorous quality assurance procedures throughout the development lifecycle of the AI system. Therefore, the following table delineates the requirements, objectives and key performance indicators necessary to address these considerations in our case study.

TABLE 4: Outline of a key performance indicator (KPI) catalogue for the selected case study

REQUIREMENTS FROM TABLE 1	EXAMPLE OBJECTIVE	KPI FOR THE CASE STUDY	
1. TRUSTWORTHY AI:			
1.A HUMAN AGENCY AND OVERSIGHT	As our target group is vulnerable, the degree of its "agency and oversight" is verifiable primarily with KPIs and objectives from requirement (4).		
1.B TECHNICAL ROBUSTNESS AND SAFETY	See Table 1 (no need for specialised KPIs)		
1.C PRIVACY AND DATA GOVERNANCE			
1.D TRANSPARENCY	Faithfulness (Göllner et al., 2023)	Divergence-based faithfulness	Moradi et al.'s (2021) approach works on sentence-level translation. Each translated word is aligned to the corresponding word in the original sentence. Subsequently, the faithfulness is measured by the

REQUIREMENTS FROM TABLE 1	EXAMPLE OBJECTIVE	KPI FOR THE CASE STUDY	
		objective (Moradi et al., 2021)	accuracy of the alignment. This enables the user to quantify the degree of the translational faithfulness per word. This metric additionally penalises the model for the lack of connection between the translated and source words. To eventually force the model to learn better translations by forcing it to justify each output in a <i>right answer for the right reason</i> paradigm.
1.E DIVERSITY, NON-DISCRIMINATION AND FAIRNESS	<i>see (4) Inclusive and diverse AI system design and participatory, inclusive and diverse development</i>		
1.F ENVIRONMENTAL AND SOCIETAL WELL-BEING	<i>see (2) Environmental sustainability</i>		
1.G ACCOUNTABILITY	Objective-accountability (Krafft et al., 2022)	PACT (Bondi et al., 2021)	PACT (Bondi et al., 2021) offers the opportunity to capture the existing capabilities of the target group in a participatory manner through a question-answer-based process. Afterwards, desirable new capabilities are defined. Their degree of fulfilment can serve as a performance indicator. This gives us an objective with which the project can be held accountable. PACT also considers the heterogeneity of the target group to the extent that different (sub-)communities can be formed, which can stand up for their respective interests.
2. ENVIRONMENTAL SUSTAINABILITY	No need for specialised KPIs, please see Table 1		
3. PROMOTION OF AI LITERACY	Non-experts AI literacy (Laupichler et al., 2023)	No directly applicable literature on this topic at present	"AI literacy" is a relatively new concept. Initial approaches are only generally designed for non-experts. However, these non-experts have a high level of comprehension and the capability of abstraction. As our approach is aimed at people with cognitive impairments, current approaches fail. Therefore, there is a research gap in the form of an "AI literacy"-measure for people with cognitive impairments.
4. INCLUSIVE AND DIVERSE AI SYSTEM DESIGN AND PARTICIPATORY, INCLUSIVE, AND DIVERSE DEVELOPMENT	Diversity attributes (Shams et al., 2023)	-	Diversity is traditionally measured using various diversity attributes. Gender is often used as such an attribute. There are already functioning approaches for this aspect. In the text simplification case study, however, the project works with a very heterogeneous group of people with cognitive impairments. These impairments are difficult to categorise in attributes. Even if verifiable attributes are available, previous diversity metrics do not provide any information about the real level of participation of the target group. These metrics therefore currently harbour more of a risk to be used for so-called participation washing than being a genuine indicator for the degree of fulfilment.
	Participation quality (Gupta et al., 2023)	-	Gupta (Gupta et al., 2023) confirmed that current participation measures focus largely on these measures' existence, rather than their quality. But to measure the quality of the participation, it needs to be (better) defined. There is a need for data based on objective measures (for example speaking time of different participants) as well as subjective ones (participant views on whether they felt listened to).
5. ASSESSMENT AND PREVENTION OF NEGATIVE	Quality of Life (Schalock et al., 2003)	KidsLife scale (Gómez et al., 2016)	As pointed out by the (High-Level Expert Group on Artificial Intelligence, 2019): "No commonly accepted or widely agreed legal definition of vulnerable persons exists, due to their heterogeneity." This lack of definition makes it particularly hard to quantify the

