D2.2

Evolution of supranational institutional success criteria in post-2018 AI guidelines

Delaram Golpayegani, Marta Lasek-Markey, Arjumand Younus, Aphra Kerr, Dave Lewis, and Alexandros Minotakis

FORSEE

Forging Successful AI Applications for European Economy and Society

Forging Successful AI Applications for European Economy and Society

D2.2. Evolution of supranational institutional success criteria in post-2018 AI guidelines

Project Information
Project Number: 101177579
Project Title: Forging Successful AI Applications for European Economy and Society - FORSEE
Funding Scheme: HORIZON-CL2-2024-TRANSFORMATIONS-01-06
Project Start Date: 1 February 2025

Deliverable Information	
Title	Evolution of supranational institutional success criteria in post-2018 AI guidelines
Work Package	WP 2 - Mapping of social expectations: institutional/formal approaches to success
Lead Beneficiary	Trinity College Dublin

Funded by the European Union

Due Date	31/01/2026
Revision Number:	V2.0
Authors	Delaram Golpayegani[1], Trinity College Dublin Marta Lasek-Markey[2], Trinity College Dublin Arjumand Younus[3], University College Dublin Aphra Kerr[4], University College Dublin Dave Lewis[5], Trinity College Dublin Alexandros Minotakis[6], University College Dublin
Reviewers	John D. Kelleher Esther Keymolen
Dissemination Level	PU - Public
Deliverable Type	Report

[1] https://orcid.org/0000-0002-1208-186X
[2] https://orcid.org/0000-0002-0183-3982
[3] https://orcid.org/0000-0001-7748-2050
[4] https://orcid.org/0000-0001-5445-7805
[5] https://orcid.org/0000-0002-3503-4644
[6] https://orcid.org/0009-0001-9294-4564

Funded by the European Union

Version History				
Version	Changes implemented by	Revision Date	Reviewed by:	Brief description of changes
V1.0	Delaram Golpayegani, Marta Lasek-Markey, Arjumand Younus, Aphra Kerr, Dave Lewis, Alexandros Minotakis	28/11/2025	John D. Kelleher, Esther Keymolen	Initial version
V2.0	Delaram Golpayegani, Aphra Kerr, Dave Lewis, Arjumand Younus	22/01/2026		Revised based on reviewers' feedback, with major revisions including: addition of themes, refining the terminology, enhancing the discussions of sociology of expectations and topic modelling)

Executive Summary

A number of supranational institutions have published policies and guidelines since 2018 that aim to ensure Artificial Intelligence (AI) systems deployed in Europe are safe, trustworthy and in line with fundamental rights and other shared values.

This study investigates the themes, priorities and dimensions of success set forth by a selected number of European and non-European supranational institutions in documents that are the result of formal policymaking processes and consensus building across different stakeholders. These institutions are some of the most important and powerful supranational bodies shaping formal approaches to defining successful AI.

This study aims to discern the set of baseline criteria for AI success from select policies, guidelines, and regulations issued by European and non-EU supranational institutions from 2018 to 2025. Our sample includes 13 different types of documents of varying lengths. There are three main groups of documents: EU bodies (8), OECD/G20 (2) and UNESCO, UN and Council of Europe (3). Soft instruments of governance (e.g. codes of ethics and guidelines) dominate the corpus with the AI Act the only regulatory instrument examined.

To analyse these documents, we utilise a combination of qualitative thematic analysis and unsupervised approaches to topic modelling, ensuring that a contextual understanding of concepts within documents are captured through human-based analysis and can be used to frame and interpret patterns gleaned from across a large and diverse document set using an unsupervised machine learning algorithm. The findings from thematic analysis and topic modelling were analysed by a multidisciplinary team to identify the themes that signal the success criteria set forth by supranational bodies.

Across the documents analysed in this report there is a positive expectation that **AI innovation** can lead to benefits while harms can be minimised through **AI governance**. There is however a divergence as to the appropriate mode of governance and what harms and rights to prioritise.

There are **four meta-themes** across all the documents: ***AI Technical Issues***, ***AI Uses***, ***AI Risks and Harms*** and ***AI Governance*** into which we have grouped **10 themes**. These are outlined in section 7.1. Key findings include:

1) Across the documents there is growing emphasis on technical features of AI: including computer infrastructure, computing resources, data and cybersecurity as well as AI characteristics framed in narrow technical terms including AI system bias, accuracy and reliability.

2) We identified an evolution in the EU documents over the period under analysis, away from AI ethics/ethical principles and towards trustworthy AI, risk and regulatory compliance. Fundamental rights also come to the fore.

3) Later EU documents place a greater emphasis on both **individual and social risks and harms**, vulnerable groups, health and safety and children's rights but socio-ecological factors, especially environmental and social wellbeing are less evident.

4) By contrast, human rights and diversity and inclusion remain important in documents from the non-EU supranational institutions, especially cluster 3, the UN, UNESCO and the Council of Europe. Many of these documents also identified the importance of sustainable development, environmental protection and democracy as key considerations related to AI innovation.

5) The OECD/G20 documents foreground sustainable development, innovation, the economy and labour and employment.

6) Across the documents we find a strong emphasis on AI governance with multistakeholder collaboration and stakeholders to the fore. However, the mode of governance varies in the documents. The OECD/G20 talks generically about AI governance while later EU documents and the AI Act narrow to focus on EU directed top-down governance and practical issues of legal enforcement, compliance and standards. This signals a move away from self-regulation.

7) Across the four meta-themes we identify ten themes that align with **micro** level expectations on individual organisations, **meso** level expectations on sectors or national government policy and **macro** level expectations that span sectors, states and borders. This multiscalar analysis signals emerging consensus as to who is responsible or will be held responsible for governing and ensuring successful AI.

A wide range of potential sectoral uses of AI are mentioned across the documents but it is clear that the impacts and uses of AI can be positive or negative depending on how the technical aspects and risks are managed and governed.

While there is a shared social expectation that AI innovation can be beneficial and can be governed by stakeholders, there are competing imaginaries as to the underlying values and appropriate mode of governance in the corpus.

Overall, it would appear from these documents that the criteria of success are evolving and vary across the three groups of documents. As such, success will be context dependent. From a European perspective AI success will require compliant governance of technological affordances, protection of rights, and the careful minimisation of risks.

Acronyms

AI	Artificial Intelligence
CSO	Civil Society Organisation
EU	European Union
HLEG	High Level Expert Group
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
OECD	Organization for Economic Cooperation and Development
SME	Small and Medium-sized Enterprise
SOE	Sociology of Expectations
UN	United Nations

Terminology

The following provides the definitions of the key concepts in this deliverable.

Code	A concept that is manually identified in the thematic analysis process as being both present and important within the document.
Document	A formally issued textual artefact, as a whole, issued by supranational bodies, technical standardisation organisations, academic institutions, or professional bodies, and intended to articulate principles, requirements, or procedures relevant to the governance, development, or deployment of artificial intelligence.
Thematic Analysis	A popular method of manual qualitative data analysis that systematically analyses datasets and identifies patterns of meaning.
Theme	A broad interpretable semantic concept derived through expert interpretation, either from the topics generated by a topic modeller or from the codes generated during the thematic analysis process.
Topic Modelling	A statistical method used to identify prevalent topics in large corpora of text using unsupervised machine learning techniques.
Topic	A concept that has been identified by a computational model (e.g., LDA or BERTopic) as a latent concept associated with a subset of documents (in our work a subset of sentences within a document). Topic is typically summarised by salient words that tend to appear when that concept is present.

Table of Contents

Funded by
the European Union

1. Introduction

This deliverable (D2.2) is part of a four-part study in work Package 2 (WP2) of the FORSEE project, shown in Figure 1. Together, these studies investigate the viewpoints and criteria set forth by a range of supranational bodies, technical organisations, academic institutions, and professional bodies in defining successful AI applications. Supranational bodies, technical organisations, academic institutions, and professional bodies collectively play a pivotal role in shaping perceptions of success in AI by establishing standards, guidelines, and ethical frameworks that influence the development, deployment, and evaluation of AI applications. These bodies serve as crucial actors therefore in shaping formal approaches, criteria and social expectations as to AI success.

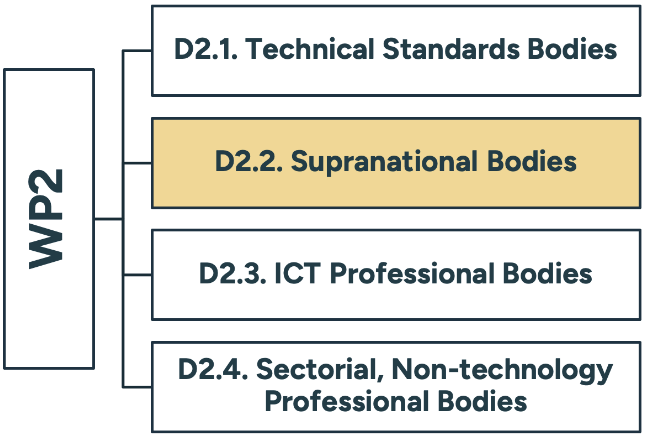


Figure 1. The structure of work package 2 deliverables

For context, the other three parts of WP2 address respectively AI governance outputs from:

- **Technical Standards Bodies (D2.1)**: representing international and national bodies developing technical and process standards for AI,
- **ICT Professional Bodies (D2.3)**: While the field of ICT does not require a professional affiliation to practice, bodies representing ICT professionals internationally that have engaged in guidelines or rules addressing AI governance as part of their broader remit to guide professional practice.

- **Sectorial, Non-technology Professional Bodies (D2.4)**: Bodies representing professionals internationally in sectors that are not primarily ICT-based but which are sufficiently impacted by AI technology to engage in sector-specific AI governance.

This deliverable (D2.2) aims to map the viewpoints and criteria set forth by supranational institutions in formal documents that seek to define success criteria for Artificial Intelligence (AI). The research question we explore in D2.2 is: ***What are the prevalent themes in existing AI policies, guidelines, and regulations published by prominent supranational bodies between 2018 and 2025?***

D 2.2 is focussed on the expression of AI success criteria offered in documents developed by **Supranational Bodies** addressing issues of AI governance. The types of Supranational Bodies studied consist of parties that represent, or who are selected to be representative of, interests of multiple different countries in developing guidelines or rules for the governance of AI. Examples of such bodies are the United Nations, European Commission, G20, the Organisation for Economic Co-operation and Development (OECD).

These institutions were selected because they represent both wide global supranational representation as well as bodies with influence in the European context. This study does not aim to draw any comparison between individual countries, but rather looks at the documents as the result of multistakeholder consensus processes between and across countries that attempt to shape positive and negative social expectations as to how AI systems can be deployed in Europe and beyond.

Overall, D2.2 aims to discern the set of baseline criteria for AI success produced by these different classes of supranational institutions. We seek to identify and categorise recurring themes, priorities, and nuanced dimensions of success criteria outlined in the documents issued by **supranational bodies.** D2.2. therefore aims to shed light on the evolving perspectives and priorities that shape institutional discourse on the responsible development and deployment of AI.

To do so, this study employs a combined methodological approach that integrates thematic analysis, originating from social sciences (Braun & Clarke, 2021; Terry et al., 2017), and unsupervised topic modelling approaches to identify key themes that signal success criteria for AI within a corpus of policies, guidelines, and regulations issued by supranational bodies.

Combining thematic analysis and topic modelling ensures that a contextual understanding of concepts within documents are captured through human-based analysis and can be used to frame and interpret patterns gleaned from across a large and diverse document set using an unsupervised machine learning algorithm.

This study selects documents from 2018 to 2025, as 2018 followed a period of significant negative discourse about AI and whistleblowing about practices in major AI companies, including the Cambridge Analytica scandal (see Kerr et al., 2020). From a Sociology of Expectations perspective one could discern that these negative expectations were impacting upon public and citizen willingness to accept AI in their everyday lives. Thus from 2018 there has been a period of rapid growth in multistakeholder AI policymaking during which a number of supranational institutions have reached agreement on some basic criteria for AI governance. This period is also of sufficient duration for us to discern some shifts in the criteria set forth by these influential entities.

This report specifically aims to identify common themes, both shared and divergent, within the **AI policies, guidelines, and regulations issued by supranational bodies *between 2018 and 2025.*** Addressing our research question, we first create a corpus of the AI policies guidelines, and regulations issued by prominent supranational bodies, including the European Union (EU), the United Nations, OECD, and G20 in the time frame identified.

Particular emphasis is placed in the corpus on EU documents, as it was an early leader, with the EU-High Level Expert Group (HLEG)'s "Ethics Guidelines for Trustworthy AI" (HLEG, 2019a) followed later by the enactment of the EU AI Act (Regulation (EU) 2024/1689, 2024), both of which have significantly shaped the discourse around AI governance both within the EU and globally. Therefore, **EU-issued documents**, both binding and non-binding, are analysed as a distinct group. These documents range from binding legislation, i.e. the EU AI Act, to non-binding instruments such as the guidelines issued by the European Commission in 2025 on prohibited AI practices (European Commission, 2025a).

We also included policies and guidelines issued by non-EU **supranational institutions**, including the OECD, the United Nations, and G20, and these are analysed as a distinct category. These represent attempts to achieve consensus on AI policy at an inter-governmental level but without the expectation that the supranational bodies will be involved in any enforcement of such policies, this instead remaining in the remit of individual states. For states that are EU members, they also participate in consensus building of binding policy through directives and regulations agreed at an EU level. Analysis of the corpus using this EU/non-EU categorisation provides a clear view on the European perspective, as the development of the AI Act represents a significant and currently unique manifestation of the EU's commitment to establishing legally binding success criteria for AI. On the other hand, analysis of non-EU supranational institutional bodies, reflects points of consensus and divergence on AI principles, norms, and governance priorities. This analytical categorisation also captures the differences in policy scope and priorities, normative framing, regulatory ambitions between legal authorities.

While common codes (from manual thematic analysis) and topics (from unsupervised topic modelling) are identified from these two sets of documents separately, we consolidate the results to identify themes and analyse AI success factors articulated by supranational bodies through the lens of sociology of expectations.

The remainder of this deliverable is structured as follows. In section 2, we provide essential information and literature regarding sociology of expectations and AI governance. In section 3, we review the prior work on analysis of policy documents issued by supranational bodies using both manual and automated approaches. Section 4 describes our novel methodology to identify themes that signal AI success within AI documents through combination of topic modelling and thematic analysis approaches. Section 5 presents our corpus of AI policies, guidelines, and regulations issued by supranational bodies from 2018 to 2025. We present the results of our manual thematic analysis and automated machine learning-based topic modelling in section 6. Informed by sociology of expectation, section 7 provides an interpretation of these findings to identify themes. The deliverable is concluded in section 8.

2. Sociology of Expectations and Institutional Approaches to AI Governance

WP2 is informed by the Sociology of Expectations (SoE). This is a strand of the Social Construction of Technology Paradigm (Pinch & Bijker, 1984) that posits technologies are shaped by social forces, including the values, beliefs, interests, and power dynamics of the actors involved in their development and use. Specifically, SoE highlights the prominent role that expectations play in shaping technological processes. Expectations are perceived as socio-technical visions of the future that shape the present. As (Brown et al., 2003), note:

"Expectations mobilize the future into the present, they do so with varying success and according to different time frames and forms of organisational relationship...so while expectations can be formulated as (probabilistic) predictions,...there is always a performative aspect to them."

Within this process, discourses of hope often serve as vectors, embedding promissory futures with real world effects (Brown, 2015). While expectations are primarily articulated through narratives, once they gain traction they "may materialise in experiments and prototypes" (van Lente et al., 2013). By focusing on the potential material impact of promises about the future, scholars working within the SoE paradigm shift attention away from verifying the truth of specific claims towards examining their meaningfulness and their capacity to mobilise resources, organisations, and people to act.

Expectations can be categorised according to their level—micro (e.g. research groups), meso (technological fields), or macro (societal contexts) (van Lente et al., 2013). They can also be distinguished by their content, since "expectations may concern technical, commercial, or societal aspects, and probably a mix of these" (van Lente, 2012). In practice, particular technological artefacts and areas of scientific research are rarely associated with a single, neatly bounded set of expectations. Rather, they tend to attract multiple and overlapping expectations that vary in scope, content, and degree of internal coherence.

In summarising the potential force that expectations yield, Van Lente (van Lente, 2012) notes the following key aspects:

- "First, what expectations do is to raise attention and legitimise investments: a project or programme can be defended by referring to a promising future.
- Second, expectations provide direction to the search processes of science and technology...Typically there are many possible paths while choices have to be made. The optimal direction cannot simply be calculated.
- Finally, there is a coordination effect of expectations...Technical development is not solitary work, but the work of networks of companies and research institutions. When a central control is lacking, as is usually the case, expectations indicate pieces of work and stipulate roles".

AI has a long and well-documented history, characterised by recurrent cycles of hype and subsequent disappointment, often followed by recalibration of expectations in order to sustain funding and public attention (Galanos, 2023). These cycles, commonly referred to as “AI winters” and periods of renewed optimism, reflect the complex interplay between technological capabilities, societal expectations, and industrial interests. In the past three years, the commercial applications of large language models (LLMs) have reignited this pattern, situating AI at the centre of a hype cycle. Hype is further reinforced by the technical characteristics of AI; its adaptability and the promise of applications across multiple industries strengthen initial expectations and increase the chances of sustaining them beyond phases of disappointment (van Lente et al., 2013).

