DIVERSIFAIR

D2.1 Deliverable

INTERACTIVE MAPPING OF THE AI REGULATION LANDSCAPE

Co-funded by
the European Union

DIVERSIFAIR

Co-funded by the European Union

D2.1 INTERACTIVE MAPPING OF THE AI REGULATION LANDSCAPE

TABLE OF CONTENTS

INTRODUCTION

STATE OF THE ART, OBJECTIVES AND CONTRIBUTION

With accelerating deployment of artificial intelligence (AI) in applications across private and public sectors, there is mounting evidence of potential risk. Historical injustices may, for instance, be learned by machine learning (ML) algorithms from training data, and these may be perpetuated in society through biased and discriminatory AI systems. To mitigate risks of AI systems, many countries are introducing regulations and ethical guidelines on developing ethical, legal and robust AI also known as trustworthy AI.

The resulting AI related policies, strategies, guidelines and legislation span a large spectrum of hard and soft law, including for example binding rules, voluntary codes of conduct or industry-specific standards. These efforts can be tracked at several levels of governance, ranging from local and national policies to regional and international instruments. In addition, numerous public organisations have commissioned expert reports and think-tanks, and NGOs have actively intervened in the field through position papers, adding important literature to the field.

Navigating this rich regulatory landscape is difficult for interested parties who do not necessarily have a background in law or public affairs but nevertheless, have AI-related queries. To facilitate traversing this field, multiple parties have published overview reports or have set up observatories. For example, the AI Democratic Values Index published by the Center of AI & Digital Policy, the AI Policy Observatory by the Organization for Economic Cooperation and Development (OECD) and the AI Policy Portal by the United Nations Institute for Disarmament Research (UNIDIR) provide information about existing regulations and policy frameworks worldwide.

Our gap analysis shows, however, that these initiatives have several shortcomings that this DIVERSIFAIR project deliverable aims to overcome. For example, the AI Policy Portal (UNIDIR) database is under construction and shows a world map that does not easily offer visual information, nor any insight concerning discrimination. The OECD.AI Policy Observatory provides, in relation to OECD countries, a link to multiple documents, summaries and a timeline per document but offers an impractical download function that allows access to excel files with numerous gaps or broken links. It is also unclear which policies have an accessible (English) underlying file. The AI Democratic Values Index offers a textual report that also analyses national AI policies but no comprehensive visually accessible mapping of existing regulatory initiatives.

Furthermore, and of relevance to this project, these initiatives do not offer specific insights into how existing policies address issues of discrimination and related concepts such as bias, fairness, equality and intersectionality. Hence, we position our deliverable as an interactive mapping of the AI regulation landscape. We aim to map and highlight how discrimination, bias and related harms are addressed (or not) by the current governance landscape. We also aim to employ search and information retrieval (IR) techniques to provide a platform to support exploration and analysis of texts related to AI governance.

The interactive mapping enables users to view the context in which a given keyword is used (e.g. discrimination); providing an understanding of how, to what extent and why a keyword is used. Our contribution is thus to offer an interactive visualisation tool that can help to identify, understand and assess the main perspectives and gaps in relation to discrimination, including intersectional discrimination, in AI regulations globally. Furthermore, to support interactive exploration of texts within regulatory documents we used natural language processing techniques with the aim of increasing the accessibility and interpretability of AI legislation.

TARGET AUDIENCE AND IMPACT

The platform developed as part of this project aims to support the increasing range of groups who need to understand the rapidly changing landscape of AI governance globally. For this project we have focused primarily on the needs of students, researchers and policy makers.

EDUCATION

A central aim of the DIVERSIFAIR project is to support the education of students from a range of backgrounds in issues relating to bias and discrimination in AI. Our visualisation tool will serve as an open-access educational resource for students interested in the regulation of AI, in particular in relation to issues of inequality and discrimination, and as a pedagogical tool for teachers in this sector. The interactive mapping can be used to show, in a visually accessible manner, to what extent, where, when and how different issues (bias, ethics, discrimination, intersectionality, etc.) are addressed. It could be used to support class discussions and research assignments, allowing students to become more informed and to develop critical skills. This tool ensures that information about regulations and the primary text from countries across the world is centralised in one repository and accessible to students to support their work. The kind of studies that this tool aims to support includes (e.g. How do given regulations address the issue of bias?), and to assess the effectiveness, quality and limitations of existing regulatory tools (e.g. Are these tools binding or is compliance voluntary? How is discrimination/bias defined? Does a given instrument acknowledge the existence of AI-driven intersectional discrimination?). The interactive mapping will be used in the trainings planned under the DIVERSIFAIR project (e.g. with partners like Women in AI, Women4Cyber, Turing College), and by the teachers involved in the project at Sciences Po and UCD.