REQUIREMENTS FROM TABLE 1	EXAMPLE OBJECTIVE	KPI FOR THE CASE STUDY	
IMPACT ON VULNERABLE PERSONS, ACCESSIBILITY, GENDER EQUALITY			impact on the group. We argue that the growing scientific community around the "Quality of life"-concept (Schalock et al., 2003) provides a suitable construct for metrics.

As the table demonstrates, important aspects are not captured by this process-management and quantification-oriented formulation of key performance indicators. Furthermore, the evident proximity to legally hazardous scenarios (jeopardising fundamental rights of vulnerable persons, binding "hard" law requirements, communication with public service providers) necessitates passing near the high-risk zone. But without posing "significant" risks of harm when performing "a narrow procedural task" (Art. (3)(a) AI Act), the system is nonetheless likely to be classified as reduced/non-high-risk, provided that codes of conduct are implemented. Due to the lack of a legal or at least commonly recognised definition of vulnerable persons, the system provider or deployer must make allowance for the context, such as market and other economic factors, age, gender, language, disability, religion, culture, illness, etc. (High-Level Expert Group on Artificial Intelligence, 2019). Moreover, the participatory model development and implementation require standards that are yet to be drawn-up.

6. Main findings and outlook

We tentatively classify the EU's approach to regulating non-high-risk AI systems as voluntary co-regulation due to its hybrid character with both, hands-off and hands-on elements. In view of the concrete and detailed substantive legislative input, the role of the AI Board, frequent evaluations, and the possibility of sanctions, it would be misleading to speak of a pristine self-regulatory approach. Moreover, the ubiquitous risk-orientation needs to be reflected in the context of a comprehensive AI liability framework. Viewed overall, the AI Act's governance set-up and broader legal context has, in principle, the potential to support differentiating, risk-proportional rulemaking for non-high-risk AI systems. The experience from the EU General Data Protection Regulation with its industry-developed codes of conduct for specific sectors is probably what the EU Commission had in mind, though it is a model with only slow progress (von Grafenstein, 2022).

The legal reconstruction of the EU's soft governance approach created the basis for analysing central aspects of codes of conduct. It is unclear how much flexibility

remains once the AI Office /EU Commission starts to “encourage and facilitate” (Art. 95 (1) AI Act) and what the concrete design of frameworks that respond to technology development dynamics will look like. Several characteristics of the co-regulatory strategy were carved out in Sections 2 and 3, among them a possibly strong indirect enforcement, a reliance on the market, and a “curated” substantive content. The dynamic reference to the EU’s Ethics Guidelines for Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019) is a flexible gateway to a more stakeholder-oriented approach. Still, a fair risk allocation and equal participation in the co-regulatory processes is not *per se* present in the EU’s regulatory design, not only because it is voluntary. At the same time, the likelihood of not being classified as a high-risk system increases if the installation of “strong” codes of conduct reduces the related hazards. Thus, if risk categorisation is also contingent on voluntary and effective risk reduction measures consented by stakeholders, such as codes of conduct, then the risk levels’ demarcation discussion is likely to also turn up the heat on these soft governance measures. Regarding non-high-risk systems an incentive is created for autonomous “good” AI system governance, with the reclassification as high-risk acting as a deterrent. Resorting to ethics should therefore not be misinterpreted as just another way to steer clear of binding regulation. Rather, this can motivate operators to adopt rules voluntarily.