Media narratives and investor discourse frequently emphasise the transformative potential of these technologies, fostering a techno-solutionist ethos in which AI is presented as an inevitable fix for a wide range of social, economic, and scientific challenges (Lindgren & Dignum, 2023). In a similar vein, AI is treated as an “inevitability”, an unstoppable force that is not subject to debate; thusly the techno-deterministic narrative of “permissionless innovation” (Dotson, 2015) is echoed in the discourse around AI, fueling aversion to regulation and democratic governance of emerging technologies. The same effect is created by positive-negative ideal types of expectations that associate AI either with utopian or dystopian visions of the future, obscuring limitations and a grounded debate on AI’s actual impact (Vicsek, 2020).

By contrast, the SoE approach carries a democratisation potential by identifying different expectations, associated with different social positions and perspectives. For example, (Kerr et al., 2020) identify a change in expectations surrounding AI when focus shifts from the individual to the societal level as well as a demand from the public for public authorities to govern the design and use of steering of AI by public authorities. This finding illuminates potential conflicts between organisational-professional expectations surrounding AI and broader societal concern and ambivalent positions (Kerr et al., 2020). In order to address potential conflicts, a more grounded perspective is necessary, including insight into a diverse range of stakeholders’ expectations (Vicsek, 2020) as well as a disengagement from performative understanding of ethics (Kerr et al., 2020), that tend to embed them into the business strategies of AI developers (Steinhoff, 2024).

The Sociology of Expectations, in common with SCOT, identifies that different social groups or stakeholders may have competing narratives or expectations (Jasanoff & Kim, 2009). These narratives and expectations may stabilise over time, and some narratives, concepts and meanings may become dominant, and win out at a particular moment in time over alternative views and perspectives. Social expectations can be positive or negative, and they can be merely performative (discursive) or have real impact and shape actions in the world (Kerr et al., 2020). Additionally, the concept of **social imaginaries** provides a complementary theoretical lens to the sociology of expectations for interpreting how different social groups imagine and understand AI, as well as how it is used, governed, and made meaningful. Emerging in STS in relation to technologies such as nuclear power (Jasanoff and Kim, 2009) and the internet (Mansell, 2012)social imaginaries refer to collectively held, negotiated, and stabilised visions of how technologies are expected to function in society, including how responsibilities and benefits should be distributed.

(Mansell, 2012) notes that despite the origins of the internet, a market-based imaginary, with governance from above, has become dominant and alternative imaginaries based on the commons, networked governance and governance from below have become less valued. Finally, as Bareis and Katzenburg note (Bareis & Katzenbach, 2022) in the original conceptualisation of social imaginaries the role of the state and state power is crucial to understanding the power of state driven social imaginaries embedded in policies and regulations.

In the context of AI, supranational bodies are proposing competing imaginaries to provide legitimation for AI innovation and set the terms of engagement for policy discourses including setting regulatory priorities and assumptions about appropriate modes of governance, oversight and intervention. In analysing the documents of supranational bodies we are analysing documents which are backed by political and discursive power, and in some cases legal power. Many are the result of **formal** policymaking processes that involve explicit processes of engagement with a range of stakeholders (e.g. foresight exercises, member voting) and varying degrees of public participation (e.g. public surveys). These processes aim to improve legitimacy and trust in policies and the resulting documents should be viewed as consensus documents. Each document is attempting to reduce uncertainty about the future, and guide AI innovations in a particular direction. Some documents and concepts may have more impact than others in the long run, but in our analysis we are focussed on what themes are articulated in these documents and how these may relate to each supranational body. These formal mechanisms and documents contrast with more informal documents and position statements from the media, on social media and from companies or individuals which are analysed elsewhere in the FORSEE project.

A final consideration that emerges from historical studies in science and technology studies is that despite expectations that emerging technologies may significantly change economies and societies, many technologies fail in the marketplace. Thus, understanding negative expectations and the role that policies and regulations can play in reassuring users and mitigating potential harms is crucial to success in the marketplace. **Governance** has emerged as a key term to signal that there can be multiple forms of regulation (e.g. statutory, co-regulation and self-regulation) that operate at multiple levels, from local to regional to global. Thus, governance is often taken to include vertical (state, regional, supranational) and horizontal actors (private sector). In many media and communication fields there has been a move from top down government and statutory regulation to co-regulation with a variety of non-state stakeholders over the past decades, particularly in the EU since the 2000s (Puppis et al., 2024). In social media and some information technology fields there has been a move from self-regulation to co-regulation. AI technologies have emerged from a range of companies in sectors that have historically been self-regulated. Supranational AI document analysis prior to 2018 points to a proliferation of national strategies, codes of ethics, principles and guidelines. However, the period of time under analysis in this deliverable saw the introduction of the first significant AI legislation in Europe and a significant discourse on what forms of governance and factors are appropriate to ensure AI success.

3. Prior Work Analysing AI Policies, Guidelines and Regulations

There is a body of work that analyses the content of AI policies, guidelines, and regulations to determine the degree of consensus (or divergence) as well as the gaps in AI governance and ethical AI principles within the growing body of AI policies. These studies apply qualitative thematic analysis, quantitative content analysis (often using topic modelling algorithms such as LDA and BERT), or a combination of both. In the following, we first review the existing studies that conducted manual thematic analysis. Then, we refer to some pioneering work that utilises automated approaches from machine learning and natural language processing (NLP) to analyse policy documents. Finally, we review the studies that, similar to our work, integrate manual thematic analysis and automated topic modelling approaches.

3.1. Analysis of Existing Thematic Reviews of AI Policies and Guidelines

Prior to proceeding with the empirical component of the project, we conducted a preliminary review of the existing academic literature (both peer-reviewed and pre-prints) on the topic of AI governance with a view to identifying some key recurring themes. Results of this literature review are briefly summarised in appendix A.

Early reviews, such as (Jobin et al., 2019) and (Hagendorff, 2020), approached the task of identifying key ethical principles underpinning AI governance descriptively, listing frequently mentioned principles without detailing how themes were derived. Both identify transparency, fairness, accountability, privacy, and human control or autonomy as recurring anchors. More structured analyses, such as (Fjeld et al., 2020) and (Palladino, 2023), systematise these principles into distinct clusters or hierarchies. Fjeld's eight thematic groupings, spanning privacy, accountability, safety, transparency, fairness, human control, professional responsibility, and human values, demonstrate an effort to map ethical aspirations to operational domains. Palladino's NVivo-based thematic coding reaches a similar configuration, framing principles as nested under overarching values and requirements. (Birhane et al., 2024) compare academic and non-academic audit practices, situating "accountability" within institutional and socio-technical contexts rather than as an abstract principle. Their findings suggest that meaningful oversight depends less on who audits and more on how audits are designed, communicated, and embedded within power relations. Similarly, (Reuel et al., 2025) shift focus toward the taxonomy of technical capacities: assessment, verification, security, and ecosystem monitoring, bridging normative discourse with concrete mechanisms for implementation. Quantitative text analyses, such as (Roche et al., 2023), reveal the dominance of data-centric and human-centric terms in European AI discourse ("data", "human", "rights", "public", "law"), confirming a continued preoccupation with legal and societal embedding of AI systems.

Overall, the most recurring principles underpinning ethical AI governance were: 1) transparency, 2) accountability, 3) fairness, 4) privacy, and 5) human oversight.

3.2. Discovering Topics from AI Policies and Guidelines

This section reviews the body of work that employs topic modelling approaches to identify prevalent topics from AI policies and guidelines. The focus of this section is on the technical and methodological aspects of identification of topics, as these inform the methodology we used in this work.

A pioneering work (Papadopoulos & Charalabidis, 2020) within the realm of NLP-based AI policies' analysis uses topical distributions from LDA to perform clustering over 12 national AI strategies, and use the cluster information to determine the strategic priorities of various governments in the AI race.

It is important to note that this analysis was performed over some of the earliest available documents covering strategic directions laid by national governments with respect to AI; and hence there is also more coverage on leadership in AI deployment for innovation. Another work built on LDA-based analysis and visualisations mines China's data governance policies to discover evolution paths within these (Yang, 2022). One of the uncovered themes **is AI-based decision-making for data governance**. Similarly, (Wang et al., 2025) pursue via structured topic models a comparative analysis of AI policies from China, EU and US uncovering significant differences in priority areas within the AI policymaking landscape.

(Roche et al., 2023) used a combination of automated data analytics tools and manual content analysis on a corpus of AI policies and governance frameworks to investigate the inclusivity of ethical AI approaches. Specifically, inclusivity here focuses on the contribution of voices from the Global South within the AI ethics discourse. (Schiff et al., 2021) analysed the degree of global consensus on ethical AI principles by manual coding of AI documents published by public sector, private organisations, and NGOs, followed by a quantitative analysis of the coding. Works by Saheb and Saheb (Saheb & Saheb, 2023, 2024) perform a mixed method analysis of AI ethics guidelines (i.e., application of topic modelling followed by manual extraction of themes) issued by national governments to identify the most pressing concerns with respect to the ethical aspects of AI with the most worrying aspect being emergence of marginalised AI policy topics specifically when it comes to concrete governance recommendations.

A very recent work by (Kajava et al., 2025) attempts to track the discourse surrounding increased risks in the AI landscape following the launch of ChatGPT. Their method comprises the use of guided BERTopic in conjunction with qualitative content analysis; and this was done to ensure scalability and trust in the analysis of the discourse. Similarly, (Suter et al., 2025) track AI discussions in various parliamentary debates through a combination of BERTopic and manual classification into interpretable themes by two independent coders. They specifically analyse debates from US Congress, EU Parliament, Parliament of Singapore, and Swiss Federal Assembly and this helps them uncover distinct national priorities in a post-LLM world. Essentially, the findings of both works (Kajava et al., 2025; Suter et al., 2025) reveal a **greater focus on increasing literacy and ethical awareness** within the adoption of AI policies in a world that has been shaken by software such as ChatGPT.

A key difference between our analysis and the existing ones lies **in the way we focus on the progress made by significant international organisations with concrete implications for the EU's framing of AI governance**. While similar to the analysis by (Kajava et al., 2025), this represents a deep analysis of emerging themes within AI governance regulatory landscape; and pursuing such an analysis through a study of major actors in the area (i.e., supranational bodies pursuing AI governance and regulatory goals) **significantly helps in understanding EU's competitiveness and unique positioning in the area of assessing AI success criteria**. Moreover, our analysis is conducted by experts from law, computer science and social science, resulting in **refined themes** that effectively capture the complexities and nuances of supranational policymaking efforts regarding AI governance.

4. Methodology

In this section, we introduce how we selected our corpus of documents and how our methodology for analysing these documents combined manual thematic analysis, automated topic modelling, and expert analysis.

4.1. Selecting Documents and Compiling the Corpus

Documents to be examined in this deliverable are selected based on the following criteria:

- Scope: Policies, guidelines, or regulations with an explicit focus on AI.
- Timeline: Documents published during the period of 2018 - 2025, corresponding to

the period in which AI governance emerged as a policy area within the EU and globally.
- Issuing body: Major supranational bodies who have issued AI related governance documents in the specified time period. There are three main groups of issuing bodies: EU bodies (8 documents), OECD/G20 (2 documents) and UNESCO, UN and Council of Europe (3 documents).
- Legal status: both binding (e.g. regulations) and non-binding (e.g. policies, recommendations, and guidelines) are included in the corpus. Soft instruments of governance (e.g. codes of ethics and guidelines) dominate the corpus with the AI Act the only regulatory instrument examined.
- Length of document: No restrictions were applied in respect to the length of the documents.

The corpus of selected documents is detailed in section 5.

4.2. Combining Thematic Analysis and Topic Modelling

The terms "**code",** "**topic**", **"theme"** are used extensively throughout this deliverable. Although they may appear as synonyms, they are treated as distinct concepts and are not used interchangeably. To clarify this distinction, in the following, we provide background information about how we combine **thematic analysis** and **topic modelling** in this deliverable.

In this deliverable, we consider a **document** as a formally issued textual artefact published by a supranational body regarding AI. Such documents can range from regulations to policies to official guidelines and are collectively compiled within a corpus (see section 5).

In our methodology, we first apply **thematic analysis,** which is a popular method of qualitative data analysis that systematically organises datasets, helps to identify patterns of meaning (Squires, 2023). First described in the 1970s by Houlton – albeit the term itself had been in use even earlier – it became more prominent in the late 1990s with researchers like Boyatzis and Hayes (Braun & Clarke, 2021; Terry et al., 2017). In recent times, thematic analysis has been understood as an umbrella term for different approaches ranging from reflexive thematic analysis to codebook and coding reliability approaches. No coding reliability tests were used in this project although the combining of thematic analysis and topic modelling required negotiation between a fully qualitative and a more quantitative analysis of the data. The disciplinary backgrounds of the team and their subjectivities were crucial however to the initial manual coding, the fine tuning of the topic modelling and the final thematic analysis.

In our work, we use manual thematic analysis to inductively code a sample of our corpus of documents. The output of this process is a set of **codes. Code refers to a concept that is manually identified as being both present and important within the domain.** This form of thematic analysis relies on the coder's judgements and expertise. It produces contextually informed codes grounded in the documents.

However, manual thematic analysis is a time-consuming and resource-intensive process which makes it difficult to conduct thematic analysis of a large corpus of documents. Further, it does not provide insights on statistical frequency of the codes.

Addressing these specificities, we also employ **topic modelling,** which is a statistical method used to identify prevalent topics in large corpora of text. Topic modelling uses unsupervised machine learning techniques for uncovering topics within large sets of documents. There are different unsupervised topic models (see (Churchill & Singh, 2022)) for evolution of such models). Currently, BERTopic (Grootendorst, 2022) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are the most popular topic models. The output of these models are clusters of keywords representing topics. **Topic** is a concept that has been identified by a computational model (e.g., LDA or BERTopic) as being present in the corpus consisting of multiple documents. In both LDA and BERTopic, a topic can be understood as a latent concept associated with a subset of documents (in our work subset of sentences within a document as explained later) and typically summarised by salient words that tend to appear when that concept is present. It should, however, be noted that the underlying mathematical object defining a topic differs across LDA and BERTopic (a probability distribution over words in LDA versus a cluster of document embeddings in BERTopic).

Though topic modelling is usually applied to a set of documents, we apply it to each document separately to gain insights into each single document. Therefore, in this work topics are associated with a subset of sentences within a document.

Finally, to gain insights, we interpret the codes and topics through the lens of sociology of expectation to identify **themes. Theme** is a broad interpretable semantic concept derived through expert interpretation, either from the topics generated by a topic modeller or directly from the codes generated during the thematic analysis process. By illustrating the overall methodology wherein thematic analysis and topic modelling are integrated, Figure 2 shows the key concepts in the methodology and how they are related to each other.

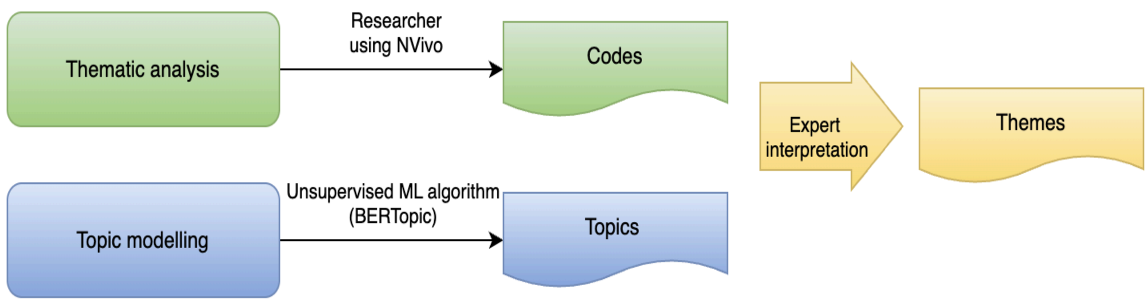


Figure 2. Overall methodology that combines thematic analysis and topic modelling to identify themes

4.2.1. Thematic Analysis

The qualitative analysis of a subset of the corpus (6 documents out of 13), including the EU AI Act and the HLEG "Ethics Guidelines for Trustworthy AI", and UNESCO's guideline on the ethics of AI (UNESCO, 2021) were coded with the use of NVivo software (see section 6 for the findings).

As explained by (Dhakal, 2022), NVivo is a CAQDAS (Computer-assisted qualitative data analysis software) programme that assists, rather than replaces, a human researcher. NVivo analysis is, thus, expert-led and in this project it was performed by a legal researcher. The documents were imported into NVivo and thematic coding was performed by selecting a section of text and then tagging it with a code. The codes were assigned inductively by the human researcher.

4.2.2. Topic Modelling

The topic modelling component of the methodology comprises two key steps:

- **Step 1**: Determining the suitable unsupervised topic model through comparisons of automated results with manual thematic analysis results conducted for a subset of

corpus, including the AI Act, the HLEG "Ethics Guidelines for Trustworthy AI", and UNESCO's guideline on the ethics of AI.
- **Step 2**: Identifying the prevalent topics by analysing the results generated by the selected topic modeller for each document.

Step 1 - Determining Suitable Topic Modeller: Given the variety of topic modelling techniques, this step aims to identify the most suitable topic model algorithm for the objective of our study. As shown in Figure 3, this step runs in parallel with thematic analysis and applies several different topic modellers on the subset of the corpus mentioned in the previous section. For this, in consultation with NLP experts, we initially considered Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and BERTopic (Grootendorst, 2022) among the existing topic models. As mentioned earlier, both LDA and BERTopic uncover topics that occur within a corpus. While they both leverage unsupervised machine learning techniques, i.e. they do not need label data, they fundamentally differ in how they derive topics. LDA is a probabilistic model and identifies latent topics based on a probability distribution over words whereas BERTopic leverages language embeddings to generate semantically coherent topics based on meaning rather than word frequency alone.