To support exploration of the texts of documents pertaining to AI governance, the interactive exploratory platform will enable students to submit queries in language form. The model's interactive nature allows students to receive near-immediate responses to questions related to AI governance. An AI literacy component will be delivered to support student's ability to critically engage with the platform. The objective of this is to increase the accessibility and interpretability of legislative text to students from a broad range of backgrounds. This can be achieved for instance through the generation of accessible summaries of regulatory texts along with searching for specific topics such as legislation concerning bias.

RESEARCH

The interactive mapping of the AI regulation landscape offers a starting point for conducting further research and qualitative analysis by researchers of all disciplines. The tool allows selecting given keywords related to discrimination to gain insights into their actual use throughout AI regulations worldwide. In addition to this quantitative and geographical overview, the integrated context-viewing functionality facilitates the deployment of complementary qualitative analysis through methods like discourse/policy/legal analysis to better understand the way in which a concept is used (e.g. as part of a binding provisions that prohibits a given practice like “discrimination”, a soft-law commitment to “ethics” or declaration signaling general awareness of the existence of technical “bias”). As these concepts tend to be used interchangeably and overlap, facilitating robust qualitative assessment of their intended meaning in existing policy frameworks is key. Researchers involved in the DIVERSIFAIR project will also use the interactive mapping tools to inform their ongoing and future research projects.

Increasingly, AI governance regulation concerns multiple sectors and therefore multiple areas of research. While legal text is incomprehensible to many without a legal background, the use of natural language processing techniques within an information retrieval (IR) platform aims to support tasks such as summarisation. The aim of this is to improve increased accessibility to legislative texts concerning AI to people from across multiple disciplines.

POLICY MAKERS

This interactive mapping tool allows policymakers to compare existing regulations and policy frameworks in other countries or regions. Such a comparative understanding is key for informing public discussions with stakeholders at national and regional level. The objective of this tool is to also foster discussions around best practices such as which regulations are good to follow or which are good examples of explicit guidelines on discrimination and intersectionality. For policy makers the interactive IR platform aims to support rapid querying of large volumes of regulatory text in an accessible way. This functionality aims to support existing processes providing a tool to rapidly extract key information.

USER GUIDE AND DEFINITIONS

The main feature of the AI Governance IR platform is the information retrieval functionality, which allows users to enter natural language text in the form of queries and receive answers to those queries in text form. Accompanying this platform in the pilot phase users will be provided with appropriate training to support optimal use of the functionality. This feature is designed to be accessible and support rapid extraction and synthesis of information from a large dataset of literature related to regulation.

The main feature of the interactive visualisation of AI governance is the functionality to select discrimination-related keywords and identify their context of use in AI regulation. This deliverable offers interactive visualisations that show geographical trends and the evolution in time of the ways in which problems related to discrimination and intersectionality are addressed in AI regulations. To do so, it tracks the use of ten different keywords and related terms in all documents collected for the purpose of creating this tool (policy frameworks, strategies, laws, regulations, etc.). Users can interactively tailor their insights to countries or regions they are interested in. Tracing the use of these keywords offers a birds-eye view on how AI-driven discrimination is addressed globally and is a first step towards analysing, assessing and comparing these regulatory approaches. The visual tool and the underlying database (also available in open access) will enable further research initiatives. It also highlights the lack of specific and concrete attention to discrimination, and in particular intersectional discrimination, in AI regulations and policies globally. This deliverable facilitates research into the key topic of AI-driven discrimination whilst decentralising the focus on Western knowledge and countries. By giving visibility to non-Western sources of regulation and valuable policy developments in the Global South (in line with efforts to decolonise knowledge), it aims to facilitate more representative and diverse knowledge production.

Below, we provide definitions of key terms used in this report and the related deliverable.

- We use **discrimination-related keywords or simply the keywords** to refer to the set of concepts we found in AI regulations that are related to discrimination. The full list of related keywords is: fairness, discrimination, equality, inclusion, intersectional, human rights, social justice, bias, ethics and trustworthy. They all have different connotations and are on varying abstraction levels, but all these keywords were used in the documents to signal similar needs and problems in AI. See below for more details about the keyword selection and coding process.
- We use **the documents** or **the AI regulation landscape** to directly refer to our current database of 96 documents that are either national strategies or policies related to AI or international documents that function as standards, guidance or regulation for AI.
- We use the term **the portal** to refer to the publicly accessible hosted webpage where we host the interactive visualisation, key information on the use, limitations and foundation of the visualisation as well as directions to the IR platform for demonstrative educational purposes.
- We use **the term** information retrieval (IR) to refer to the interactive platform that supports natural language querying and answering in relation to the compiled AI regulation dataset.

DATABASE

The foundation of both the interactive visualisation and the information retrieval platform is a database currently containing 94 documents. The documents are either national strategies or policies related to AI or international documents that function as standards, guidance or regulation for AI.

Scope

We selected 70 documents for analysis, focusing on those issued by official bodies. At the national level, we included documents published by governments or national authorities that address AI strategies and policies (e.g., USA's AI Accountability Framework). At the international level, we limited the selection to intergovernmental organisations like the OECD, UN, European Union, and African Union, excluding those from independent research institutes or think tanks.