Having demonstrated that achieving a robust governance requires a balanced approach, we plead for integrating hard and soft metrics, qualitative assessments, and stakeholder engagement. The discussion of the key performance indicators has shown that the operationalisation of indispensable qualitative evaluation criteria is fraught with challenges. At the same time, the availability of reliable and informative benchmarks for measuring the objectives pursued by AI system operators is decisive to the success of the EU’s co-regulatory approach. Although technical metrics deliver quantifiable measures, it is essential to add soft metrics as well as qualitative assessments to gain a sufficiently comprehensive depiction of societal expectations and user perspective(s), in other words of the trustworthiness and human-centeredness of a system. The case study has enabled us to take a look at what such an approach might look like and how this would help safeguard fundamental rights of an assistive AI application’s vulnerable addressees. Thus, a fundamental rights-oriented risk analysis, the conceptual backbone of the AI Act, cannot rely solely on “hard” quantifiable metrics. Additionally, addressing the challenges in benchmark availability for measuring AI objectives is crucial for the success of such governance approaches, emphasising the importance of collaborative efforts among stakeholders to establish robust evaluation standards. In addition, the results presented here could be integrated into holistic approaches, such as

world-model based approaches (Dalrymple et al., 2024).

Recognising the dynamic nature of technological advancements and societal needs, continuous refinement and adaptation of governance mechanisms are imperative to ensure the responsiveness and relevance of regulatory frameworks in addressing emerging challenges and opportunities in AI development and deployment. At this point in time, doubts about the balance of the regulatory implementation are certainly justified, namely whether sufficient legal certainty on the one hand and sufficient adaptability and flexibility on the other are in an appropriate balance. Beyond that and despite the insistence on “clear objectives” and “key performance indicators”, providers and operators of AI systems are not required to establish effective control and accountability mechanisms. Thus, the role of co-regulation in addressing societal challenges such as digital inclusion and algorithmic bias is tackled only indirectly.

Looking ahead from a technological perspective, advancements in AI will continue to innovate the range of assistive applications, offering many opportunities to enhance accessibility and inclusivity for individuals with impairments. However, ensuring alignment with user needs will often require collaboration with stakeholders. With the case study in Section 5 we wanted to demonstrate what user involvement in co-designing AI solutions to inclusivity and usability could look like in this context. Additionally, interdisciplinary collaboration among stakeholders is essential to addressing the complex challenges. By embracing a collaborative and user-centred approach, assistive AI technologies can better meet the diverse needs of users and promote accessibility and autonomy. It remains to be seen whether the chosen co-regulatory path will have this effect.

The tension between regulatory flexibility and referenced substantive principles is striking and may be due to the constraints of political consensus-building in the trilogue between EU Commission, EU Parliament, and the member states. A closer analytical look revealed not only limitations concerning the ostensible flexibility but also some substantial flaws in the governance design. The hierarchisation between providers (and deployers) of AI systems on the one hand and stakeholders on the other is palpable. The decision as to whether, to what extent, and how codes of conduct are to be drawn-up and applied is the sole responsibility of the provider and/or deployer, despite pertinent guidelines referenced by the AI Act. Equity and fairness of the policy response needs to bear in mind that benefits and costs can deviate between different groups in society. Transparency is crucial to assure that the party responsible for drawing-up the code of conduct does not act solely in its own interests. The wide scope of the AI Act's non-compulsory provi-

sions for non-high-risk AI systems paired with a not always transparent network of institutions and rulemaking bodies will leave many questions unanswered for quite some time. Further, the role of the EU Commission in its role as the AI Board has yet to be found. Its success depends on both political will and funding.