These models are usually used to identify common topics in a collection of documents; however, this work only deals with a limited number of documents that should be processed **individually**. Therefore, to enable applying topic modelling algorithms on a single document, each document was segmented into a set of **sentences. This means that topics identified by the topic modeller correspond to sentences within one single document, rather than the entire corpus of documents.** This helps to gain detailed insights to each specific document in the corpus.

For implementation, we used `PyMuPDF` library in Python to extract the text from the PDF file and then split the text into sentences and stored them in a CSV file. After applying appropriate pre-processing steps required for LDA and BERTopic, we applied both of these topic modelling algorithms to the AI Act, the HLEG "Ethics Guidelines for Trustworthy AI", and UNESCO's guideline on the ethics of AI. We then compared the outputs of LDA and BERTopic against the findings of manual thematic analysis. The comparison indicates the outputs produced by BERTopic are more consistent with the codes derived through thematic analysis than the outputs produced by LDA. On this basis, we conclude that **BERTopic outperforms LDA in the context of AI policy texts**, offering higher accuracy and better interpretability. This is aligned with the findings of (Kajava et al., 2025), which compared the interpretability of these two topic modellers for EU-related documents and news articles on AI.

We also experimented with recent LLM-based topic models (Pham et al., 2024) but their use involves heavy costs in the two stages of tokens' processing and topical refinement with the results essentially not showing a major improvement over BERTopic. Certain LLM enhancements of topic models (Li et al., 2025) may help assess AI policies and guidelines through a "sociology of expectations" lens, however we consider the investigation of this as a topic for future research.

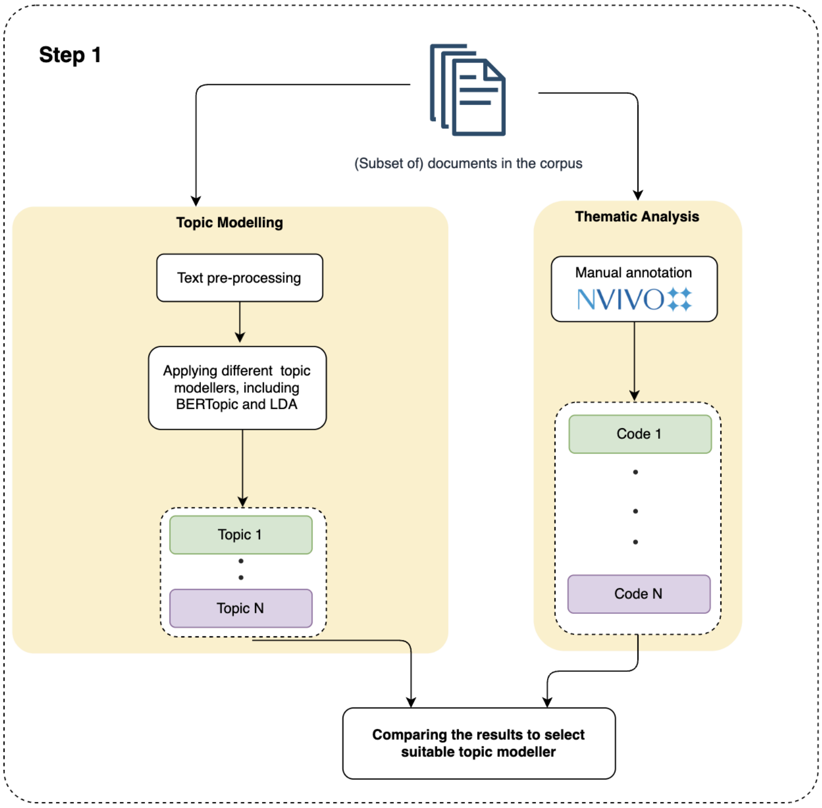


Figure 3. Step 1 of the topic modelling part of the methodology to identify suitable topic modeller

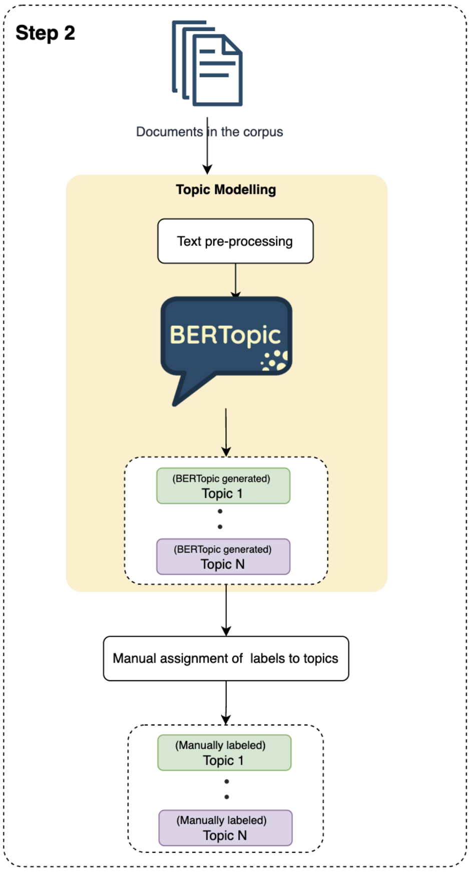


Figure 4. Step 2 of the topic modelling part of the methodology to identify topics from AI policies, guidelines, and regulations included in the corpus themes by combining BERTopic and expert analysis

Step 2 - Applying BERTopic to Determine Topics: As shown in Figure 4, this step comprises three phases to identify topics from each document in the corpus:

1. **Text preprocessing**: Before application of BERTopic, we employed common NLP approaches to apply minimal preprocessing using the `nltk` library for:
    - Lowercasing,
    - Removing non-alphabetic tokens (e.g. punctuation marks),
    - Removing tokens that consist solely of numbers, and
    - Removing tokens shorter than two characters.

Note that we did not remove stop words or perform lemmatisation prior to embedding, given that BERTopic uses transformer-based embeddings, which require the text to be persevered for better accuracy. As explained later, we removed the stop words, after generating embeddings, using the `CountVectorizer` when initialising BERTopic.

2. Applying BERTopic: Since this work uses BERTopic at the sentence-level, the description of how BERTopic works is made consistent with the wording used throughout this deliverable. The following describes how BERTopic generates topics:
    a. Each sentence is converted to its embedding representation (numerical representation) using a pre-trained language model. We used `all-MiniLM-L6-v2`[7] which is a pre-trained mode that converts sentences and paragraphs into vector-based representations.

```python
embedding_model = SentenceTransformer("all-MiniLM-L6-v2")
```

    b. To group sentences into semantically similar clusters, the dimensionality of the resulting embeddings is reduced using UMAP (Uniform Manifold Approximation and Projection) and HBDSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Since applying BERTopic default parameters on short documents did not generate any coherent topics, we identified two hyperparameter configurations: one for longer texts and another for shorter documents. These configurations only differ in the `HDBSCAN` clustering model, with a lower minimum number of samples per word cluster is set for shorter documents. These hyperparameter configurations were used uniformly (based on the length of document) to ensure consistency and reproducibility of results.

[7] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Python
```python
map_model = umap.UMAP(n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine',
random_state=42)
hdbscan_model = HDBSCAN(min_cluster_size=10, min_samples=5) #longer documents
#hdbscan_model = HDBSCAN(min_cluster_size=10, min_samples=5) #shorter documents
```

c. For topic representations, we used `CountVectorizer` to enable frequency analysis, without down-weighting the common words in the document. Instead, we removed some common words, such as "AI", "Artificial Intelligence", and latin numerals in this step.

Python
```python
vectorizer_model = CountVectorizer(ngram_range=(1, 1), stop_words= stop_words)
```

The snippets used for preprocessing and topic modelling are available on GitHub under the MIT licence at https://github.com/DelaramGlp/forsee_ topicmodelling. The source code and the results are available in an archived release with a persistent digital object identifier (DOI): https://doi.org/10.5281/zenodo.18303112.

3. Topic labelling: An example of BERTopic output for UNESCO's Recommendation on the Ethics of Artificial Intelligence in Figure 5. In this, "topic" denotes a unique numeric identifier for each topic generated by the model. The "Count" column shows the frequency of each topic, i.e. the number of sentences assigned to the topic. "Name" shows a short topic label, however, as shown in the figure, BERTopic does not automatically assign a semantic label and only shows top keywords. "Representation" provides a list of keywords representing the topic and "Representative Sentences" is a set of sentences that are associated with the topic.

   As noted earlier, the *Name* value generated by BERTopic does not constitute meaningful semantic labels for topics. Consequently, assigning interpretable labels as topics was conducted manually. In our work, two researchers involved in WP2 (see their profile in the positionality statement) independently labelled topics based on the keywords. The proposed labels were then discussed in a joint session and consolidated into their final form. An example of label assignments is shown in

Figure 6, with the “Topic Label” demonstrating the agreed-upon topic label.

Topic	Count	Name	Representation	Representative_Sentences
0	21	0_instruments_translation_substantive_agreements	['instruments', 'translation', 'substantive', 'agreements', 'languages', 'territory', 'compendium', 'international', 'organisation', 'challenges']	['the oecd member countries are australia austria belgium canada chile colombia costa rica the czech republic denmark estonia finland france germany greece hungary iceland ireland israel italy japan korea latvia lithuania luxembourg mexico the netherlands new zealand norway poland portugal the slovak republic slovenia spain sweden switzerland türkiye the united kingdom and the united states', 'oecd legal instruments since the creation of the oecd in around substantive legal instruments have been developed within its framework', 'all substantive oecd legal instruments whether in force or abrogated are listed in the online compendium of oecd legal instruments']
1	18	1_policy_governments_observatory_initiatives	['policy', 'governments', 'observatory', 'initiatives', 'stakeholders', 'social', 'community', 'academia', 'workers', 'expertise']	['the network of experts one ai is an informal group of ai experts from government business academia and civil society that provides policy expertise and advice to the oecd', 'the working party supports the implementation of the oecd recommendation on artificial intelligence develops practical guidance for implementation provides forum and online hub for exchanging information on ai policy and activities through the policy observatory and foster and interdisciplinary dialogue through the expert groups of the network of experts', 'to provide an inclusive forum for exchanging information on ai policy and activities and to foster and interdisciplinary dialogue the oecd launched the ai policy observatory as well as ii the informal oecd network of experts on ai one ai in february']
2	15	2_recommendation_dpc_developments_technological	['recommendation', 'dpc', 'developments', 'technological', 'policy', 'international', 'revised', 'committee', 'governance', 'facilitate']	['further work to support the implementation of the recommendation in addition to reporting to the council on the implementation of the recommendation the dpc is also instructed to continue its work on ai building on this recommendation and taking into account work in other international fora such as unesco the european union the council of europe and the initiative to build an international panel on ai see intelligence and', 'instructs the digital policy committee through its working party on ai governance to continue its important work on artificial intelligence building on this recommendation and taking into account work in other international fora and to further develop the measurement framework for ai policies develop and iterate further practical guidance on the implementation of this recommendation to meet evolving developments and new policy priorities provide forum for exchanging information on ai policy and activities including experience with the implementation of this recommendation and to foster and interdisciplinary dialogue to promote trust in and adoption of ai and report to council in consultation with other relevant committees on the implementation dissemination and continued relevance of this recommendation no later than five years following its revision and at least every ten years thereafter', 'on the basis of the report to council on its implementation dissemination and continued relevance the recommendation was revised by the oecd council meeting at ministerial level on may to reflect technological and policy developments including with respect to generative ai and to further facilitate its implementation']

Figure 5. An excerpt of the output generated by BERTopic for OECD recommendations

Topic	Count	Representation	Topic Label	R1_Topic	R2_Topic
0	78	['systems', 'democratic', 'learning', 'examples', 'validate', 'interacting', 'knowledge', 'improve', 'environment', 'behaviour']	N/A	democracy	?
1	50	['vulnerable', 'groups', 'factors', 'disabilities', 'social', 'situations', 'children', 'education', 'exclusion', 'accessibility']	Vulnerable groups	vulnerable groups	vulnerable groups
2	43	['rights', 'intelligence', 'guidelines', 'ethics', 'values', 'principles', 'ethical', 'charter', 'citizens', 'global']	Ethical AI	ethical guideline	ethical AI
3	43	['bias', 'fairness', 'unfair', 'diversity', 'prejudice', 'biases', 'substantive', 'hr', 'accountability', 'inclusive']	Bias, Fairness, Diversity and inclusion	Bias, fairness	bias
4	40	['tensions', 'principles', 'harmony', 'endeavour', 'glossary', 'relevance', 'requirements', 'contexts', 'distinctions', 'paragraphs']	N/A	??	?
5	36	['risks', 'safety', 'malicious', 'risk', 'attacks', 'mitigate', 'security', 'foreseen', 'resilience', 'vulnerabilities']	Risk, safety and security risk	safety and security	risk
6	35	['ethics', 'compliance', 'ethical', 'laws', 'moral', 'guidance', 'normative', 'principles', 'philosophy', 'standards']	ethical AI, compliance	normative guideline, standard	ethical AI / compliance
7	33	['rights', 'charter', 'treaties', 'convention', 'disabilities', 'articles', 'directives', 'copyright', 'legal', 'enforceable']	rights	(fundamental) rights enforcement	rights
8	33	['privacy', 'protocols', 'consumer', 'gdpr', 'processing', 'governance', 'integrity', 'qualified', 'officer', 'prevention']	privacy	privacy	privacy
9	32	['ethical', 'lawful', 'trustworthy', 'components', 'ensuring', 'compliance', 'regulations', 'norms', 'principles', 'securing']	ethical AI, compliance	regulatory compliance	trustworthy AI
10	31	['stakeholders', 'participation', 'organisations', 'stakeholder', 'developers', 'develop', 'deployers', 'limitations', 'capabilities', 'practitioners']	stakeholders	stakeholder participation	stakeholders

Figure 6. An example of assigning of labels to the outputs generated by BERTopic by experts

(Experts #1 and #2 as described in the positionality statement) for the HLEG's trustworthy AI guideline

4.2.3. Interpreting Codes and Topics to Identify Themes

Following inductive manual coding and automated topic analysis the research team went through a process of iterative interpretation aimed at generating, refining and naming themes. This process of meaning making involved both colour coding of topics and codes, clustering, referring back to the original texts and a structured analysis. In line with prior work on the Sociology of Expectations, we also categorised the themes according to whether the themes primarily referred to actions that might be taken by actors at a micro, meso and macro levels. The multidisciplinary team then reviewed, discussed and refined the themes and clustered them into meta-themes guided by the overall research question, our disciplinary backgrounds and our grounding in the data and the prior literature.

5. Corpus of AI Policies, Guidelines, and Regulations Issued by Supranational Bodies

The documents analysed in this deliverable include key AI policies, guidelines, and regulations issued by supranational bodies between January 2018 and August 2025, capturing the period from early ethical and trustworthy AI guidelines through to the recent regulation setting efforts.

There are three main groups of documents: EU bodies (8), OECD/G20 (2) and UNESCO, UN and Council of Europe (3). Accordingly, our corpus includes legislation, official guidelines, and codes of practice directly related to AI and issued by relevant EU institutions. Soft instruments of governance (e.g. codes of ethics and guidelines) dominate the corpus with the AI Act the only regulatory instrument examined. All of the documents were published in English, which we used for our analysis.

As noted earlier, there is a particular emphasis on EU documents in the corpus given the relatively long history of EU initiatives on AI trustworthiness and governance that has evolved into the first binding AI regulation. In addition to EU documents, the corpus includes AI policies and guidelines from prominent supranational bodies, including the Council of Europe, UN, UNESCO, OECD, and G20. The list of documents that constitute our corpus is provided in Table 1.

The resulting corpus encompasses documents with varying legal statuses ranging from a political commitment to binding legal force. The only directly effective legal measure among them is the EU AI Act (Regulation (EU) 2024/1689, 2024).

The EU chose to lay down the AI governance rules via a Regulation, a type of secondary legislation which as per Article 288 of the Treaty on the Functioning of the European Union (TFEU) does not require any further implementing measures from the Member States. Moreover, once in force, it enjoys what is known as direct horizontal effect, i.e., individuals may rely directly on its provisions in ordinary courts of the Member States. Unlike any other instrument of international law, EU Regulations are automatically binding not only on the governments, but on all persons operating on the EU single market, including private companies. Following its enactment in 2024, the EU AI act has been accompanied with a series of additional implementing measures, such as the European Commission Guidelines on prohibited AI practices (European Commission, 2025a) and the Code of practice for General-Purpose AI models (European Commission, 2025c), as well as relevant Guidelines on the scope of practice (European Commission, 2025b). It should be noted that the AI Act is not the only legal instrument in the EU affecting development, deployment, and use of AI systems. Other components of the EU digital acquis, such as the General Data Protection Regulation (GDPR) (EU 2016/679, 2016), Digital Services Act (DSA) (EU 2022/2065, 2022), and Digital Governance Act (DGA) (EU 2022/868, 2022), do have implications for AI. However, these are not included in the corpus as their regulatory scope does not directly focus on AI per se.

The proposal for the EU AI Act was preceded by "Ethics guidelines for trustworthy AI" (HLEG, 2019a) issued in 2019 by the High-Level Expert Group, an independent expert group that was set up by the European Commission the previous year. The non-binding guidelines defined seven requirements for "Trustworthy AI" (human agency & oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; accountability), and laid down the groundwork for the EU AI Act. The ethics guidelines were later accompanied by the 'Assessment List for Trustworthy AI' tool (*The Assessment List for Trustworthy*

Artificial Intelligence (ALTAI) for Self Assessment, 2020), as well as 'Policy and investment recommendations for trustworthy AI' (HLEG, 2019b) and 'Sectoral considerations on policy and investment recommendations for trustworthy AI' (HLEG, 2019c).