Data collection

For the collection procedure we have started with the yet established OECD.AI policy observatory[1] and added additional sources from missing countries via the AI policy portal and a few additions through previous exposure of the DIVERSIFAIR team to reach a more global representation of the AI regulation landscape. We do not assume this scope is exhaustive. Given that we build on established databases we do assume it is sufficient as a first start and are dedicated to updating the database at the end of 2025 based on a new search and suggestions from users.

Access and Descriptives

The data set is available in the portal hosted on the webpage for users to peruse[2]. Here you can find descriptive statistics of all documents included. The fields available to sort and filter are:

Document Title: The official English title of the document
Region: The country or region the document applies to.
Owner/Initiative: The owners or parties responsible for the document
Year: The year it was published.
URL: The website page where either the document itself can be publicly accessed or more information may be found.
Classification: This classification pertains to whether a document is national or international.

Most documents in the database were written in the English language and were publicly accessible in a PDF format. However, for 24 documents this was not the case.

Fourteen documents were written in another language (e.g., Latvian, Spanish, Finnish, Romanian), where an English translation was not found or accessible.
For 5 unfinished documents, it was signaled on the OECD.AI policy observatory or on their respective governmental websites that they are in progress but not yet published.

For 7 other documents, it was stated they exist, however, they either could not be found within reasonable effort or the pdf file format publicly available was not processable for a search function or in text analysis within reasonable effort.

INTERACTIVE VISUALISATION

QUANTITATIVE KEYWORD ANALYSIS

The first component of our deliverable is a quantitative keyword analysis of the documents including an interactive visualisation. To demonstrate the presence of guidance, regulation or warning of bias, discrimination or related concepts, we have counted how many times the documents explicitly mention these

[1] https://oecd.ai/en/dashboards/overview)

[2] https://bloomingdata.com/diversifair/

concepts via keywords. This tailors to the first question one may have while navigating the AI regulation landscape through the lens of discrimination, bias and related concepts: are these covered at all?

KEYWORD LIST

Due to the numerous ways the concepts of discrimination and bias may be described as well as the dynamic nature of language where buzzwords come and go, the DIVERSIFAIR team has iterated over multiple sets of keywords and variations of keywords (see Table 1). This was done until the list of keywords covered a diverse set of language styles and most sections in documents related to the relevant concepts were detected.

Keyword selection procedure: Based on input of the DIVERSIFAIR team we started this analysis with a short list: bias, fairness, discrimination and intersectionality. Through perusal of documents using the search function with these terms we noted that too many documents had zero to few hits while they had a significant section on similar concepts. As a result, for documents with more abstract sections on similar topics we included keywords such as "ethics", "human rights" or "trustworthy" and for documents with a different language style we first had in mind we added keywords such as "inclusive", "equality" and "social justice".

Variations: Moreover, we choose to see each keyword in the list as a categorisation of all relevant variations of said keyword. That is, when the data or visualisation shows how many times the keywords: bias, fairness, discrimination are found, this also includes variations such as biases, unbiased, unfair, fairly, non-discrimination and discriminatory. After examining multiple documents for possible variations via word search and reading a few documents fully for possible variations, the list for the submission of the deliverable is as follows:

Keyword	Keyword variations			
Ethics	ethical	unethical	ethically	
Fairness	fair	unfair	unfairness	
Inclusion	inclusive	inclusiveness	inclusivity	
Human Rights	fundamental rights	civil rights		
Equality	equal	inequality	inequalities	
Bias	biases	biased	debiased	unbiased
Discrimination	discriminate	discriminatory		
Trustworthy	trustworthiness	untrustworthy		
Social justice	injustice			
Intersectional	intersectionality			

Table. 1 Keywords and Variations

DOCUMENT PROCESSING

To facilitate the processing of the large set of documents (pdfs) for all keywords (and their variations) a natural language process procedure for sentence extraction was applied. This procedure was done via a small Python script. Each pdf is first processed into machine-readable text and thereafter split into sentences (Table 2). Then, each sentence containing a keyword or multiple is then stored in a table alongside the sentence before and after (to store additional context for eventual interpretation). The columns of the table are the pdf title, document title, the region, the sentence, the three sentence context as well as a column for each keyword. Each keyword column indicates with a 1 whether the sentence contains at least one mention of the keyword and 0 if the keyword is not mentioned (Table 3). Finally, the output is formatted in a common interoperable table format known as a csv file (csv stands for comma separated values). See below an example of the output. As we encourage the public use, contribution and feedback from the community for this educational resource, the python notebook is also available upon request.