References

- Antonucci, M. C., & Scocchi, N. (2018). Codes of conduct and practical recommendations as tools for self-regulation and soft regulation in EU public affairs. *Journal of Public Affairs*, 18(4). <https://doi.org/10.1002/pa.1850>
- Baum, S. D. (2017). On the promotion of safe and socially beneficial artificial intelligence. *AI & Society*, 32(4), 543–551. <https://doi.org/10.1007/s00146-016-0677-0>
- Beckert, B. (2021). The European way of doing Artificial Intelligence: The state of play implementing Trustworthy AI. *2021 60th FITCE Communication Days Congress for ICT Professionals: Industrial Data – Cloud, Low Latency and Privacy*, 1–8. <https://doi.org/10.1109/FITCE53297.2021.9588560>
- Bernisson, M. (2021). *The Public Interest in the Data Society* [Doctoral thesis, Karlstad University]. <https://kau.diva-portal.org/smash/record.jsf?pid=diva2%3A1544799&dswid=-2596>
- Black, J. (2001). Decentering regulation: Understanding the role of regulation and self-regulation in a “post-regulatory” world. *Current Legal Problems*, 54(1), 103–146. <https://doi.org/10.1093/clp/54.1.103>
- Bondi, E., Xu, L., Acosta-Navas, D., & Killian, J. A. (2021). Envisioning communities: A participatory approach towards AI for social good. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 425–436. <https://doi.org/10.1145/3461702.3462612>
- Brown, I., & Marsden, C. T. (2013). *Regulating code: Good governance and better regulation in the information age*. The MIT Press.
- Budescu, D. V., & Budescu, M. (2012). How to measure diversity when you must. *Psychological Methods*, 17(2), 215–227. <https://doi.org/10.1037/a0027129>
- Cambridge University Press. (2024a). Key performance indicator. In *Cambridge Business English Dictionary*. <https://dictionary.cambridge.org/dictionary/english/key-performance-indicator>
- Cambridge University Press. (2024b). Metric. In *Cambridge Business English Dictionary*. <https://dictionary.cambridge.org/dictionary/english/metric>
- Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. *Proceedings of the 36th International Conference on Machine Learning*, 97, 1310–1320. <https://proceedings.mlr.press/v97/cohen19c.html>
- Cowls, J., & Floridi, L. (2018). *Prolegomena to a white paper on an ethical framework for a good AI society* [Prolegomena]. <http://dx.doi.org/10.2139/ssrn.3198732>
- Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*, 3(2), 111–115. <https://doi.org/10.1038/s422>

56-021-00296-0

Dalrymple, D. 'davidad', Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Zhi-Xuan, T., Wing, J., & Tenenbaum, J. (2024). *Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2405.06624>

Dimitrakakis, C., Liu, Y., Parkes, D. C., & Radanovic, G. (2019). Bayesian fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 509–516. <https://doi.org/10.1609/aaai.v33i01.3301509>

DIN-Normenausschuss Ergonomie. (2023). *DIN SPEC 33429:2023-04, Empfehlungen für Deutsche Leichte Sprache [DIN SPEC 33429:2023-04, Recommendations for German Simple Language]* [Technical rule outline]. Beuth. <https://doi.org/10.31030/3417293>

European AI Alliance. (n.d.). *Welcome to the ALTAI portal!* Futurium – European Commission. <https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal>

European Commission. (2020). *White paper on artificial intelligence. A European approach to excellence and trust* (White Paper No. COM(2020) 65 final; pp. 1–26). https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission. (2021). *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* (Proposal No. COM/2021/206). <https://eur-lex.europa.eu/legal-content/EN/TEXT/?uri=celex:52021PC0206>

European Commission. (2023). *Guidelines on the applicability of Article 101 of the Treaty on the Functioning of the European Union to horizontal co-operation agreements* (Communication from the Commission No. 2023/C 259/01). https://eur-lex.europa.eu/legal-content/EN/TEXT/?uri=uriserv:OJ.C_.2023.259.01.0001.01.ENG

European Union Agency for Fundamental Rights. (2020). *Getting the future right: Artificial intelligence and fundamental rights* [Summary]. Publications Office of the European Union. <https://data.europa.eu/doi/10.2811/155>

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>

Gille, M., Schomacker, T., von der Hülls, J., & Tropmann-Frick, M. (2023). Der Einsatz von Neural Language Models für eine barrierefreie Verwaltungskommunikation: Anforderungen an die automatisierte Vereinfachung rechtlicher Informationstexte [The use of neural language models for accessible administrative communication: Requirements for the automated simplification of legal information texts]. 6. *Fachtagung Rechts- und Verwaltungsinformatik*, 144–158. <https://doi.org/10.18420/RVI2023-013>

Goldenfein, J., & Griffin, D. (2022). Google Scholar – Platforming the scholarly economy. *Internet Policy Review*, 11(3). <https://policyreview.info/articles/analysis/google-scholar-platforming-scholarly-economy>