Outside of the EU, the OECD Council Recommendation on AI (OECD, 2019) was the first intergovernmental document on AI, adopted at ministerial level on the proposal of the Committee on Digital Economy Policy in 2019 and revised in 2023-2024 to accommodate the most recent AI developments. It also served as a basis for the G20 AI Recommendation (G20 Trade Ministers and Digital Economy Ministers, 2019). The Recommendation was signed by all 38 OECD Members and eight non-members. Like many instruments of international law, the OECD Recommendation is addressed to governments and not to individual persons. Furthermore, as a 'Recommendation', it is not legally binding, but it represents a political commitment of the adherents to make their best efforts to implement it.

Closely linked to the OECD Recommendation is the G20 AI Recommendation (G20 Trade Ministers and Digital Economy Ministers, 2019) issued by the Group's leaders at the 2019 summit in Osaka. The Recommendation reflects a political commitment signalling an emerging global consensus around AI governance, yet it carries no direct legal obligation on organisations or states.

UNESCO's Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021) was adopted in November 2021 at UNESCO's General Conference in Paris. Upon proclamation, UNESCO recommended that Member States apply it on a **voluntary basis** through legislative or other measures. It also urged the Member States to engage with all relevant stakeholders, including companies, as well as bring it to the attention of national authorities, research and academic organizations involved in AI. The UNESCO Recommendation, therefore, is a commitment on the part of the Member States, which has been signed by 193 countries. Importantly, unlike Conventions, a UNESCO Recommendation is not legally binding, but it solidifies the political commitment to implement solutions in AI governance. Furthermore, like the majority of international law, it is addressed to governments of the signatories, and not to private actors, such as tech companies, albeit it does offer 'ethical guidance' to all AI actors, whether public or private. More recently, the United Nations also published a report on 'Governing AI for Humanity' (United Nations, 2024).

The most recent supranational initiative was the 2024 Council of Europe (Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law,

Funded by the European Union

2024). To date, it has been signed by the European Union and a number of other Council of Europe countries (e.g. the UK, Switzerland, Ukraine), as well as some non-members, such as the USA, Canada and Japan. Importantly, it has not yet entered into force as it has not been ratified by any of the signatories, i.e. it is non-binding as of Jan. 2026. However, once ratified, it will constitute the first-ever international legally binding treaty in the field of AI, with 17 countries, in addition to the European Union, having signed it. The Framework Convention is indicated to be technology-neutral, meaning that it does not focus on the technicalities of AI, but rather shifts the focus onto democracy, rule of law, and the impacts of AI.

We initially considered the AI, Data, Robotics Association (ADRA) as a supranational body. However, it was excluded from our corpus since it is funded by private European organisations in partnership with the EU and is not situated within the remit of supranational bodies.

Table 1. AI policies, guidelines and regulations included in the corpus

ID	Document	Issuer	Type	Year
SuB_01	Policy and investment recommendations	EU HLEG	Recommendation	2019

Funded by the European Union

ID	Document	Issuer	Type	Year
	for trustworthy AI			
SuB_02	Ethics guidelines for trustworthy AI	EU HLEG	Guideline	2019
SuB_03	Assessment List for Trustworthy AI (ALTAI)	EU HLEG	Guideline	2020
SuB_04	Sectoral considerations on policy and investment recommendations for trustworthy AI	EU HLEG	Recommendation	2020
SuB_05	EU AI Act	EU Council and Parliament	Legislation	2024
SuB_06	Commission Guidelines on prohibited AI practices established by the AI Act	European Commission	Guideline	2025
SuB_07	Commission Guidelines on the scope of the obligations for general-purpose AI models established by the AI Act	European Commission	Guideline	2025
SuB_08	EU Code of Practice for General-Purpose AI Models, including transparency, copyright, and safety and security chapters	European Commission	Code of practice	2025
SuB_09	Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law	Council of Europe	Treaty	2024
SuB_10	OECD, Recommendation of the Council on Artificial Intelligence	OECD	Guideline	2019-amended in 2024
SuB_11	UNESCO Recommendation on the Ethics of Artificial Intelligence	UNESCO	Guideline	2022
SuB_12	United Nations Governing AI for Humanity: Final Report	United Nations	Guideline	2024
SuB_13	G20 Ministerial Statement on Trade and Digital Economy	G20	Guideline	2019

6. Findings

6.1. Codes Identified in NVivo-supported Thematic Analysis

6.1.1. Codes Identified from Key EU Documents

As mentioned earlier, we manually coded the Ethics Guidelines for Trustworthy AI issued by the EU's High-Level Expert Group in 2019 (HLEG, 2019a), the EU AI Act (Regulation (EU) 2024/1689, 2024). Figure 7 shows the most recurring codes identified manually in NVivo supported thematic analysis of the Ethics Guidelines for Trustworthy AI issued by EU's High-Level Expert Group in 2019 and Figure 8 shows the same for the EU AI Act. in 2024.

The HLEG Guidelines introduced the notion of **trustworthy AI** with strong emphasis falling on its ethical dimension. Reflecting the document's non-binding nature, the language was inclusive and aspirational, articulating values such as human agency, fairness, transparency, and societal well-being as foundations for responsible innovation. This discourse framed ethics as a space for participation, encouraging dialogue among the relevant stakeholders - policymakers, developers, and citizens about how AI should serve collective and human interests.

The AI Act represents an attempt to legally operationalise this ethical vocabulary, but also a shift in emphasis. Many of the same terms appear, such as transparency, accountability, oversight, yet they are contextualised within a risk-based regulatory framework. Comparing figures 4 and 5 the most radical transformation perhaps is with regard to the changing prevalence of **risk and health and safety** in the later documents. Codes that in the HLEG Guidelines invited moral reflection are redefined in procedural terms: transparency becomes an obligation to document system design, fairness becomes a question of data quality, and accountability becomes a matter of conformity assessment. Thus, the AI Act institutionalises the aspiration for trustworthy AI, giving it legal force, but at the same time narrows its ethical scope by emphasising procedural compliance.

However, some codes that featured prominently in the HLEG Guidelines, such as sustainable **development and democracy**, are not significant in the EU AI Act. It could be argued that references to sustainability and environment reflected the broad mandate of the HLEG, which could treat AI ethics as an integrative field, drawing on values from multiple policy spheres. By the time of the AI Act, however, AI governance had been absorbed into a more specialised legislative process, where competence boundaries are tightly defined. Issues such as climate, environment, and sustainable growth fall under separate frameworks like the Green Deal or the Digital Decade strategy.

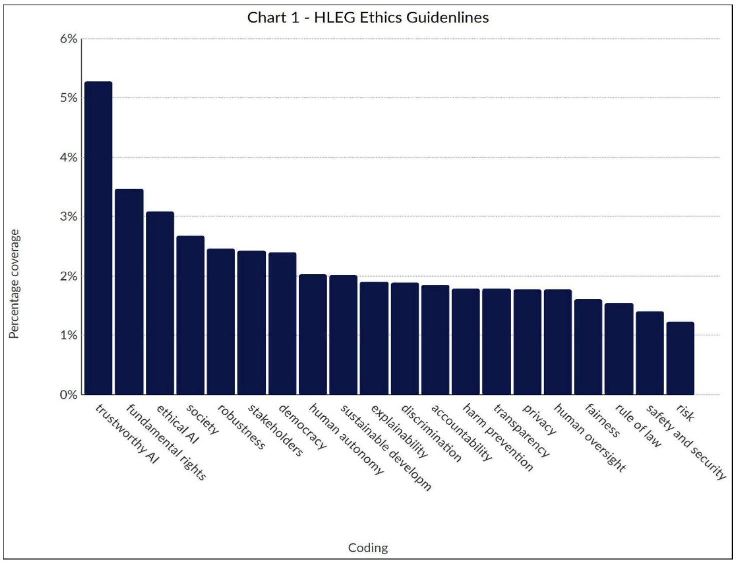


Figure 7. NVivo results (codes) of manual thematic analysis HLEG's Ethics Guidelines for Trustworthy AI

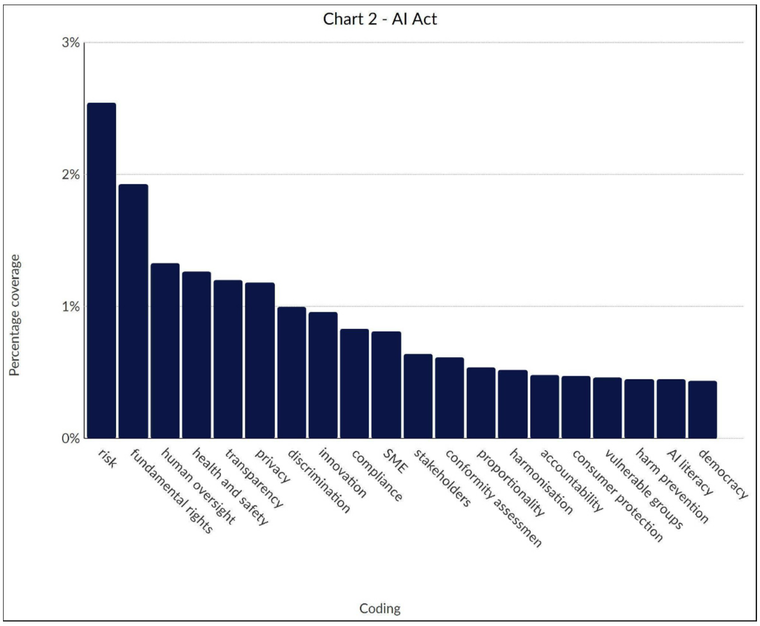


Figure 8. NVivo-supported results (codes) of manual thematic analysis for the EU AI Act

6.1.2. Codes Identified from Supranational Bodies outside of the EU

Table 2 shows an overview of the key codes identified manually through NVivo supported thematic analysis of AI policies issued by other supranational bodies outside of the EU. These include the UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021), the Council of Europe (Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024), the G20 Ministerial Statement on Trade and Digital Economy relative to AI (G20 Trade Ministers and Digital Economy Ministers, 2019), and the OECD Recommendation of the Council on Artificial Intelligence (OECD, 2019).

Table 2 presents an overview of the key codes identified in each document using manual analysis and coding. The overlapping codes that were highlighted as key in all the documents include:

1. Privacy / data protection
2. Sustainable development

3. Stakeholders
4. Environmental protection

The next tier, i.e. the codes that feature prominently in three out of four documents, showing emerging global consensus areas include:

1. Human rights
2. Gender equality
3. AI literacy and education
4. Transparency
5. Society / societal impact

Table 2. Overview of the key codes identified in NVivo-supported manual thematic analysis

ID	Document	No. of codes	Key codes identified through NVivo-supported manual thematic analysis
SuB_09	Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law	30	**Human Rights**, **Democracy**, Rule of Law, Human Oversight, Gender Equality, Risk, Discrimination, Innovation, Privacy, Vulnerable Groups
SuB_11	UNESCO Recommendation on the Ethics of Artificial Intelligence	49	**Human Rights**, **Environmental Protection**, Ethical AI, Life Cycle, AI Literacy, Gender Equality, Stakeholders, Transparency, Privacy, Diversity and Inclusion
SuB_10	OECD Recommendation of the Council on Artificial Intelligence	39	Stakeholders, Privacy, Labour and Employment, **Human Rights**, Risk, State of the Art, Life Cycle, Trustworthy AI, Safety and Security, **Sustainable Development**

SuB_13	G20 Ministerial Statement on Trade and Digital Economy	45	**Sustainable Development**, Innovation, Economy, Diversity and Inclusion, Stakeholders, Benefits, Society, Safety and Security, Employment, Gender Equality

6.2. Topics Identified using BERTopic

In this section, we provide the topics identified from BERTopic, following step 2 of our topic modelling methodology. As a reminder, in this step, BERTopic is applied to a pre-processed text of each document. The output of BERTopic is stored in a tabular format (CSV), and top keywords related to each topic are visualised using a bar chart as well as a word cloud. The use of multiple visualisations enhanced assignment of meaningful semantic topic labels to BERTopic-generated topics. An example bar chart and world cloud illustrating key topics identified from the HLEG's guidelines for trustworthy AI in Figure 9 and Figure 10, respectively. Note that in the word cloud, Topic -1 represents the outlier topic (i.e sentences that are not assigned to any meaningful topic).

Figure 9. Bar charts representing the **topics** identified by BERTopic for the HLEG's guidelines for trustworthy AI

Figure 10. World clouds demonstrating **topics** identified by BERTopic for the HLEG's guidelines for trustworthy AI

Two team members (refer to the positionality statement at the end of this document for the profile of the experts involved) independently analysed the BERTopic outputs and assigned labels to each topic. Despite text pre-processing and hyperparameter optimisation of BERTopic, some of the keywords associated with a topic exhibited insufficient coherence (high noise) to allow meaningful topic labelling. Additionally, while some key words representing a topic were coherent, multiple topic labels could be assigned to them. We adopted a minimalistic approach when assigning topic labels, i.e. attributing the smallest set of labels to each topic identified by BERTopic. In cases where comprehensive representation required multiple labels, we established a constraint of assigning a maximum of 3 labels to each topic. The results of label assignment were then reviewed jointly in a consensus-building session, during which discrepancies were discussed and resolved.

As mentioned earlier, we first provide the results for EU policies (SuB_01 to SuB_08) to examine how the AI policies have evolved within the EU, given its central role in shaping global AI governance. Then, we provide results for the supranational bodies outside the EU (SuB_09 to SuB_13). Finally, we provide the consolidated results with an analysis of overlaps and differences.

6.2.1. Topics Identified from EU Policies

Applying the BERTopic model to each EU policy document produced 227 topics in total. Table 3 shows the number of topics identified by the BERTopic model and the number of labels we assigned to these topics (100 distinct topic labels in total). As noted earlier, the two figures differ because a single topic identified by BERTopic may receive multiple labels (up to three), and some topics were judged to be too noisy to warrant any meaningful label assignment. The complete set of results for all selected AI policies, guidelines, and regulations are publicly archived online at https://doi.org/10.5281/zenodo.18303112[8].

Table 3. Overview of the BERTopic-generated topics and the assigned topic labels associated with each EU document

ID	Document	No. of topics from BERTopic	No. of assigned topic labels	Key topics (from manually assigned labels)
SuB_01	HLEG policy and investment recommendations	26	23	AI literacy, SME, Education, Data governance, Ethical AI, Investment, Innovation
SuB_02	HLEG ethics guidelines	28	27	Ethical AI, Rights, Vulnerable groups, Compliance, Risk, Fairness, Diversity and inclusion, Bias, Safety and security, Oversight, Privacy
SuB_03	HLEG ALTAI	14	16	Human intervention, Bias, Discrimination, Trustworthy AI, Privacy, Data governance, Cybersecurity, Safety and security, AI standardisation
SuB_04	HLEG Sectoral considerations on policy and investment recommendations	8	12	Health Care, Innovation, Risk, Copyright, Sector-specific AI, Trustworthy AI, Investment
SuB_05	EU AI Act	70	47	Regulatory enforcement, Certification, Data protection, Monitoring, SME, Biometrics, Market surveillance, Traceability, Sandbox, AI value chain, General-purpose AI model, Harmonisation of rules, Discrimination, Documentation, Compliance, Risk,

[8] Also available on GitHub at https://github.com/DelaramGlp/forsee_topicmodelling/tree/main/results/Supranational.

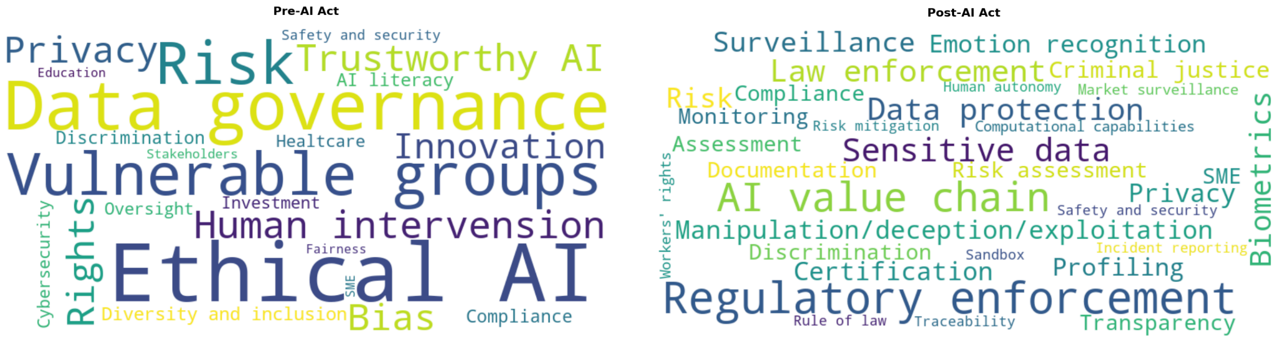
				Cybersecurity, Oversight, Data governance, AI standardisation, Sector-specific AI
SuB_06	AI Act's prohibited AI guideline	62	33	Sensitive data, Law enforcement, Surveillance, Privacy, Manipulation/deception/exploitation, Profiling, Criminal justice, Data protection, Emotion recognition, Biometrics, Transparency, Regulatory enforcement, Rule of law, Human autonomy, Vulnerable groups, National security, Psychological harm, Workers' rights, Social scoring
SuB_07	AI Act's general-purpose AI guideline	7	7	Regulatory enforcement, AI value chain, Computational capabilities, Model training
SuB_08	AI Act's general-purpose AI code of practice	12	12	Risk, Risk assessment, Safety and security, Risk mitigation, AI value chain, Documentation, Compliance, Copyright

To create the aggregated topic overview across the documents, we used the `Count` value associated with each topic, which we labelled in a manual process. Since we applied BERTopic at sentence level, `Count` value represents the number of sentences assigned to each topic. Consequently, `Count` is an indicator of topic occurrence and prevalence within each document. To assess cross-document topic prevalence, counts from the EU documents were aggregated by summing the number of sentence assignments corresponding to equivalent topics across all documents. Figure 11 provides an aggregated topic overview using a word cloud visualisation, wherein relative size of each topic reflects the cumulative number of sentences associated with that topic across the corpus of EU documents. All word cloud visualisations in this section illustrating consolidated results across multiple documents are generated following the same approach.