Document Title	Sentence Context	Three Sentence Context
UK's National Artificial Intelligence Strategy	These include concerns around fairness, bias and accountability of AI systems.	There is growing awareness in industry and by citizens of the potential risks and harms associated with AI technologies. | | These include concerns around fairness, bias and accountability of AI systems. | | For example , the report from the Commission on Race and Ethnic Disparities raised concerns around the potential for novel ways for bias to be introduced through AI.
(European) Artificial Intelligence Act	Such possible biased results and discriminatory effects are particularly relevant with regard to age, ethnicity, race, sex or disabilities.	Technical inaccuracies of AI systems intended for the remote biometric identification of natural persons can lead to biased results and entail discriminatory effects. || Such possible biased results and discriminatory effects are particularly relevant with regard to age, ethnicity, race, sex or disabilities. || In addition, the immediacy of the impact and the limited opportunities for further checks or corrections in relation to the use of such systems operating in real-time carry heightened risks for the rights and freedoms of the persons concerned in the context of, or impacted by, law enforcement activities.

Table 2: Sample Regulations

Document Title	Classification	Initiative/Owner	Region	Year	URL	Counts Any	Counts Fairness	Counts Discrimination
National Strategy for Artificial Intelligence of the Danish Government	National	Ministry of Finance and Ministry of Industry Business and Financial Affairs	Denmark	2019	https://en.digst.dk/media/19337/305755_gb_version_final-a.pdf	56	2	1
Eqypt National Artificial Intelligence Strategy	National	The National Council for Artificial Intelligence (of Egypt)	Egypt	2023	https://mcit.gov.eg/Upcont/Documents/Publications_672021000_Egypt-National-AI-Strategy-English.pdf	46	1	0
(European) Artificial Intelligence Act	International	The European Parliament and the Council of the European Union	European Union	2024	https://artificialintelligenceact.eu/	158	14	35

Table 3: Dataset Analysis

DESCRIPTIVES OF THE FREQUENCY DATA

As expected, the usage of keywords varies in the 70 documents (written in English). We highlight here a few preliminary insights with aid of Table 4. Notably, the overarching umbrella term, "ethics" (64 out of 70), signaling the motivation and need for reflective and moral considerations is used in most documents. Many documents mention "bias" (47 out of 70) and "fairness" (53 out of 70) which are common in the AI literature, where bias often is more associated with the technical or data component of AI use, and fairness has a more normative and abstract connotation. Moreover, over half of the documents (41 out of 70) mention discrimination, the more legal and everyday associated keyword. Lastly, very few (3 out of 70) mention the activating, nuanced and complex term "intersectionality".

Keyword	Number of documents with a mention	Number of documents without a mention
Social Justice	2	68
Intersectionality	3	67
Trustworthy	39	31
Discrimination	41	29
Bias	47	23
Human Rights	49	21
Equality	50	20
Fairness	53	17
Inclusion	58	12
Ethics	64	6

Table 4: Presence of keyword references

Furthermore, it is noteworthy that three national strategy documents from Estonia, Switzerland and Uganda, had no mention of the keywords in our list. Caution is needed to directly interpret this finding as disinterest to the concepts the keywords refer to. One cannot compare a dedicated ethical framework document from Malta with Uganda's national Fourth Industrial Revolution strategy document tackling multiple innovations in one. More on this caveat for interpretation is described in the limitations section of the document.