Göllner, S., Brumen, B., & Tropmann-Frick, M. (2023). Aspects and views on responsible artificial intelligence. In G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, P. Pardalos, G. Di Fatta, G. Giuffrida, & R. Umeton (Eds.), *Machine Learning, Optimization, and Data Science* (Vol. 13810, pp. 384–398). Springer. https://doi.org/10.1007/978-3-031-25599-1_29

Göllner, S., Tropmann-Frick, M., & Brumen, B. (2024). Towards a definition of a responsible artificial intelligence. In M. Tropmann-Frick, H. Jaakkola, B. Thalheim, Y. Kiyoki, & N. Yoshida (Eds.), *Information Modelling and Knowledge Bases XXXV* (Vol. 380, pp. 40–56). IOS Press. <https://doi.org/10.3233/FAIA231146>

Gómez, L. E., Alcedo, M. Á., Arias, B., Fontanil, Y., Arias, V. B., Monsalve, A., & Verdugo, M. Á. (2016). A new scale for the measurement of quality of life in children with intellectual disability. *Research in Developmental Disabilities*, 53–54, 399–410. <https://doi.org/10.1016/j.ridd.2016.03.005>

Gupta, P., Rouffy-Ly, B., Rohrer-Herold, K., Koch, K., Rao, N., Poulussen, C., Brearley, L., Abou-Taleb, H., & Rajan, D. (2023). Assessing the interactions of people and policy-makers in social participation for health: An inventory of participatory governance measures from a rapid systematic literature review. *International Journal for Equity in Health*, 22. <https://doi.org/10.1186/s12939-023-01918-2>

Hacker, P. (2023). The European AI liability directives – Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*, 51. <https://doi.org/10.1016/j.clsr.2023.105871>

Hansen-Schirra, S., Abels, K., Signer, S., & Maaß, C. (Eds.). (2021). *The dictionary of accessible communication*. Frank & Timme. <https://doi.org/10.26530/20.500.12657/52603>

Hepburn, G. (2006). *Alternatives to traditional regulation* [Report]. Organisation for Economic Co-operation and Development. <https://web.archive.org/web/20150801112055/https://www.oecd.org/gov/regulatory-policy/42245468.pdf>

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI* [Report]. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Hoffmann-Riem, W. (2020). Artificial intelligence as a challenge for law and regulation. In T. Wischmeyer & T. Rademacher (Eds.), *Regulating artificial intelligence* (pp. 1–29). Springer. https://doi.org/10.1007/978-3-030-32361-5_1

Hoffmeister, K. (2024). The dawn of regulated AI: Analyzing the European AI Act and its global impact. *Zeitschrift Für Europarechtliche Studien*, 27(2), 182–212. <https://doi.org/10.5771/1435-439X-2024-2-182>

IEEE Standards Association. (2020). *IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being* (Active Standard No. IEEE 7010-2020). <https://standards.ieee.org/ieee/7010/7718/>

Kaminski, M. E. (2023). Regulating the Risks of AI. *Boston University Law Review*, 103, 1347–1411. <https://doi.org/10.2139/ssrn.4195066>

Keller, H. (2008). Codes of conduct and their implementation: The question of legitimacy. In R. Wolfrum & V. Röben (Eds.), *Legitimacy in international law* (Vol. 194, pp. 219–298). Springer. https://doi.org/10.1007/978-3-540-77764-9_12

Krafft, T. D., Zweig, K. A., & König, P. D. (2022). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance*, 16(1), 119–136. <https://doi.org/10.1111/rego.12369>

Kusche, I. (2024). Possible harms of artificial intelligence and the EU AI act: Fundamental rights and risk. *Journal of Risk Research*, 1–14. <https://doi.org/10.1080/13669877.2024.2350720>

- Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023). Development of the “Scale for the assessment of non-experts’ AI literacy” – An exploratory factor analysis. *Computers in Human Behavior Reports*, 12. <https://doi.org/10.1016/j.chbr.2023.100338>
- Laux, J., Wachter, S., & Mittelstadt, B. (2024). *Three pathways for standardisation and ethical disclosure by default under the European Union Artificial Intelligence Act*. SSRN. <https://doi.org/10.2139/ssrn.4365079>
- Li, S., Faure, M., & Havu, K. (2022). Liability rules for AI-related harm: Law and economics lessons for a European approach. *European Journal of Risk Regulation*, 13(4), 618–634. <https://doi.org/10.1017/err.2022.26>
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. <https://doi.org/10.1145/3313831.3376727>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777. <https://dl.acm.org/doi/10.5555/3295222.3295230>
- Mantelero, A. (2022). Human rights impact assessment and AI. In A. Mantelero (Ed.), *Beyond data: Human rights, ethical and social impact assessment in AI* (Vol. 36, pp. 45–91). T.M.C. Asser Press. https://doi.org/10.1007/978-94-6265-531-7_2
- Marchant, G. (2019). “Soft law” governance of artificial intelligence. <https://escholarship.org/uc/item/0jq252ks>
- Marsden, C. T. (2011). *Internet co-regulation: European law, regulatory governance and legitimacy in cyberspace*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511763410>
- Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*, 32(2), 241–268. <https://doi.org/10.1007/s11023-021-09577-4>
- Moradi, P., Kambhatla, N., & Sarkar, A. (2021). Measuring and improving faithfulness of attention in neural machine translation. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2791–2802. <https://doi.org/10.18653/v1/2021.eacl-main.243>
- Muchlinski, P. T. (2021). *Multinational enterprises and the law* (3rd ed.). Oxford University Press. <https://doi.org/10.1093/law/9780198824138.001.0001>
- North, D. C. (1990). *Institutions, institutional change and economic performance* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511808678>
- Open Data für Leichte Sprache. (n.d.). *OPEN-LS: Open Data für Leichte Sprache – Ein Forschungsprojekt der HAW Hamburg*. <https://open-ls.entavis.com/>
- Organisation for Economic Co-operation and Development. (2015). *Industry self regulation: Role and use in supporting consumer interests* (Report No. 247; OECD Digital Economy Papers, Vol. 247). OECD Publishing. <https://doi.org/10.1787/5js4k1fjqkwh-en>
- Organisation for Economic Co-operation and Development. (2022). *Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint* (Report No. 341; OECD Digital Economy Papers, Vol. 341). OECD Publishing. <https://doi.org/10.1787/7babf571-en>

Pagallo, U., Casanovas, P., & Madelin, R. (2019). The middle-out approach: Assessing models of legal governance in data protection, artificial intelligence, and the Web of Data. *The Theory and Practice of Legislation*, 7(1), 1–25. <https://doi.org/10.1080/20508840.2019.1664543>

Pointer, M. R. (2003). *New directions—Soft metrology requirements for support from mathematics, statistics and software: Recommendations for the Software Support for Metrology programme 2004–2007* (Report No. CMSC 20/03). National Physical Laboratory. <http://eprintspublications.npl.co.uk/id/eprint/2617>

Regulation 2024/1689. (2024). *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)*. European Parliament and Council. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>

Schallock, R. L., Brown, I., Brown, R., Cummins, R. A., Felce, D., Matikka, L., Keith, K. D., & Parmenter, T. (2002). Conceptualization, measurement, and application of quality of life for persons with intellectual disabilities: Report of an international panel of experts. *Mental Retardation*, 40(6), 457–470. [https://doi.org/10.1352/0047-6765\(2002\)040<0457:CMAAOQ>2.0.CO;2](https://doi.org/10.1352/0047-6765(2002)040<0457:CMAAOQ>2.0.CO;2)

Schmidt, T., & Voenekey, S. (2022). Fostering the common good: An adaptive approach regulating high-risk AI-driven products and services. In S. Voenekey, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge handbook of responsible artificial intelligence: Interdisciplinary perspectives* (1st ed., pp. 123–149). Cambridge University Press. <https://doi.org/10.1017/9781009207898.011>