Figure 11. Word Cloud of topics from all the EU AI documents included in the corpus based on BERTopic results

6.2.2. Evolution of Topics in EU AI Documents According to BERTopic Results

Our analysis shows a shift in the AI policy discourse in the EU following the adoption of the EU AI Act. As shown in Figure 12(a), the pre-AI Act topic landscape predominately focused on ethical AI, data governance, and risks particularly those that impact vulnerable groups. This emphasis aligns with both the non-binding nature of the EU HLEG documents and their mission to promote ethical AI, raise awareness of AI risks, and provide a strategic roadmap, particularly in 2019–2020, when the concept of trustworthy AI is emerging as a popular yet still unfamiliar topic. Topics from the post-AI Act era (Figure 12(b)) illustrate a clear shift from aspirational ethical AI toward operationalised legal AI, with particular emphasis on regulatory enforcement. We contend that this shift offers evidence of a broader "regulatory turn" in policy discourse on ethical, responsible or trustworthy AI practice and a narrowing of considerations to foreground technical and regulatory factors.



(a) Topics pre-AI Act

(b) Topics post-AI Act

Figure 12. Word clouds demonstrating evolution of topics in the EU before and after publication of the AI Act according to BERTopic results

In addition to concrete compliance tasks such as documentation and risk assessment, the AI Act and its associated guidelines distinguish between entities that are involved in development, deployment and use of AI systems, i.e. AI value chain, indicating the need for involvement of multiple entities to achieve trustworthy AI. Post-AI Act topics also revealed the focus on specific AI use cases that impose significant risk. The transition in the AI policy discourse is clearly illustrated in the stream graph and heatmap of top 15 frequency occurring topics across the EU documents in Figures 13 and 14. An interesting observation in the BERTopic-based results is that the topic of risk is more frequent in the HLEG trustworthy AI guidelines than in the AI Act. This is not consistent with the NVivo-based thematic analysis (shown in Figures 7 and 8), which shows a drastic increase in the emphasis on risk in the transition from HLEG guidelines to the EU AI Act. Another observation on risk is related to post-AI Act EU guidelines as the European Commission guidelines on prohibited AI systems and general-purpose do not feature a distinct risk topic. Given that we applied topic modelling approaches, this should not be interpreted as the absence of risk-related considerations, as a simple keyword search shows the word risk is repeated 206 times in the guidelines on prohibited AI and 68 times in general-purpose AI guidelines. This rather reflects a limitation in using BERTopic as an unsupervised topic modelling approach, as will be discussed in section 6.2.5.

The BERTopic-based results show that the emphasis on risk and data protection remains consistent from the pre- to post- legislation period, with the latter reflecting the influence of the EU GDPR. However, certain topics appear to be overlooked in this shift. To have a better understanding of such topics, we analysed the evolution of the bottom 15 (least frequently occurring) topics in the aggregated results across all the examined EU documents, as shown in Figures 15 and 16. Both the stream graph and heap map show a waning of attention to environmental impacts of AI as well as its broader effects on the future of work in the post-AI Act era.

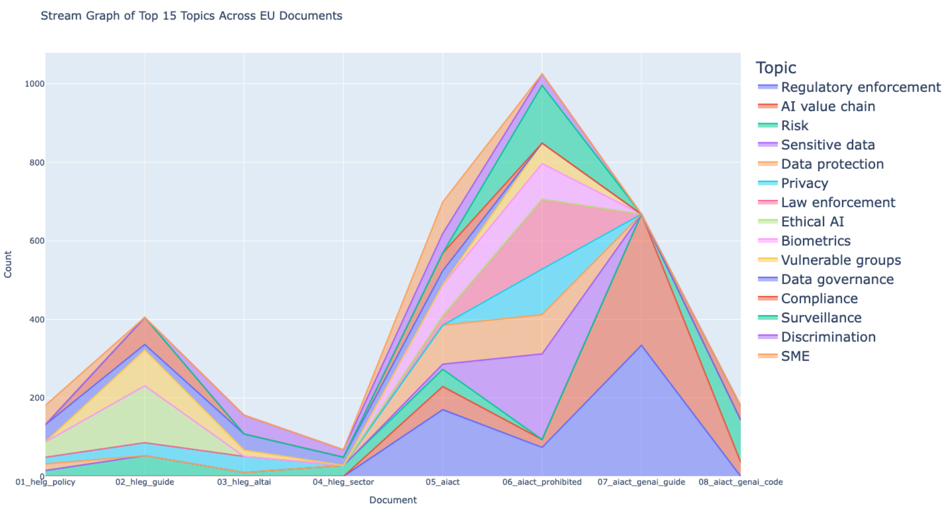


Figure 13. Stream graph showing the distribution of 15 top topics across EU documents

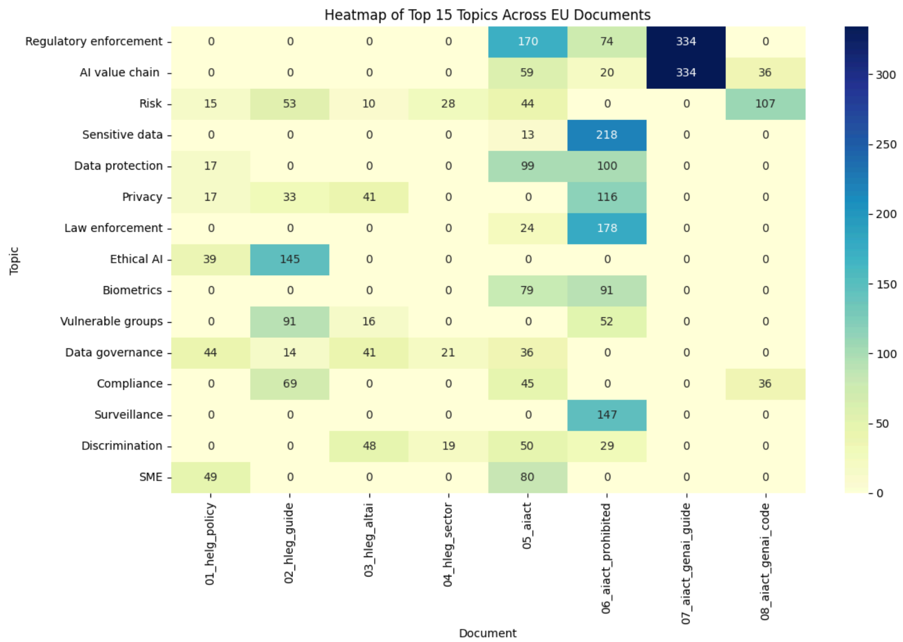


Figure 14. Heatmap showing the distribution of 15 top topics across EU documents

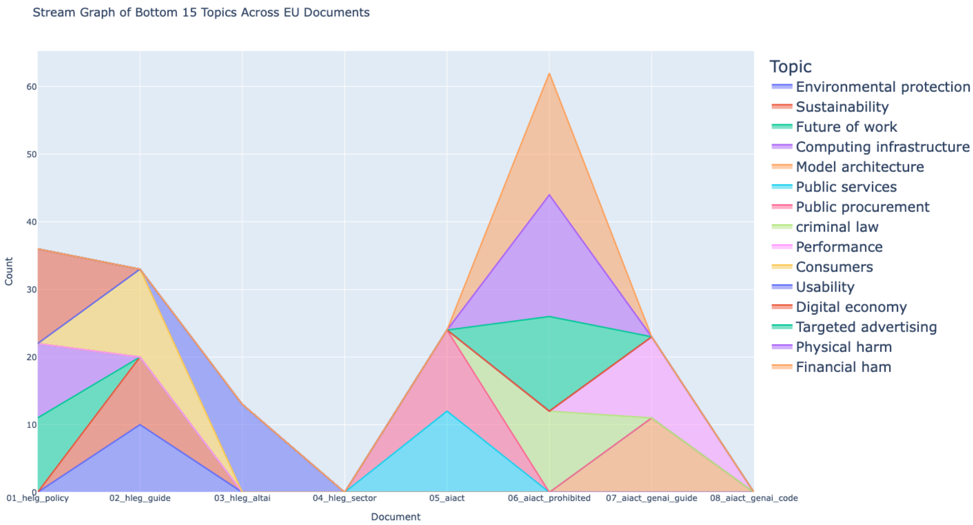


Figure 15. Stream graph showing the distribution of 15 bottom topics across EU documents

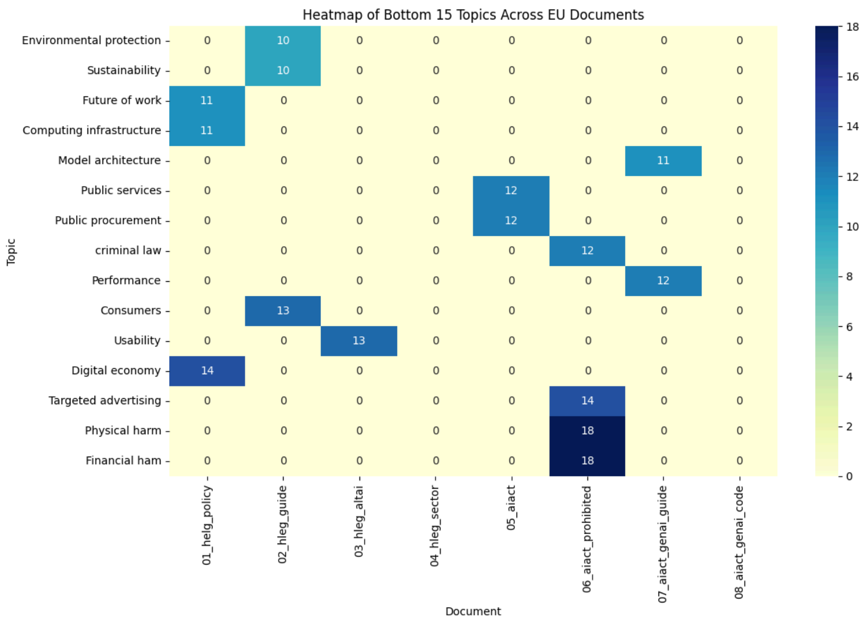


Figure 16. Heatmap showing the distribution of 15 bottom topics across EU documents

6.2.3. Topics Identified from Supranational Bodies Other Than EU

From the assignment of topic labels to BERTopic outputs for supranational bodies outside the EU (SuB_09 to SuB_13), 58 unique topic labels were identified, as illustrated in Table 4.

Table 4. Overview of the topics associated with documents issued by supranational bodies other than EU

ID	Document	No. of topics from BERTopic	No. of assigned topic labels	Key topic labels
SuB_09	Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law	8	4	**Human rights, Democracy**, Risk mitigation, Vulnerable groups
SuB_10	OECD, Recommendation of the Council on Artificial Intelligence	9	8	**AI governance**, Stakeholders, Human rights, Democracy,
SuB_11	UNESCO Recommendation on the Ethics of Artificial Intelligence	18	26	Ethical AI, **Environmental protection**, Sustainability, Trustworthy AI, Vulnerable groups, Data governance, Bias, Fairness, Stakeholders, AI research
SuB_12	United Nations Governing AI for Humanity: Final Report	44	42	**AI governance,** Computing infrastructure, Computing resources, Global collaboration/cooperation, AI standardisation/standards, Data governance, AI research, Diversity and inclusion, Multidisciplinarity, Sustainable development (SDGs), Risk, Environmental protection, Policy dialogue
SuB_13	G20 Ministerial Statement on Trade and Digital Economy	13	14	Trustworthy AI, **AI governance**, Economy, Sustainable development (SDGs), Stakeholders, Data governance

Funded by the European Union

Figure 17 illustrates the word cloud of supranational bodies outside EU with the following 10 top topics:

1. AI governance
2. Computing infrastructure
3. Computing resources
4. Data governance
5. Human rights
6. Global collaboration/cooperation
7. Diversity and inclusion
8. Stakeholders
9. Sustainable development goals (SDGs)
10. AI research

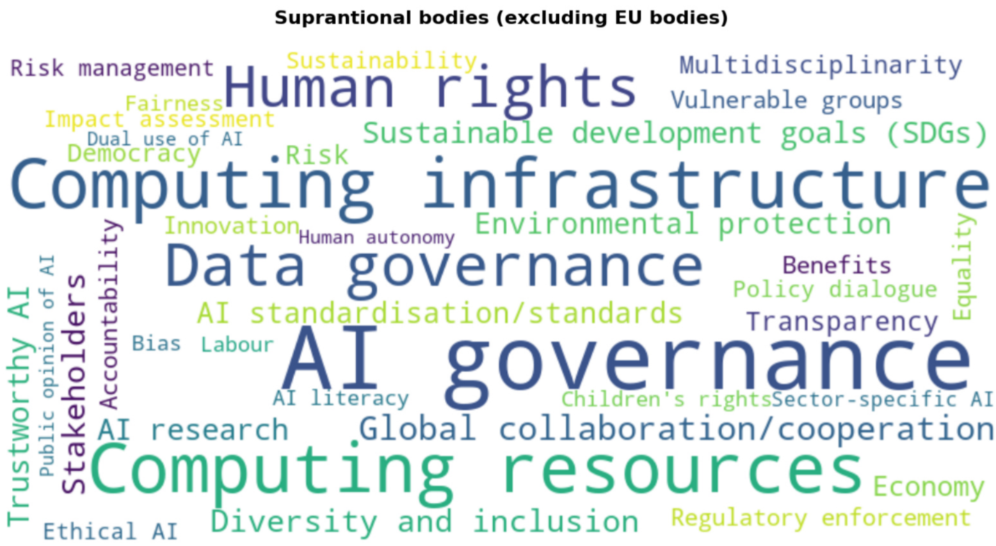


Figure 17. Word cloud of topics identified from supranational bodies other than EU based on BERTopic results

6.2.4. Consolidated Results: Topics for Supranational Bodies

So far, we analysed policies, guidelines and regulations issued by EU and non-EU supranational bodies separately. In this subsection, we consolidate our findings and discuss the overlaps and divergences between these two sets.

We integrated the results from the two sets to gain the overall view of prevalent topics in our corpus of AI documents issued by supranational bodies (presented previously in section 5). The word cloud of the consolidated results is shown in Figure 18. As mentioned earlier, to create the word cloud the frequency of each topic is calculated as the sum of associated `counts` over the entire corpus.

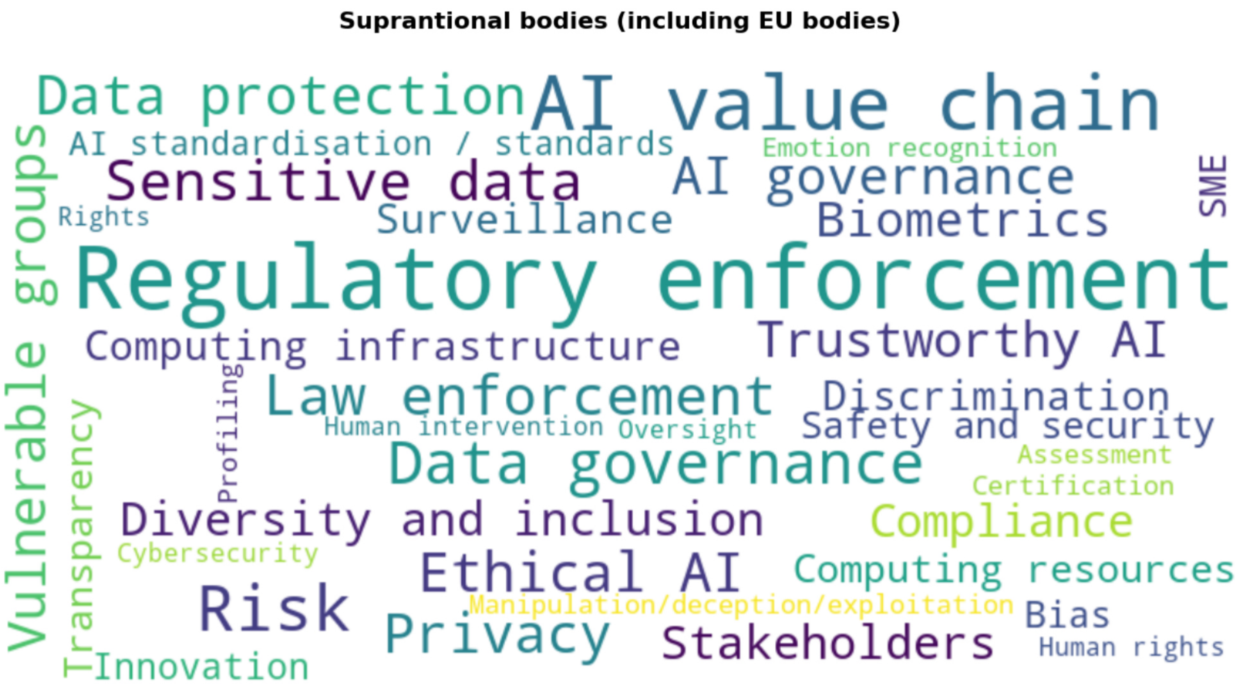


Figure 18. Word cloud of **topics** from all (EU and non-EU) supranational bodies

In terms of overlaps, we identified 26 topics that are prevalent in both sets of documents. The overlapping topics and their frequency are shown in Figure 19. For each topic, the numerical values displayed in the heatmap represent the aggregated frequency of that topic across each document category, based on `counts` calculated by BERTopic.