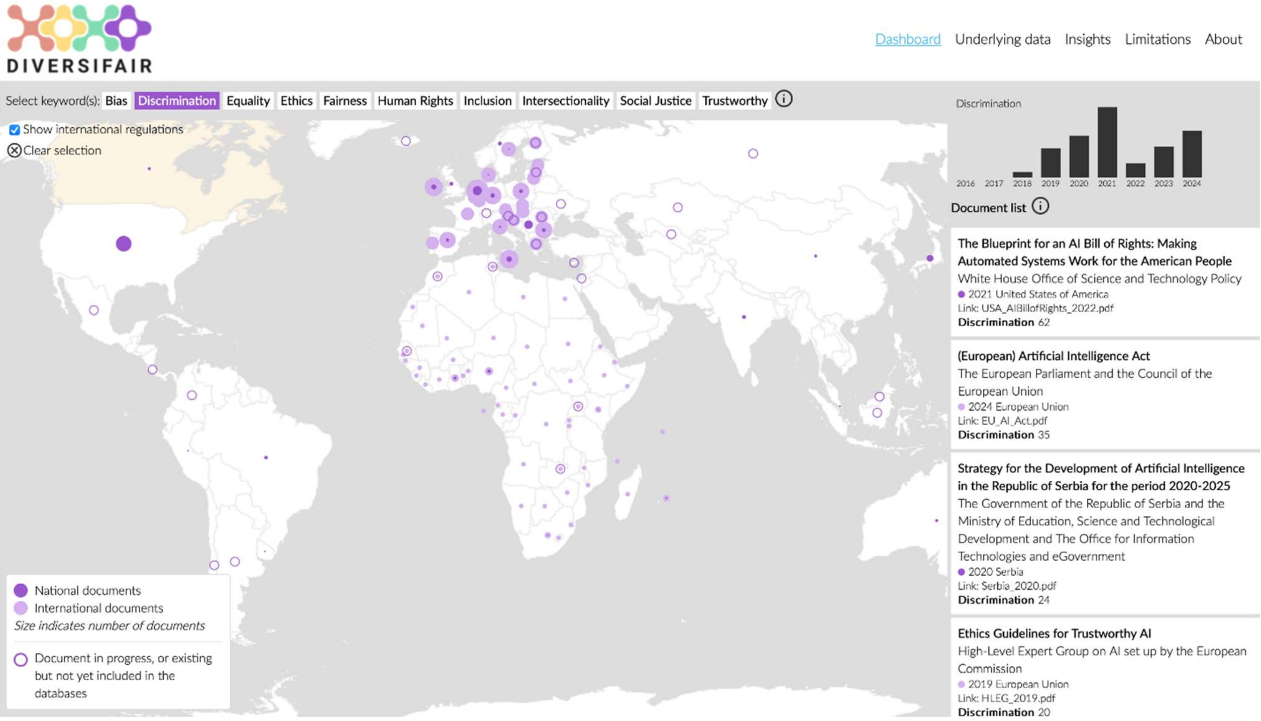
PORTAL

We provide public access to the portal where we host, the dashboard containing the interactive visualisation, the database, as well as a few sections providing additional information on the methodology and aims of the visualisation. Next to that, the portal will also refer to and provide directions to reach the AI Regulation IR Platform.

DASHBOARD

The dashboard hosts the visualisation where students, policymakers and peer researchers can explore and access the AI regulation landscape interactively by clicking on countries, selecting keywords, observing regional patterns as well as over time. The world map shows via the size of bubbles (purple solid circles) the relative frequency of keyword usage. On top of the world map you find all the keywords which also function as filters. The analysis can be shown for a single keyword or multiple keywords simultaneously. The panel on the right shows the document titles and frequencies, as well as the frequencies per year on top. Clicking on one of the countries instead of the keywords, provides the user a list of all related documents with short descriptive information including a link to the document for further perusal.

The shade of the purple bubbles indicates whether the frequency pertains to a national (dark) or a specific international document (light). Purple rings rather than the solid circles signal the presence of either a document with non-English text, that the national strategy or policy is currently in progress or that the document has been published but not found or was not processable.



The Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People
White House Office of Science and Technology Policy
● 2021 United States of America
Link: USA_AIBillofRights_2022.pdf
Discrimination 62

(European) Artificial Intelligence Act
The European Parliament and the Council of the European Union
● 2024 European Union
Link: EU_AI_Act.pdf
Discrimination 35

Strategy for the Development of Artificial Intelligence in the Republic of Serbia for the period 2020-2025
The Government of the Republic of Serbia and the Ministry of Education, Science and Technological Development and The Office for Information Technologies and eGovernment
● 2020 Serbia
Link: Serbia_2020.pdf
Discrimination 24

Ethics Guidelines for Trustworthy AI
High-Level Expert Group on AI set up by the European Commission
● 2019 European Union
Link: HLEG_2019.pdf
Discrimination 20

Figure 1: The Dashboard

DATABASE

All access to our processed data is found in the database section of the portal. First of all, the quantitative frequency data per document as shown before, where one can make standard table customisations such as filtering or sorting the data.

Second, the sentence data for the keywords: bias, discrimination, fairness and intersectionality (885 sentences). Given the aim of the DIVERSIFAIR project, we have chosen to focus our efforts on cleaning the valuable sentence data for these keywords closer to the concept of discrimination rather than the keywords which are more abstract or broader (e.g. "ethics", "human rights", "trustworthy"). The processing was done in two steps. First, we deleted 51 sentences that were just references to other sources, such as titles of research articles. Second, we manually shortened 56 sentences longer than 1000 characters, by deleting part of the sentence whilst maintaining the phrase's essence. The full sentence is still accessible in the Three Sentence Context column of the database.

Third, an example of possible qualitative analysis on the sentence data. To illustrate and inspire users we demonstrate for six documents a preliminary analysis on sentences with the keyword bias may be "coded", summarised concisely for its essence, as part of a thematic analysis.

INFORMATIONAL SECTIONS

To make the portal self-explanatory we have decided to include more descriptive information as guidance in the insights, limitations, and about sections. The sections also invite users to provide their feedback and direct them to this full report for more information. Based on user feedback and updates to the visualisation to come, the sections will be revised.

LIMITATIONS AND MITIGATION STRATEGY

This interactive mapping aims to empower and inform students, researchers and policy makers. However, there are certainly drawbacks to our approach. In the following section we highlight these limitations and what we have done to reduce their impact on the learning experience.

First, the primary added value of the database is to offer frequency data – the number of occurrences of chosen keywords – which is a form of quantitative data. Although valuable on its own, this only signals the acknowledgment of discrimination issues in regulatory and policy frameworks and does not *per se* provide other information regarding how they are addressed. For instance, frequency data does not offer immediate information as to what perspective the concept is approached from, whether bias is framed as an issue of AI or AI is posited as a solution for bias, or even how those two terms are articulated together.