Schomacker, T., Gille, M., Tropmann-Frick, M., & von der Hülls, J. (2023). Data and approaches for German text simplification – Next steps toward an accessibility-enhanced communication. *Proceedings of the 19th Conference on Natural Language Processing*, 63–68. <https://aclanthology.org/2023.konvens-main.6>

Schomacker, T., Gille, M., Tropmann-Frick, M., & von der Hülls, J. (2024). Self-regulated and participatory automatic text simplification. In M. A. Jabbar, S. Tiwari, F. Ortiz-Rodríguez, S. Groppe, & T. Bano Rehman (Eds.), *Applied Machine Learning and Data Analytics* (Vol. 2047, pp. 264–273). Springer. https://doi.org/10.1007/978-3-031-55486-5_19

Senden, L. A. J. (2005). Soft law, self-regulation and co-regulation in European law: Where do they meet? *Electronic Journal of Comparative Law*, 9(1), 1–27. <https://ssrn.com/abstract=943063>

Shams, R. A., Zowghi, D., & Bano, M. (2023). AI and the quest for diversity and inclusion: A systematic literature review. *AI and Ethics*. <https://doi.org/10.1007/s43681-023-00362-w>

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. <https://doi.org/10.1109/SP.2017.41>

Stuurman, K., & Lachaud, E. (2022). Regulating AI. A label to complete the proposed Act on Artificial Intelligence. *Computer Law & Security Review*, 44. <https://doi.org/10.1016/j.clsr.2022.105657>

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, 361(6404), 751–752. <http://doi.org/10.1126/science.aat5991>

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C. M., Ezer, D., Haert, F. C. V. D., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., Wever, W. D., ... Clopath, C. (2020). AI for social good: Unlocking the opportunity for positive impact. *Nature Communications*, 11(1), 2468. <https://doi.org/10.1038/s41467-020-18488-8>

i.org/10.1038/s41467-020-15871-z

van Schooten, H., & Verschuuren, J. (2008). *International governance and law: State regulation and non-state law*. SSRN. <https://ssrn.com/abstract=1291162>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>

von Grafenstein, M. (2022). Co-regulation and competitive advantage in the GDPR: Data protection certification mechanisms, codes of conduct and data protection-by-design. In G. González, R. Van Brakel, & P. De Hert (Eds.), *Research handbook on privacy and data protection law: Values, norms and global politics* (pp. 402–432). Edward Elgar Publishing. <https://doi.org/10.4337/9781786438515>

Wörsdörfer, M. (2024). Mitigating the adverse effects of AI with the European Union's artificial intelligence act: Hype or hope? *Global Business and Organizational Excellence*, 43(3), 106–126. <https://doi.org/10.1002/joe.22238>

Zhang, W., Dimiccoli, M., & Lim, B. Y. (2022). Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–32. <https://doi.org/10.1145/3491102.3517522>

Züger, T., Faßbender, J., Kuper, F., Nenno, S., Katzy-Reinshagen, A., & Kühnlein, I. (2022). *Civic coding: Grundlagen und empirische Einblicke zur Unterstützung gemeinwohlorientierter KI [Civic Coding: Fundamentals and empirical insights to support AI for the common good]* [Research report]. Initiative Civic Coding vom Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz, Bundesministerium für Arbeit und Soziales, Bundesministerium für Familie, Senioren, Frauen und Jugend. https://www.civic-coding.de/fileadmin/civic-ai/Dateien/Civic_Coding_Forschungsbericht.pdf

Züger, T., & Asghari, H. (2023). AI for the public. How public interest theory shifts the discourse on AI. *AI & SOCIETY*, 38(2), 815–828. <https://doi.org/10.1007/s00146-022-01480-5>

Published by



ALEXANDER VON HUMBOLDT
INSTITUTE FOR INTERNET
AND SOCIETY

in cooperation with



CREATE



centre
— internet
et —
societe



R&I
IN3
Internet
interdisciplinary
Institute
Universitat Oberta de Catalunya



UNIVERSITY OF TARTU
Johan Skytte Institute of
Political Studies