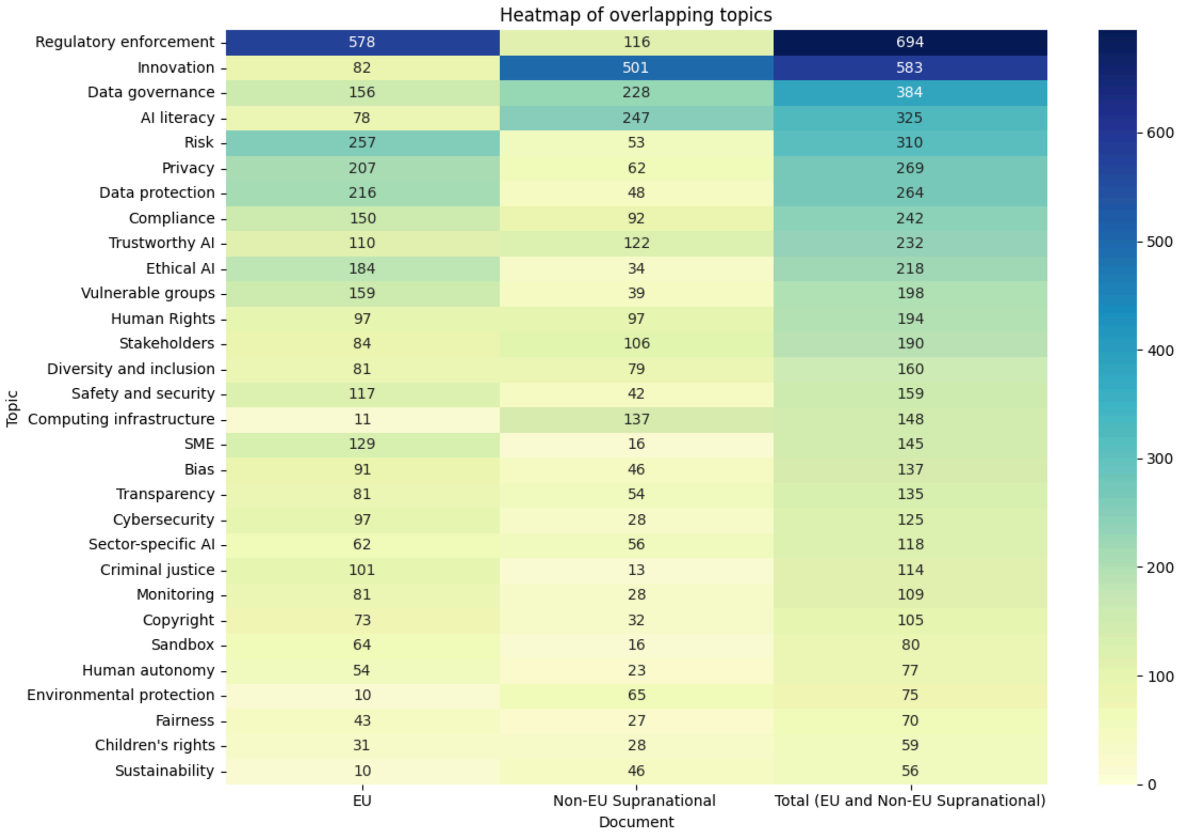


Figure 19. Heatmap of overlapping topics between EU and non-EU supranational bodies

Compared to the topics from the EU documents, analysis of non-EU supranational bodies revealed several distinctive emphases. Similar to the EU policies, other supranational bodies' documents highlight the risks of AI, with explicit mentions of **AI innovation** and **AI literacy. Some of these documents place increasing responsibility on organisations and users to innovate and educate themselves and their users.** There is particular attention to data governance and AI governance and stakeholders, global cooperation and collaboration to address AI challenges and opportunities. Sustained policy dialogue and engagement of a wide range of stakeholders, including the public, and including **inclusive design** of AI. Further, Sustainable development (SDGs), future of the environment, climate, and peace emerge as other notable topics. There are differences between the non-EU documents with the OECD, G20 and UN aligning on many topics, while the Council of Europe and UNESCO diverged, including their emphasis on vulnerable groups which they held in common with

EU documents. The EU and non-EU documents align on stakeholders but AI governance as a term is more common in the non-EU documents while the EU has shifted to the language of compliance and regulatory enforcement.

By referring to both benefits and risks, supranational bodies attempt to balance positive and negative expectations compared to EU policies wherein the focus has shifted in later documents to risks. In addressing the risks, the EU policies put emphasis on technical approaches. The divergences across documents are more clearly evident in Figures 20 and 21.

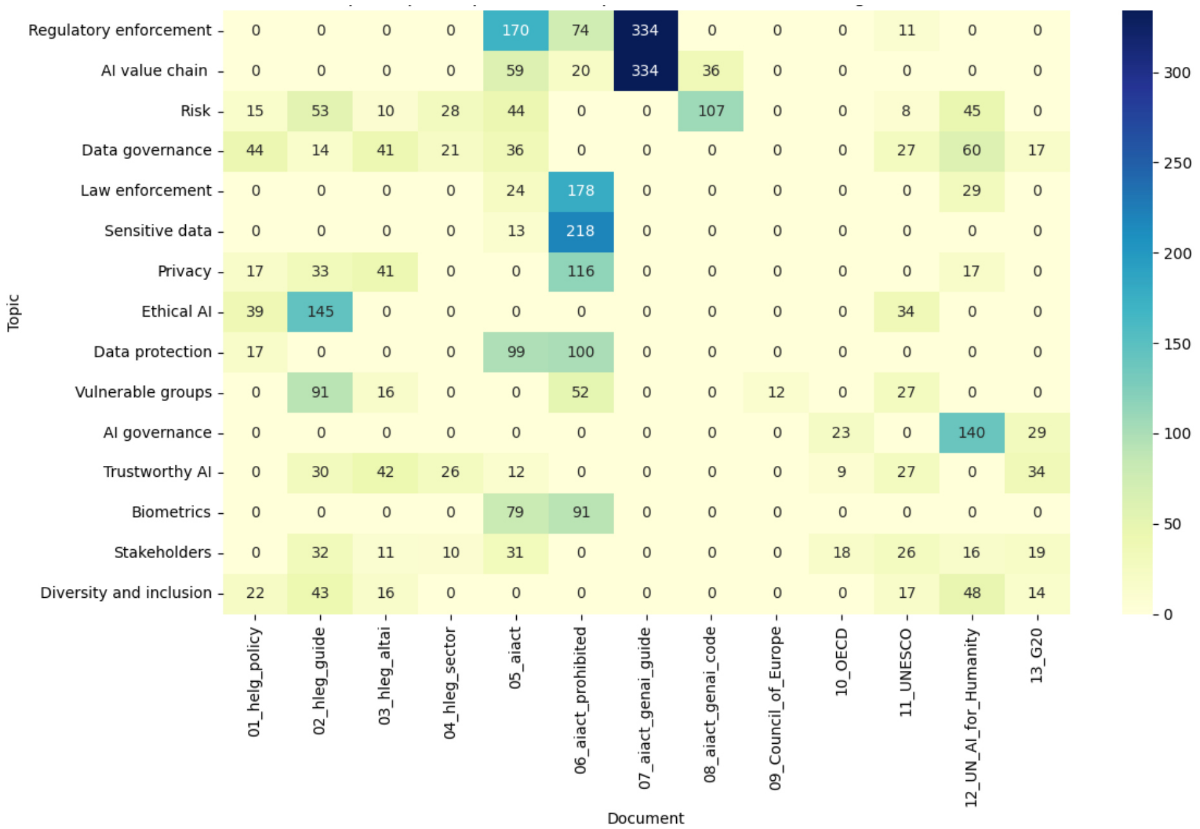


Figure 20. Heatmap of top 15 frequent topics across supranational bodies (including EU and non-EU)

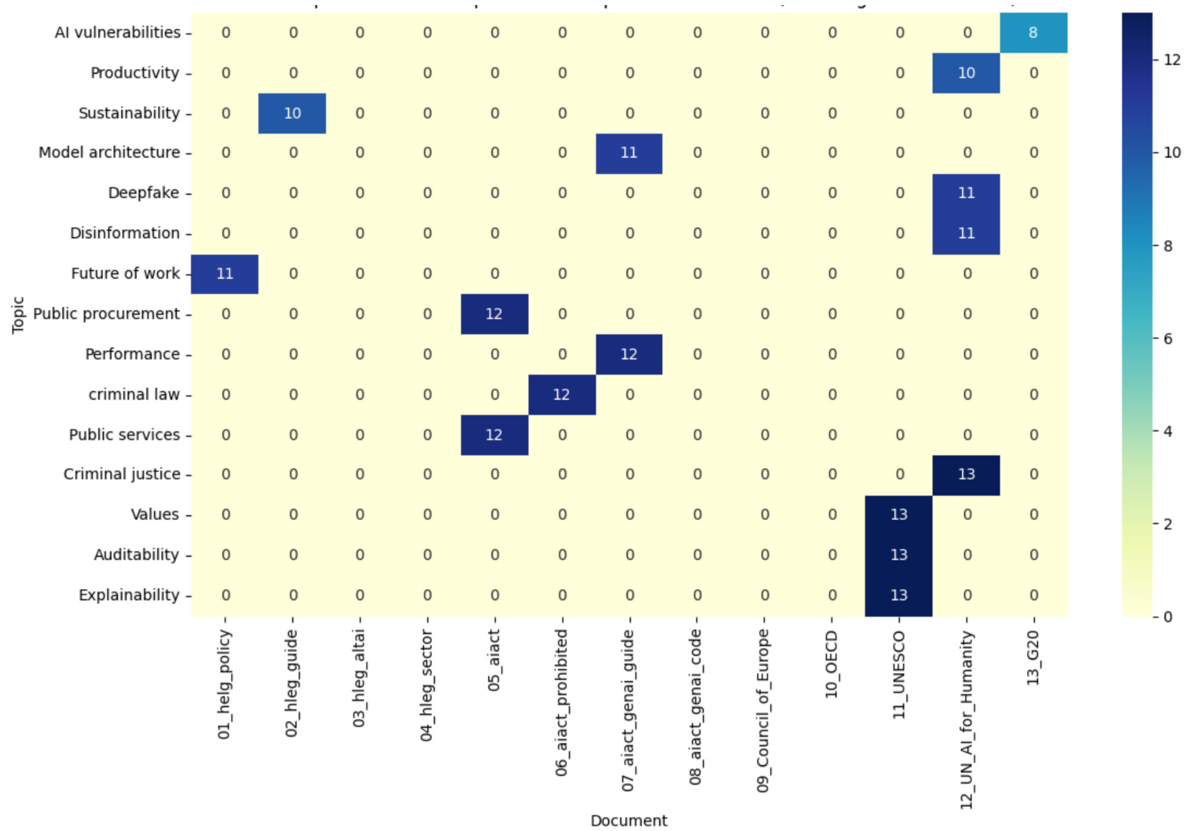


Figure 21. Heatmap of bottom 15 frequent topics across supranational bodies (including EU and non-EU)

6.2.5. Validation of Topics from BERTopic through Comparison with Codes from NVivo-supported Thematic Analysis

To validate the results of topic modelling, we compare the topics with the codes from manual expert analysis encoded in NVivo. This comparison involves identification of overlaps and divergences of the two sets of results for 6 of the documents whose codes were identified manually (as presented in section 6.1). In Appendix B, we provide a detailed comparison of the results of NVivo-supported codes with topics identified from the BERTopic outputs. Since the NVivo-supported thematic and BERTopic analysis were conducted in parallel, some topics and codes shared the same meaning but differed slightly in wording. These minor differences were resolved by treating them as a single concept.

Our comparative analysis shows that:

- For all cases, the number of topics identified from BERTopic are fewer than the coded identified fully manual NVivo-supported thematic analysis. While for longer documents the BERTopic-based results are more comprehensive and accurate, for documents with shorter length they are very limited.
- The order of conducting NVivo-supported coding and labelling of BERTopic results can have influence on expert's judgements, e.g. when the expert first analysed the BERTopic results and engaged consensus building process, the results of manual NVivo-supported coding resulted in more overlapping concepts.
- BERTopic is helpful in gaining insights regarding the documents under investigation, however, it does not replace substantive manual expert-based thematic analysis. Instead, BERTopic supports identification of latent topics that may be missed by experts due to bias, information overload, or even fatigue. BERTopic also enables rapid and scalable analysis of large corpora of documents, which would be labour-intensive if conducted manually.

6.2.6. Limitations of BERTopic-based Topic Identification

Sensitivity to hyperparameter settings: BERTopic results heavily rely on several key hyperparameters, including those determining how the text converts into numerical representations (embedding model), and those controlling the way these representations group into clusters with semantic similarity (parameters such as UMAP and HDBSCAN). In this work, the hyperparameter tuning was conducted at an initial stage using a subset of representative documents, with the objective of yielding a sufficient number of relevant topics from each document while limiting the stochastic variability. As mentioned earlier, two hyperparameter configurations were used in this work: one for longer documents and another for shorter ones which only differ in the number of minimum samples in each cluster of keywords. These hyperparameter configurations were used uniformly (based on the length of document) to ensure consistency and reproducibility of results. However, a caveat in following this approach is that the performance of BERTopic might not be equally optimal for each document, for instance it might have resulted in a higher **proportion of outliers** or **topic omission or merging in one document compared to another**. Further, we applied BERTopic to individual documents by segmenting each into sentences, though BERTopic is typically used to identify latent topics across a large corpora of documents.

This is mainly done to deal with the issue of small-scale data that BERTopic struggles with; and the approach in this case essentially enables BERTopic to overcome sparsity in data issues.

Subjectivity of the topic labelling process: the manual interpretation and labelling of BERTopic results to identify topics can be to some extent subjective (Cisek & Kelleher, 2023). To reduce the element of subjectivity, in this deliverable the manual label assignment was conducted and reviewed by three domain experts from different disciplines (see the positionality statement).

7. Key Themes, Insights, and Discussion

7.1. From Codes and Topics to Key Themes

In this section, we aim to discover the key themes within our corpus of documents issued by supranational bodies through analysis of codes identified in the NVivo-supported thematic analysis and the topics extracted by applying BERTopic. To this end, inspired by knowledge engineering methodologies, we take a bottom-up approach to classify related codes and topics into meaningful groups. Then, through interpretation of the concepts grouped together, a theme indicating AI success criteria is assigned to each category.

By integrating the codes of topics, we identified 157 unique concepts, which are categorised into **4 meta-themes**: **AI technical features, AI Uses**, **AI Risks and Harms** and **AI Governance**. These are further subdivided into 10 constituent themes, as shown in Table 5.

These constituent themes are also broadly categorised against a micro-meso-macro framing. This framing aims to offer a sense of the expectations of supranational bodies as to where the concepts collected in each theme would be best addressed. The **micro** categorisation indicates that the concepts would be best addressed by individual organisations, but with some degree of freedom in how they do so and therefore accommodating of diversity and innovation in approaches taken. The **meso** categorisation indicates that concepts would be best addressed at the level of specific sectors or within a particular government policy area. Similarly this would accommodate different approaches as appropriate between sectors or areas. The **macro** categorisation indicates that concepts would be best addressed at a societal level in a way that spans approaches taken by organisations, sectors and policy areas.

The meta-theme **AI Technical Issues** assembles concepts that are grounded in the characteristics of AI technology. This is broken down into a theme of **AI Technical Features** such as 'computing resources' and 'model training' and of **AI Quality Characteristics** that convey features that supranational bodies expect AI to exhibit, such as 'transparency', 'explainability' and 'robustness'. The expectation for these themes is that they would be best addressed by individual organisations at the *micro level*. For AI Technical Feature theme, this categorisation is selected because it is made up of concepts that are grounded in the technical design choices made by AI developers and deployers that are understood to have a bearing on the economic and societal impact of those systems. For AI Quality Characteristics, the concepts have emerged from multi-stakeholder discourse (with HLEG being one of the earliest) on the types of system qualities that should be measured and controlled for AI systems to be developed and deployed in a responsible manner. While this group of concepts represent a means for advancing discourse and protocols for advancing how these characteristics are defined and monitored between oversight authorities and AI developers and deployers, the expectation is that the latter will be charged with determining and reporting these characteristics.

The meta-theme **AI Uses** assembles concepts that are used to group AI under specific areas of application and policy responsibility. This is broken down into a theme of **Areas AI Uses** which identifies established application areas such as 'healthcare' and 'education' and of **Areas AI Benefits** such as 'digital transformation' and 'digital economy'. Areas of AI Uses represent concepts at the *mesa level* while Areas of AI Benefits are assessed as being best addressed at the *macro level*. The Areas of AI Uses theme corresponds to established definitions of sectors or governmental function, where existing systems of coregulation and other public policy implementation already exist and therefore provide institutional anchor points for implementing AI-related policy initiatives. The Area of AI Benefit theme is categorised at the macro level since they are broader functions of public policy, such as 'investment' and 'economy' that typically require a cross government approach spanning multiple government departments and often strategic interaction with other national governments.

The meta-theme **AI Risks and Harms** assembles concepts related to the identification of possible harms arising from AI and the management of the risk of such harms materialising.