Second, whilst the use of text analytics led to more efficient analysis and therefore allowed us to include a greater number of sources and documents, there was a cost to quality relative to manual data collection. For example, making certain assumptions is necessary order to conduct a large-scale quantitative analysis – for instance assuming that a given keyword is used in a relevant manner in the first place as opposed to e.g. its presence in a footnote or reference – and this simplifies, and reduces nuance in, the analysis. Keywords are also treated as signaling attention to issues of discrimination, fairness and social inequality. Yet, language is dynamic and regional, and different wording could capture similar notions and fall outside the scope of our research.

Third, it is important to acknowledge that documents have different aims, scope and audiences. Some AI strategies offer guidance to the national industry in the form of soft law, other instruments are pieces of hard law and provide binding regulations at national or regional level. Beyond their different scope, the documents also vary in length. Longer documents might address discrimination issues in more depth and therefore have higher frequencies reported. We have chosen *not* to normalise frequency of keywords by the number of pages because we consider length to be a decision of policy-makers and therefore a significant variable when choosing how to address issues of discrimination in AI.

Fourth, the research team has made choices as to which documents were to be processed in this first iteration of our visualisation tool. Our position as researchers based in Western Europe necessarily influences our understanding of, and choices in relation to, this field of research. We prioritised instruments that were readily available in English. This may exclude less globally oriented yet relevant national policies that were only published in the national language. In further iterations, we plan to reflect on using automated translation to extend frequency analysis to non-English documents. Of course, we are aware that automated translation comes with its own drawbacks and that the quality may heavily be influenced by the type of translation and tool chosen. We plan to reflect on further mitigation strategies in the course of future iterations.

As explained above, we have sought to limit these shortcomings by introducing various safeguards. We have laid out a clear and detailed methodology for keyword selection and measurement and coding choices to foster awareness of potential limitations among users. We have implemented a post-processing strategy in relation to core keywords to ensure that we only measure relevant occurrences (eg by excluding keywords used only in references to scholarly articles). Any automated processing also needs to be accompanied by quality inspections. We have therefore conducted sanity checks to ensure the soundness of the processing technique (see our "Under the hood" section for more information). Finally, introducing further information about the context of use of specific keywords and a direct link to the corresponding documents also allows mitigating the shortcomings of quantitative analysis by facilitating further qualitative analysis by users.

AI GOVERNANCE IR PLATFORM

The creation of the AI Governance IR platform was inspired by the challenge that many regulatory documents are dense, complex, and difficult for non-experts to navigate. By using a comprehensive dataset of government sources on AI policy and regulation from around the world (as detailed in the Database section), the IR platform distils diverse legal and ethical frameworks into a user-friendly IR interface. This allows users to explore and compare approaches to AI regulation from different countries and regions with ease. This system is thus particularly valuable for those interested in AI governance but who lack a legal background, as it lowers the barrier to accessing and understanding regulatory documents. We envision, however, that the tool can also be of use for users with legal knowledge – e.g. in a seminar or a training setting, in which the near-immediate responses to comparative and regulatory questions facilitate classroom discussion. Further, engagement with such a novel teaching tool in the latter setting allows students of humanities to come into a meaningful interaction with AI innovations revenant for their future work reality.

The IR platform works alongside the interactive visualisation component of this deliverable, which highlights keyword trends like "bias" and "discrimination" across various regions and time periods. The visualisation offers a quantitative and interactive view of where and how often key regulatory terms appear, and the IR platform enables users to dive deeper into these frameworks and concepts. Users can ask specific questions, compare regulations across countries, or explore parts of the legal text that might otherwise be difficult to interpret.

DATA COLLECTION AND PREPARATION

Co-funded by the European Union

The development of the AI Governance IR Platform relied on a carefully curated dataset of government-published AI policies, strategies, and regulatory documents (for more information, see the Database section of this report). These documents were sourced from open databases and were primarily in English, with a few translated documents included in this model that were not used for the interactive visualisation.

A multi-step pipeline prepared the data for use in the IR platform. All PDFs were converted into text that could be processed by the model. A text cleaning process removed irrelevant or non-informative content (e.g. page numbers and document-specific notations) while preserving the core legal structures of the documents. Consistent formatting was maintained across the data. Legal documents often follow a specific hierarchy, such as sections, articles, and clauses. The data processing system identified these structures and segmented the text accordingly, ensuring that the model could accurately interpret the legal context and references within each document. Given the length and complexity of many legal documents, the text was divided into smaller, manageable pieces or “chunks” of information. This step ensured that the model could process the data efficiently while maintaining a coherent understanding of the larger document. Each chunk was converted into a format that a Large Language Model (LLM) could understand, known as an “embedding.” This step transformed the text into numerical representations, allowing the model to find similarities, retrieve relevant information, and answer user questions. The result of this process was a dataset ready to be integrated into the AI Governance IR Platform.