This represents the technocratic approach to governance employed when the impact and risk of harms is not well understood and therefore uncertainty is present at many levels of society on the appropriate actions to take. This is broken down into a theme that addresses concepts related to **AI Risks Addressed by Organisations,** such as 'discrimination' and 'confidentiality'. This is therefore placed in the *micro category.* This is because it contains concepts where there exists some understanding of how risks represented can be assessed and as a result these can be undertaken, with oversight, by organisations with a higher degree of confidence that the risks will be effectively managed. The other theme addresses concepts related to **Shared AI risks/harms**, such as 'democracy' and 'human dignity'. These are concepts where the level of understanding of harms and risks is not yet mature enough to delegate its management to individual organisations (micro-level) with confidence in the outcomes. Further, this set of risk/harm concepts do not yet consistently fall entirely into specific sectors or government policy domains, so this theme spans *the mesa and macro categorisations.*

The **AI Governance** meta-theme invokes concepts that relate to modes, mechanisms and principles that relate to how responsibilities for AI governance are undertaken at different levels. It is subdivided into four themes. Some concepts appear in more than one theme when they relate to activities that entail interaction between actors at different levels, e.g. 'audit' which could be invoked at both the micro and meso level. The *micro level* theme **Organisation-level governance mechanism** includes concepts that are needed to fulfill AI governance responsibilities, such as 'risk management' and 'testing' at an organisational level. The *meso level* theme, **Sectoral/domain oversight mechanisms** assembles concepts such as 'certification', 'standards' and 'incident reporting' that would be employed by bodies related to specific sectors or with specific government mandates in contributing to systems of AI governance. These reflect in particular the governance mechanisms brought in via the AI Act. Concepts deemed most appropriate to the macro level are split into two themes. The macro theme Society-level governance mechanisms collects concepts related to legal rights that are intended to be protected via AI governance, including concepts such as 'intellectual property' and 'workers rights' expressed right. The macro theme, **Society level governance principles** bring together concepts that do not represent legally protected rights, but which have been introduced to help coordinate high level AI governance policy initiatives, e.g. 'trustworthy AI' and 'AI literacy'.

Funded by
the European Union

Table 5. Meta-themes and themes identified from supranational bodies

High level (meta) Themes:	**Micro** themes where expectations are attended to by individual organisations, but they may do so in ways different to other organisations	**Meso** themes where expectations are attended to at a specific sectoral or national policy level	**Macro** themes where expectations are attended to at a societal level, and are expected to span sectoral or policy domains
AI Technical Issues	Theme 1: AI Technical Features		
	Theme 2: AI Quality characteristics		
AI Uses		Theme 6: Areas of AI use	Theme 5: Areas of AI benefit
AI Risks and Harms	Theme 3: Risks addressed by organisation	Theme 7: Shared risks/harms	
AI Governance	Theme 4: Organisation-level governance mechanism	Theme 8: Sectoral/domain oversight mechanisms	Theme 9: Society-level governance mechanisms Theme 10: Society level governance principles

7.2 Applying a Sociology of Expectation Lens to Supranational AI Documents

This document forms one part of a four-part study of how different classes of bodies are addressing challenges of AI Governance, which itself form part of a broader study that encompasses analysis on AI success criteria narratives in traditional and social media, as well as with stakeholders representing SMEs and Civil Society Organisations (CSO). This broader study will be coordinated through the lens of Sociology of Expectations which

Funded by the European Union

allows the narratives on AI observed in these different arenas to be analysed together to better understand how these different stakeholder types are responding to and indeed attempting to shape dominant narratives of the inevitable and transformative nature of AI technology.

The analyses of supranational body documents addressed in this document represents the outcome of formal processes and a class of macro expectations. They are expressions of agreements formed between governments in response to the public, and some industry expectations, that governments and a range of public and private organisations can exert some degree of governance over the potential harms of AI technology as it advances. Implicit in our selected focus on supranational documents is the recognition that the challenges of AI governance across the value chain are complex and poorly understood, and success in mitigating harms is therefore very uncertain. As a result, most European governments seek solutions through supranational measures rather than national measures. The supranational AI governance initiatives studied in this document, therefore, explicitly aim to represent a wide range of stakeholder concerns. They also place the more technical discussions about AI in a wider social and political context.

In initiatives such as the HLEG Guidelines on trustworthy AI (HLEG, 2019a), this involved selecting and assembling a range of industry, academic and civil society stakeholders and attempting to reach consensus on a shared output. This process was not without conflict (including resignations), and the final outputs were questioned for their narrow framing of the issues, particularly their failure to address democratic and fundamental rights issues (Pierson et al., 2023), and the inadequacy of their policy recommendations (Veale, 2020). The other initiatives are results of consensus between representatives at the governmental level. In some cases, such as for the OECD Recommendation (OECD, 2019), consensus is formed again between selected experts. In others, governmental representation includes democratic representatives, e.g. European Parliament in the development of the AI Act (Regulation (EU) 2024/1689, 2024), and integrates stages of open consultations potentially open (albeit accessed with different degrees of resources) to industry, civil society and even individual citizens. The expectations established by these supranational initiatives, therefore, represent the results of different forms of multistakeholder representation, consultation and deliberation to reach a form of consensus in each case. This is important as all these bodies represent different constituencies and while there are some overlaps in their membership, the documents bring different values and emphases to the fore. They are each trying to shape public expectations of AI and manage the trajectory of future AI innovation.

7.3 Supranational Approaches to AI Governance and Potential Criteria for Success

Our analysis includes documents developed by the EU and those by other supranational bodies, specifically the UN, UNESCO, the OECD, the Council of Europe and the G20 from 2018-2025. While these documents reveal some differences in emphasis, the combined corpus shows a clear positive expectation from these bodies that the benefits of **AI innovation**, in terms of economic benefits and in accelerating solutions to societal problems, can be realised while also putting in place protections against a variety of risks and potential harms. There is also a shared determination that **trustworthy AI** is possible and that multi-stakeholders should be involved in AI governance. Where the institutions and their documents differ is on the **mode of governance** (self or co-regulation) and main criteria that should underpin AI governance.

When the documents in our corpus are analysed together there is an emphasis on the importance of computing infrastructure, data governance, AI literacy, multi-stakeholder collaboration and regulatory enforcement (fig. 12). **Regulatory enforcement** is particularly emphasised in the EU AI Act and related documents, and when we remove the later EU documents from our corpus the term **AI governance** more broadly is preferred, especially by the OECD and G20. The latter place a greater emphasis on AI literacy. Similarly, accountability, a core principle for the EU's high level expert group on AI in 2018, is less prevalent in later EU documents including the AI Act and related documents replaced by regulatory enforcement, AI value chains and compliance. This signals different approaches to the mode of governance, who governs AI and where responsibility lies.

The EU documents, and in particular the AI Act, emphasises the expectation that protections can be implemented through regulatory compliance systems based on a risk management framework, inspired by existing health and safety protection for products, but extended to include fundamental/human rights protections. This impact of this "innovation-with-protections" expectation for AI governance is also evident in the shift we observe in themes identified in documents before and after the agreement of the AI Act, where self-regulation and themes associated with ethical AI or ethical principles, are replaced by themes associated with technical aspects of AI (e.g. data protection, privacy, biometrics) and operationalised regulation, e.g. around regulatory enforcement and minimising risk. The legacy of prior EU legislation related to data is strong, with data protection, privacy and vulnerable groups still to the fore in the EU AI regulatory documents as well as a new focus on safety.

An important divergence in our document sample relates to the emphasis on sustainable development, democracy and human rights. More recent EU AI policy documents pay much less attention to broader socio-ecological impacts such as environmental and social sustainability compared to the other supranational documents. Human rights are to the fore in the Council of Europe and UNESCO documents, and still evident in the OECD recommendations, while fundamental rights, and children's rights are more to the fore in the EU AI Act.

Overall, we see a convergence on a narrow range of micro and internal technical criteria highlighted across this sample of documents and some variance in the meso and macro level regulatory, and governance approaches with the EU emphasising risks and rights. We also see a variance in the emphasis on the environment, sustainability and democracy.

These findings would suggest that **AI success criteria will vary from country to country** as different contexts and their populations place greater or lesser degrees of importance on different factors or are more sensitive to the impact of AI on certain aspects of their lives (e.g. privacy). Further, the analysis shows that the European approach to governing AI and their priorities have been changing over the period under analysis and we would suggest the European emphasis on risk and standards requires further analysis. Will the shift from ethics to risk be taken up by other supranational bodies? Who will determine what is risky and in what context, and who will ensure regulatory enforcement and compliance? Finally, who exactly are the 'stakeholders' who will be involved in AI governance. The answers to these questions may yet determine if the European approach to AI governance provides the necessary assurance for those considering investing time and resources in AI innovation and satisfies societal expectations that the benefits of AI innovation outweigh its risks.

8. Conclusion

This deliverable (D2.2) examined how success in AI articulated by key supranational bodies through an analysis of 13 major AI policies, guidelines, and regulations published between 2018 to 2025. This examined corpus comprised documents issued by EU bodies (8), OECD/G20 (2) and UNESCO, UN and Council of Europe (3).

Utilising a mix-method approach that combined thematic analysis and topic modelling, we identified prevalent codes and topics within documents and further interpreted them to define "successful AI" through the lens of sociology of expectations. Across the corpus, four meta-themes emerged: AI Technical Issues, AI Uses, AI Risks and Harms and AI Governance. Within these, 10 distinct themes were identified.

In terms of limitations, the selection criteria for the corpus excludes relevant national-level policies and additional international instruments that do not have an explicit focus on AI, such as the GDPR and DSA. Second, topic modelling was conducted at the level of individual documents, which may limit the capture of cross-document topics. Finally, while the combination of thematic analysis and topic modelling strengthened the analysis through inclusion of both qualitative and quantitative methods, further methodological refinement and triangulation could enhance the outcomes.

This deliverable (D2.2) provides the conceptual and methodological framing for the other WP2 deliverables, namely D2.1 (technical standards bodies), D2.3 (ICT professional bodies), and D2.4. (non-technology professional bodies). Methodologically, it establishes the use of BERTopic as an appropriate topic modelling approach. Substantively, the meta-themes and themes identified in D2.2 inform the subsequent document analysis in the other deliverables by providing a shared baseline for theme identification, enabling systematic comparison of similarities and differences in formal approaches to AI success across different institutional perspectives.

In the remainder of the broader study undertaken by the FORSEE project, we aim to explore how the EU's "innovation-with-protections" expectation serves: to raise attention and legitimise investments more broadly, e.g. in media discourse and attitudes of stakeholder classes such as professional bodies, SMEs and CSOs; in providing direction to the search processes of science and technology, e.g. in stakeholder expectations for further investments to realise this narrative; and in its coordination effect, both in the related coordination expectations of stakeholders and the forms of coordination observed in related consensus-based outputs from standards bodies and different professional bodies. Further, we will compare the set of themes identified in this document with the themes identified in document sets selected from other types of bodies.

Positionality Statement

The experts directly involved with identification of codes, topics, and themes are:

1. Expert #1: Holds a PhD in EU law and has 3 years of postdoc experience,
2. Expert #2: Holds a PhD in Computer Science, has experience in AI risk modelling, EU AI policies and standards, with 1 year of postdoc experience.
3. Expert #3: Holds a PhD in Communication Studies, a Full Professor of Digital Media & Communication with 25 years experience.
4. Expert #4: Holds a PhD in Computer Science, a Full Professor of Computer Science, with 35 years of research experience

- The thematic analysis to identify codes was conducted by Expert #1.
- The analysis of the BERTopic outputs to label topics were conducted by Expert #1 and Expert #2 and was reviewed by Expert #3.
- Analysis of codes and topics through the lens of sociology of expectation to identify themes were conducted by Experts #3, #4, and #2.

Acknowledgements

This work has received funding from the European Commission's Horizon Europe Research and Innovation Programme under grant agreement No. 101177579 (FORSEE) and from Research Ireland at ADAPT, the Research Ireland Centre for AI-Driven Digital Content Technology at TCD and UCD #13/RC/2106_P2. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.
We would like to thank Victoria Wiegand and Yuening Li for their contributions to this deliverable. We are grateful to Prof. John D. Kelleher and Dr. Vasudevan Nedumpozhimana for providing valuable advice on topic modelling. We would like to extend our gratitude to Prof. Esther Keymolen and Prof. John D. Kelleher for reviewing this manuscript.

References

Bareis, J., & Katzenbach, C. (2022). Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values*, *47*(5), 855–881. https://doi.org/10.1177/01622439211030007

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024). *AI auditing: The Broken Bus on the Road to AI Accountability* (No. arXiv:2401.14462). arXiv. https://doi.org/10.48550/arXiv.2401.14462

Braun, V., & Clarke, V. (2021). *Thematic Analysis: A Practical Guide*. SAGE.

Brown, N. (2015). Metrics of hope: Disciplining affect in oncology. *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine*, *19*(2), 119–136. https://doi.org/10.1177/1363459314555239

Brown, N., Rip, A., & Van Lente, H. (2003). Expectations in & about science and technology. *A Background Paper for the 'Expectations' Workshop of June*, *177*, 371–380. https://www.researchgate.net/profile/Harro-Van-Lente/publication/46671758_Expectations_in_About_Science_and_Technology/links/0c960527b59d75b7b6000000/Expectations-in-About-Science-and-Technology.pdf

Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10s), 1–35. https://doi.org/10.1145/3507900

Cisek, K., & Kelleher, J. D. (2023). Current Topics in Technology-Enabled Stroke Rehabilitation and Reintegration: A Scoping Review and Content Analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 3341–3352. https://doi.org/10.1109/TNSRE.2023.3304758

Corrêa, N. K., Galvão, C., Santos, J. W., Pino, C. D., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & Oliveira, N. de. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, *4*(10). https://doi.org/10.1016/j.patter.2023.100857

Dotson, T. (2015). Technological Determinism and Permissionless Innovation as Technocratic Governing Mentalities: Psychocultural Barriers to the Democratization of Technology. *Engaging Science, Technology, and Society*, *1*, 98–120.

European Commission. (2025a). *Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689* (No. C(2025) 5052 final). https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act

European Commission. (2025b). *Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act | Shaping Europe's digital future* (No. C(2025) 5045 final). https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act

European Commission. (2025c). *The General-Purpose AI Code of Practice | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

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 Research Publication, 2020–1*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482

Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (2024). https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

G20 Trade Ministers and Digital Economy Ministers. (2019). *'G20 Ministerial Statement on Trade and Digital Economy.'* https://www.mofa.go.jp/files/000486596.pdf

Galanos, V. (2023). *Expectations and expertise in artificial intelligence: Specialist views and historical perspectives on conceptualisation, promise, and funding*. https://doi.org/10.7488/era/3188

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

HLEG. (2019a). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

HLEG. (2019b). *Policy and Investment Recommendations for Trustworthy AI*.

HLEG. (2019c). *Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence*. Publications Office of the

European Union. https://data.europa.eu/doi/10.2759/733662

Jasanoff, S., & Kim, S.-H. (2009). Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea. *Minerva*, *47*(2), 119–146. https://doi.org/10.1007/s11024-009-9124-4

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kajava, K., Öhman, E., Takagi, N. M., Nakajima-Wickham, E., & Vitiugin, F. (2025). Post-GPT Policy: Risk and Regulation in EU AI Discourse. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*, 994–1006. https://ojs.aaai.org/index.php/ICWSM/article/view/35856

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society*, *7*(1), 2053951720915939. https://doi.org/10.1177/2053951720915939

Li, Z., Calvo-Bartolomé, L., Hoyle, A., Xu, P., Dima, A., Fung, J. F., & Boyd-Graber, J. (2025). *Large Language Models Struggle to Describe the Haystack without Human Help: Human-in-the-loop Evaluation of Topic Models* (No. arXiv:2502.14748). arXiv. https://doi.org/10.48550/arXiv.2502.14748

Lindgren, S., & Dignum, V. (2023). *Chapter 14: Beyond AI solutionism: toward a multi-disciplinary approach to artificial intelligence in society*. https://www.elgaronline.com/edcollchap/book/9781803928562/book-part-9781803928562-19.xml

Mansell, R. (2012). *Imagining the Internet: Communication, innovation, and governance*. Oxford University Press.

Micheli, M., Hupont, I., Delipetrev, B., & Soler-Garrido, J. (2023). The landscape of data and AI documentation approaches in the European policy context. *Ethics and Information Technology*, *25*(4), 56. https://doi.org/10.1007/s10676-023-09725-7

OECD. (2019). *Recommendation of the Council on Artificial Intelligence* (No. OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Palladino, N. (2023). A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices. *Telecommunications Policy*, *47*(5), 102479.

https://doi.org/10.1016/j.telpol.2022.102479

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., & Iyyer, M. (2024). *TopicGPT: A Prompt-based Topic Modeling Framework* (No. arXiv:2311.01449). arXiv. https://doi.org/10.48550/arXiv.2311.01449

Pierson, J., Kerr, A., Robinson, S. C., Fanni, R., Steinkogler, V. E., Milan, S., & Zampedri, G. (2023). Governing artificial intelligence in the media and communications sector. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1683

Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, *14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Puppis, M., Mansell, R., & Van den Bulck, H. (2024). *Handbook of media and communication governance*. Edward Elgar Publishing. https://books.google.com/books?hl=en&lr=&id=M2gTEQAAQBAJ&oi=fnd&pg=PR1&dq=Handbook+of+Media+and+Communication+Governance&ots=k1q9NabbDG&sig=vZWwIekcyY7TIKc4zIiLoWNzG4k

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), 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) (Text with EEA Relevance), 277 OJ L (2022). http://data.europa.eu/eli/reg/2022/2065/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) (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance and Amending Regulation (EU) 2018/1724 (Data Governance Act) (Text with EEA Relevance), 152 OJ L (2022).

http://data.europa.eu/eli/reg/2022/868/oj

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., … Trager, R. (2025). *Open Problems in Technical AI Governance* (No. arXiv:2407.14981). arXiv. https://doi.org/10.48550/arXiv.2407.14981

Roche, C., Wall, P. J., & Lewis, D. (2023). Ethics and diversity in artificial intelligence policies, strategies and initiatives. *AI and Ethics*, *3*(4), 1095–1115. https://doi.org/10.1007/s43681-022-00218-9

Saheb, T., & Saheb, T. (2023). Topical review of artificial intelligence national policies: A mixed method analysis. *Technology in Society*, *74*, 102316.

Saheb, T., & Saheb, T. (2024). Mapping Ethical Artificial Intelligence Policy Landscape: A Mixed Method Analysis. *Science and Engineering Ethics*, *30*(2), 9. https://doi.org/10.1007/s11948-024-00472-6

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, *2*(1), 31–42.

Squires, V. (2023). Thematic Analysis. In J. M. Okoko, S. Tunison, & K. D. Walker (Eds.), *Varieties of Qualitative Research Methods* (pp. 463–468). Springer International Publishing. https://doi.org/10.1007/978-3-031-04394-9_72

Steinhoff, J. (2024). AI ethics as subordinated innovation network. *AI & SOCIETY*, *39*(4), 1995–2007. https://doi.org/10.1007/s00146-023-01658-5

Suter, V., Ma, C., Pöhlmann, G., & Meckel, M. (2025). When Politicians Talk AI: Issue-Frames in Parliamentary Debates Before and After ChatGPT. *Policy & Internet*, *17*(3), e70010. https://doi.org/10.1002/poi3.70010

Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic analysis. *The SAGE Handbook of Qualitative Research in Psychology*, *2*(17–37), 25.

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. (2020). Publications Office of the European Union. https://data.europa.eu/doi/10.2759/002360

Toney, A., Curlee, K., & Probasco, E. (2024). Trust Issues: Discrepancies in Trustworthy AI Keywords Use in Policy and Research. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2222–2233.

https://doi.org/10.1145/3630106.3659035

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence* (No. SHS/BIO/PI/2021/1). https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

United Nations. (2024). *Governing AI for humanity: Final report*. United Nations.

van Lente, H. (2012). Navigating foresight in a sea of expectations: Lessons from the sociology of expectations. *Technology Analysis & Strategic Management*, *24*(8), 769–782. https://doi.org/10.1080/09537325.2012.715478

van Lente, H., Spitters, C., & Peine, A. (2013). Comparing technological hype cycles: Towards a theory. *Technological Forecasting and Social Change*, *80*(8), 1615–1628. https://doi.org/10.1016/j.techfore.2012.12.004

Veale, M. (2020). A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence. *European Journal of Risk Regulation*, *11*(1), e1. https://doi.org/10.1017/err.2019.65

Vicsek, L. (2020). Artificial intelligence and the future of work – lessons from the sociology of expectations. *International Journal of Sociology and Social Policy*, *41*(7–8), 842–861. https://doi.org/10.1108/IJSSP-05-2020-0174

Wang, J., Huo, Y., Mahe, J., Ge, Z., Liu, Z., Wang, W., & Zhang, L. (2024). Developing an Ethical Regulatory Framework for Artificial Intelligence: Integrating Systematic Review, Thematic Analysis, and Multidisciplinary Theories. *IEEE Access*, *12*, 179383–179395. https://doi.org/10.1109/ACCESS.2024.3501332 A: Literature related to Thematic Reviews of AI Policies and Guidelines

Appendix A: Literature Related to Thematic Reviews of AI Policies and Guidelines

Table A.1: List of existing thematic review as baseline

Paper	Coding Method	Sample Size	Themes
(Corrêa et al., 2023)	Based on a list of principles from previous literature that was subsequently both refined and expanded	200	1) Accountability/liability, 2) Beneficence/ non-maleficence, 3) Children and adolescents rights, 4) Dignity and human rights, 5) Diversity/ inclusion/ pluralism/ accessibility, 6) Freedom/autonomy/ democratic values/technological sovereignty, 7) Human formation/education, 8) Human-centeredness/ alignment, 9) Intellectual property, 10) Justice/ equity/ fairness/ non-discrimination, 11) Labour rights, 12) Cooperation/ fair competition/open source, 13) Privacy, 14) Reliability/ safety/ security/ trustworthiness, 15) Sustainability, 16) Transparency/ explainability/ auditability, 17) Truthfulness
(Micheli et al., 2023)	Not specified how themes were identified	41 out of 2237 identified for initial screening	(1) ethics, (2) quality, (3) reproducibility, (4) discoverability, (5) accountability and (6) trust
(Jobin et al., 2019)	Not specified how themes were identified, but they are listed by	84	1) transparency, 2) justice and fairness, 3) non-maleficence, 4) responsibility, 5) privacy, 6) beneficence, 7) freedom and

	frequency of mentions		autonomy, 8) trust, 9) dignity, 10) sustainability, and 11) solidarity
(Hagendorff, 2020)	Not specified how themes were identified, but they are listed by frequency of mentions	22	1) privacy protection, 2) fairness, non-discrimination, justice, 3) accountability, 4) transparency, 5) openness, 6) safety, cybersecurity, 7) common good, sustainability, well-being, 8) human oversight, control, auditing, 9) solidarity, inclusion, social cohesion, 10) explainability, interpretability, 11) science-policy link, 12) legislative framework, legal status of AI systems, 13) future of employment/worker rights, 14) responsible/intensified research funding, 15) public awareness, education about AI and its risks, 16) dual-use problem, military, AI arms race field-specific deliberations (health, military, mobility etc.), 17) human autonomy, 18) diversity in the field of AI, 19) certification for AI products, protection of whistleblowers, 20) cultural differences in the ethically aligned design of AI systems, 21) hidden costs (labelling, clickwork, content moderation, energy, resources).
(Fjeld et al., 2020)	Collated the list of principles, merging close equivalents, to form a final list of 47 principles. Then, clustered the principles, identifying ones that were closely related	"Approximately 20"	8 themes with a number of principles under each. 1) Privacy Principles: Consent, Control over the use of data, Ability to restrict processing, Right to rectification, Right to erasure, Privacy by design,

	both in terms of their dictionary meanings (e.g. fairness and non-discrimination) as well as ones that were closely linked in the principles documents themselves (e.g. transparency and explainability).		Recommends data protection laws, Privacy 2) Accountability Principles: Verifiability and Replicability, Environmental Responsibility, Evaluation and Auditing Requirements, Remedy for automated decision, Liability and legal responsibility, Accountability per se 3) Safety and security Principles: Safety, Security, Security by Design, Predictability 4) Transparency and explainability Principles: Transparency, Explainability, Open source data and algorithms, Right to information, Notification when AI makes a decision about an individual, Notification when interacting with AI, Regular reporting 5) Fairness and non-discrimination Principles: Non-discrimination and bias prevention, Representative and high-quality data, Fairness, Equality, Inclusiveness in impact, inclusiveness in design 6) Human control of technology Principles: Human review of automated decisions, Ability to opt out of automated decisions Human control of technology, Human control of technology 7) Professional responsibility Principles: Accuracy, Responsible design, Consideration of long-term

			effects, Multi-stakeholder collaboration, Scientific Integrity 8) Promotion of human values Principles: Human Values and human flourishing, Access to technology, Leveraged to benefit society
(Reuel et al., 2025)	No specific methodology, seems a large group assembly of problems according to a taxonomy of technical approaches to AI governance. Explicitly excludes works on performance, safety, robustness and ethics.	500+	Reviews literature under taxonomy of capacities: Assessment, Access, Verification, Security, Operationalisation, Ecosystem monitoring, ranges over a classification of targets: Data, Compute, Models & Algorithms; and Deployment. Under each capacity presents a set of open research questions [still need to review for relevance]. Also offers cross reference to topics by classification of technical topics, e.g. ML theory, software engineering.
(Roche et al., 2023)	Voyant Tools, a web-based reading and analysis environment for web-based texts, was used to calculate a numerical value to indicate the frequency per million of certain concepts within a document	463	Most frequently used terms: Data, Intelligence, Artificial, Use, Systems, Human, Public, Digital, Research, Technology, New, Development, Information, Rights, European, Technologies, Learning, Based, Law, Government, Used, Protection, Work, Services, Social.
(Birhane et al., 2024)	Works selected on terms related to accounting and auditing of AI: split between systematic review of 341 academic papers	341 academic+18 non academic	Analysed works were categorised by product/ model/algorithm audits, data audits, socio technical ecosystems authors and meta commentaries. Discussion focused on how impactful non-academic audits could inform academic

	and audits from 3 consulting agencies, 2 corporate, 2 journalistic, 3 law firms, 2 regulators and 6 civil society. Analysis addresses context (motivation, goals, target, harms, institutional framing), methodology and impact		research and policy with discussion points being: power relationships and auditor-stakeholder relationships; addressing audit context and reporting beyond evaluation; expand audit scope to stakeholder and AI systems interactions; specificity increases impact; widening range of audit methodologies used; appreciating diversity of audits practitioners; factors of audit goal, target, design, communication and scope outweigh auditor type and timing; recognizing limits of audits (e.g. against chilling effects and power concentration).

(Palladino, 2023)	Manual thematic analysis using NVivo	55 documents	Identified 33 elements (principles and requirements):
			* **Values**
			* **Principles**
			* **Requirements**
			
			* **Transparency**
			    * Documentation
			    * Notification
			    * Traceability
			    * Reproducibility
			    * Explainability
			* **Fairness**
			    * Bias Prevention
			    * Data Representativeness
			    * Inclusive Benefit Distribution
			    * Inclusiveness in design
			* **Accountability**
			    * Verification and Validation
			    * Assessments
			    * Auditability

			* Appealability and Remediability * Liability and Legal Responsibility * **Privacy** * Consent * Data Minimization * Data Agency/Control * Anonymization * **Reliability** * Safety * Security * Resilience * Predictability * **Human control** * Human Oversight * Human Review * Opt-outs
(Toney et al., 2024)	Manual annotation using the text analysis platform Dedoose	322.209	Initial set of 19 principles: 1) Accessibility, 2) Accountability, 3) AI Lifecycle, 4) AI System Design, 5) Discrimination, 6) Explainability, 7) Fairness, 8) Human Rights, 9) Inclusivity, 10) Law, 11) Outcomes, 12) People / Individuals, 13) Privacy, 14) Procedural Fairness, 15) Security, 16) Statistics, 17) Transparency, 18) Unjust / Unlawful, 19) Users. These were further narrowed down to key 6 principles: 1) accountability, 2) explainability, 3) fairness, 4) privacy, 5) security, 6) transparency.

(Wang et al., 2024)	Manual using NVivo	51 sources including literature on ethical frameworks, AI regulations, policies, and standards	23 themes and 19 subthemes, and categorised them into 3 categories of role, procedure, and strategy

Appendix B: Comparison of NVivo-Supported Codes and BERTopic-based Topics

Table B.1. Comparison of codes and topics for HLEG guidelines for trustworthy AI

SuB_02. HLEG Ethics Guidelines for Trustworthy AI		
No. of topics vs. codes	27 (BERTopic - topics) < 51 (NVivo - codes)	
Overlaps	16 concepts: • Trustworthy AI • Stakeholders • Ethical AI • Privacy • Fairness • Safety and security • Environmental protection • Risk • Diversity and inclusion • Human rights • Bias • Vulnerable groups • Human dignity • Reliability • Accuracy • Compliance	
Divergences	**Unique to BERTopic**	**Unique to NVivo**

	• Sustainability • Health Care • Employment • Consumers • Data governance • Audit • Reproducibility • Human intervention • Harm • Assessment • Oversight	• Fundamental rights • Society • Democracy • Robustness • Sustainable development • Transparency • Harm prevention • Human autonomy • Discrimination • Rule of law • Accountability • Human oversight • Innovation • Explainability • Values • AI literacy • Life cycle • Employment and future of work • Standards • Consumer protection • (Criminal) Justice • Impact assessment • Health and safety • Labour • Economy • Empowerment • Gender equality • Cybersecurity • Children's rights • Proportionality • Value chain • Codes of practice • Good administration • Human-friendly technology • Resilience

Funded by
the European Union

Table B.2. Comparison of codes and topics for the EU AI Act

SuB_05. EU AI Act		
No. of topics vs. codes	47 (BERTopic - topics) < 59 (NVivo - codes)	
Overlaps	11 concepts: • Risk • Health and safety • Discrimination • Compliance • Trustworthy AI • Stakeholders • Conformity assessment • Cybersecurity • SME • Monitoring • Children's rights	
Divergences	**Unique to BERTopic**	**Unique to NVivo**

	• Regulatory enforcement • Certification • Data protection • Biometrics • Market surveillance • Traceability • Sandbox • AI value chain • General-purpose AI model / GPAI • Harmonisation of rules • Documentation • Oversight • Data governance • AI standardisation / standards • Sector-specific AI • Incident reporting • Code of practice • Metrology • Registration • Military / Defence • Assessment • Employment • Workers' rights • Immigration and border control • Law enforcement • Liability • Confidentiality • Legal/Policy update • Audit • Sensitive data • Emotion recognition • Public spaces • Public procurement • Public services • Computational capabilities • Copyright	• Fundamental rights • Privacy • Transparency • Human oversight • Harm prevention • Innovation • Employment and future of work • Harmonisation • Democracy • Robustness • Proportionality • Environmental protection • AI literacy • Consumer protection • Accountability • Vulnerable groups • Legal certainty • Society • Rule of law • Economy • Value chain • Internal market • Ethical AI • Bias • Human dignity • (Criminal) Justice • Gender equality • Codes of practice • Intellectual property • Values • Good administration • Diversity and inclusion • Reliability • Standards • State of the art • Sandboxes • Safety and security • Impact assessment • Human-friendly technology • Level playing field • Sustainable development • Human autonomy • Explainability

Funded by
the European Union

		• Human rights • Empowerment • Resilience • Media and misinformation • Data science

Table B.3. Comparison of codes and topics for Council of Europe Framework Convention on AI

SuB_09. Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law		
No. of topics vs. codes	4 (BERTopic - topic) < 30 (NVivo - code)	
Overlaps	3 concepts: • Human rights • Democracy • Vulnerable groups	
Divergences	**Unique to BERTopic**	**Unique to NVivo**

	• Risk Mitigation	• Rule of law • Risk • Gender equality • Privacy • Discrimination • AI literacy • Stakeholders • Human oversight • Trustworthy AI • Harm prevention • Innovation • (Criminal) Justice • Children's rights • Human autonomy • Environmental protection • Transparency • Impact assessment • Monitoring • Sustainable development • Human dignity • Society • Employment and future of work • Accountability • Health and safety • Safety and security • Economy • Reliability

Table B.4. Comparison of codes and topics for OECD recommendations on AI

SuB_10. OECD Recommendation of the Council on Artificial Intelligence	
No. of topics vs. codes	8 (BERTopic - topic) < 39 (NVivo - code)

Overlaps	5 concepts: • Stakeholders • Trustworthy AI • Human rights • Democracy • AI lifecycle	
Divergences	**Unique to BERTopic**	**Unique to NVivo**
	• AI governance • Risk management • Sector-specific AI	• Privacy • Sustainable development • Labour • Risk • Innovation • Transparency • Employment and future of work • Safety and security • Economy • Diversity and inclusion • State of the art • Rule of law • Harm prevention • Human autonomy • Environmental protection • Society • Accountability • Explainability • Bias • Robustness • Fairness • Intellectual property • Empowerment • Freedom of expression • Knowledge • Gender equality • Human oversight • Human dignity • Media and misinformation • SME

		• Standards • Consumer protection • Sandboxes • Cybersecurity

Table B.5. Comparison of codes and topics for UNESCO recommendations on the ethics of AI

SuB_11. UNESCO Recommendation on the Ethics of Artificial Intelligence	
No. of topics vs. codes	26 (BERTopic - topic) < 49 (NVivo - code)
Overlaps	**17 concepts:** • Human rights • Ethical AI • Environmental protection • AI literacy • Stakeholders • Gender equality • Transparency • Diversity and inclusion • Impact assessment • Risk • Trustworthy AI • Vulnerable groups • Explainability • Accountability • Bias • Children's rights • Fairness

Funded by the European Union

Divergences	Unique to BERTopic	Unique to NVivo
	• Sustainability • Data governance • AI research • Equality • Healthcare • Auditability • Values • Regulatory enforcement • Global collaboration/cooperation	• Life cycle • Privacy • Monitoring • Discrimination • Human oversight • Sustainable development • Human dignity • Society • Harm prevention • Employment and future of work • Innovation • (Criminal) Justice • Media and misinformation • Health and safety • Proportionality • Rule of law • Safety and security • Democracy • Labour • Autonomy • SME • Standards • Data science • Fundamental rights • Robustness • Consumer protection • Economy • Sandboxes • Intellectual property • Human autonomy • Digital sovereignty • Human agency

Table B.6. Comparison of codes and topics for G20 statement on AI

SuB_13. G20 Ministerial Statement on Trade and Digital Economy		
No. of topics vs. codes	14 (BERTopic - topic) < 45 (NVivo - code)	
Overlaps	12 concepts: • Sustainable development goals (SDGs) • Innovation • Economy • Diversity and inclusion • Stakeholders • Benefits • Safety and security • Trustworthy AI • Gender equality • AI governance • Transparency • AI literacy	
Divergences	**Unique to BERTopic**	**Unique to NVivo**
	• Data governance • AI vulnerabilities	• Society • Employment and future of work • Privacy • Policy • SME • Vulnerable groups • State of the art • Labour • Risk • Accountability • Ethical AI • Standards • Human autonomy • Empowerment

Funded by
the European Union

		• Fairness • Life cycle • Discrimination • Robustness • Environmental protection • Consumer protection • Bias • Explainability • Democracy • Cybersecurity • Legal certainty • Rule of law • Human dignity • Intellectual property • Sandboxes • Level playing field • Human rights • Predictability • Testing