MODEL DEVELOPMENT

The AI Governance IR Platform was built using OpenAI’s GPT-4, a highly advanced Large Language Model (LLM) designed to understand and generate human-like text. This model can process and interpret complex regulatory language, making it well-suited to provide clear answers on legal documents that are typically challenging for non-experts to navigate.

RETRIEVAL-AUGMENTED GENERATION (RAG)

A core component of our IR platform is the Retrieval-Augmented Generation (RAG) framework. RAG combines the strengths of document retrieval and LLM-based generation to provide users with contextually accurate answers. The process is as follows:

1. The AI regulations data was transformed into a format that the model could understand, known as **embeddings.** Embeddings are numerical representations of text, which allow the system to compare pieces of text and identify the texts that are most relevant to a user’s query. These embeddings are stored in a **vector space**, where the distance between vectors indicates how closely related two pieces of information are.
2. When a user asks a question, the system retrieves the most relevant pieces of information from the stored embeddings using **cosine similarity** (a measure of how similar two vectors are). This means that the model doesn’t just search for keywords but understands the meaning behind the user’s query, retrieving the parts of the regulations that are more relevant.
3. Once the relevant regulatory context is retrieved, the GPT-4 model uses these documents to generate a clear and informed response to the user’s question. The result is an answer grounded in actual AI regulations, ensuring both accuracy and relevance.

As part of our commitment to transparency, the IR platform will cite the sources of information it retrieves from the regulatory dataset. This allows users to trace the origins of the regulatory content provided in the responses. Additionally, in cases where the system uses text from translated documents, a warning will be displayed that notifies users that the information may be subject to differences in interpretation due to translation.

GENERATING MULTIPLE QUERIES FOR BETTER RETRIEVAL

To further improve the system's performance, we integrated a method that generates multiple versions of a user's query, focusing on different aspects such as governance, ethics, or compliance. This approach also finds synonyms and can either widen or narrow the search, as needed. This allows the model to cast a wider net for context retrieval and can find the most relevant information for the query.

USER INTERACTION

The IR platform is designed with a user-friendly interface that simplifies interaction for non-technical users, particularly those unfamiliar with legal jargon. A demonstration of the system has been integrated using a Flask-based application:

EVALUATION AND TESTING

The evaluation and testing of the AI Regulations IR Platform will follow an iterative process, allowing us to refine the model's performance and the accuracy of its responses over time. We will conduct preliminary tests using expert evaluations, where team members with domain knowledge of AI regulations will review the model's outputs. The goal is to assess whether the responses are informative, contextually accurate, and aligned with the relevant regulatory documents. This qualitative assessment provides us with a first layer of feedback on the model's performance.

We will also evaluate the platform on several quantitative metrics like BLEU and ROUGE. These metrics, which are widely used for evaluating text generation and summarisation tasks, offer methods for assessing our model's performance. Both metrics measure the generated text with reference text (the "ground truth"). In our LLM-based system, we employ these scores to evaluate the quality of generated answers by comparing them to the documents retrieved from the user's query. When a user asks a question, our system first retrieves relevant documents, which then serves as the reference text against which we measure the BLEU and ROUGE scores of the LLM's generated answer. This process allows us to assess how effectively our model incorporates the retrieved information into its responses. In essence, these scores compare the similarity between our model's generated answers and the retrieved reference documents.

We plan to improve our model based on these evaluation metrics scores. We will analyse the scores to understand where the model can do better, then make adjustments to various parts of the system, such as how it retrieves information or formulates responses. This process of testing, analysing, and refining is repeated regularly, helping our model to steadily improve in accuracy over time.

Given the complexity of the legal and regulatory domain, the testing process will involve multiple iterations. After each round of testing, we will integrate user feedback and adjust the model's retrieval and/or generation mechanisms. This ongoing process will help ensure that the platform delivers accurate, relevant, and accessible information to its users.

INTENDED USE AND FUTURE APPLICATIONS

The AI Regulations IR Platform is designed to support users by making global AI regulations more accessible. It allows users to compare policies across countries, understand key regulatory concepts, and explore the text of complex legal documents. For example, the system can summarise entire regulatory frameworks, enabling users to quickly identify and analyse important policy. This feature becomes particularly valuable when used in combination with the visualisation tool, which highlights keyword trends within the regulation documents. Users can leverage insights from the visualisation and further investigate specific areas of interest.

In line with our collaboration with consortium partners, we aim to integrate the AI Governance IR Platform into their educational curricula, helping to promote understanding of AI governance in academic settings. By making the code for the model open source, we further support its use in education, allowing institutions to modify or expand the model for specific learning objectives.

As part of our responsible usage guidelines, a disclaimer will be provided to all users:

> *This AI Regulations Chat Model is a research and development tool created by the DIVERSIFAIR Project, based on OpenAI's GPT-4 model with approximately 1.7 trillion parameters. While the model is designed to minimise hallucinations and provide accurate information, misinformation may still occur. This system serves as an educational and research tool, and its responses should not be interpreted as legal advice or binding authority. Users are responsible for their use of the system and are encouraged to submit any bugs or inaccuracies to the administrators.*

The AI Regulations IR Platform will be publicly available after additional development and testing. All code will be open source on the DIVERSIFAIR GitHub repository after publication.

MODEL ACCESS AND FUTURE PLANS

We are actively discussing how to make the AI Governance IR platform more accessible to users by removing some of the financial barriers. One potential solution is to provide guest tokens to allow initial access without paying, with the option for users to log in to their own OpenAI account if they require continued use.

Looking ahead, we plan to further scale and update the model, ensuring it remains up to date with new regulations and legal frameworks as they emerge. There is also potential for integration with other legal or educational tools, creating a more comprehensive ecosystem for understanding AI governance.

Our long-term vision includes expanding the corpus of documents to incorporate additional literature, such as theoretical works on AI governance and ethics, while maintaining our commitment to exclusively using open-source material and adhering to relevant legislation (e.g. copyright, text and data mining or database directives). This expansion will further enrich the model's knowledge base and support its use in broader research contexts.

CONCLUSION

The interactive visualisation and AI Governance IR Platform described herein together serve as powerful tools for making complex legal and regulatory information more accessible and understandable to all researchers. The interactive visualisation provides users with a clear, quantitative view of trends in AI regulations, highlighting key terms across various regions and time periods. This tool enables students to easily explore the presence of these crucial concepts in AI governance, offering insights that can foster critical discussions and further research.

Complementing the visualisation, the AI Governance IR Platform adds depth by allowing users to engage with the underlying regulatory documents. Through its advanced retrieval and response generation capabilities, the platform provides answers to user questions, bridging the gap between surface-level trends and detailed legal frameworks. Further, it also enhances users' critical thinking and understanding of the intricacies related to the deployment of LLM-based systems in the present context. As such, the platform's continuous development, monitoring and evaluation, as well as its responsible use and reflection upon the latter, are central to fostering the necessary skills and harvesting the benefits of this educational experiment.

Together, these tools create a comprehensive resource, empowering users to better understand and navigate the landscape of global AI regulations. In alignment with the goals of Work Package 2, this deliverable contributes to a deeper understanding of the harms and discriminatory impacts of AI systems. Deliverable T2.3 has met its objective of developing an educational tool that maps regulatory, ethical, and technical standards addressing AI bias. The combination of the interactive mapping of AI regulations via the interactive visualisation, and the custom AI governance IR platform provides a comprehensive resource for understanding bias in AI. This deliverable supports ongoing efforts to align AI systems with robust regulations and ethical standards, contributing to the EU's vision of Fair AI.

FUTURE DIRECTIONS

Although we have met the promised requirement of the deliverable, we describe in this section opportunities how we or the international research community are invited to enhance the impact of the deliverable. For the interactive visualisation in the portal, specifically, we foresee the following future directions to foster impact:

- Incorporate an example of qualitative analysis on the sentences and context in a separate information page and provide this annotated data open access to the research community.
- Request feedback from peers in the research field what would make this database more valuable as a starting point for future research.
- Use the visualisation of limited input on intersectionality in the documents for communication efforts towards policymakers in our network together with dissemination partners of DIVERSIFAIR, WomeninAI and Women4Cyber.

For the AI Governance IR platform, we envision the following future directions:

- Expand the document corpus to include additional open-source regulatory and legal documents, as well as theoretical works on AI governance and ethics, to provide users with a broader and more comprehensive knowledge base.

- Explore integration with other legal and educational platforms, making it part of a larger ecosystem of tools designed to promote understanding and research into AI regulations.
- Establish an ongoing evaluation process where the model is regularly benchmarked against new developments in AI regulations to ensure it remains up-to-date and accurate.

For both the interactive visualisation on the portal and the IR platform together, we also foresee shared future directions:

- A seminar with interactive visualisation and platform to students is ideal for showing what the interactive visualisation tool can give in overview and insight and where it may fall short. With the addition of the IR platform, the interrelation can be examined with the curious question whether the LLM may (provide inspiration to) fill this gap? During such a seminar students may be inspired to create additional insights based on the fundamental database, and incorporate these, wherever desired, into the portal. To showcase the importance of critical and nuanced thinking when using data analysis and AI based resources it is also essential to activate students to think of (additional) limitations or possible misinformation that can result from the interactive visualisation and the IR platform.
- The organisation of events (such as workshops and conferences) to cultivate constructive suggestions from users is vital to foster the community connection and durability of our deliverable. With that valuable input, new updates of the portal and the IR platform are emboldened and transparently communicated on a "user feedback" information page.

ACKNOWLEDGEMENTS

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Education and Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.