Mapping AI ethics: a meso-scale analysis of its charters and manifestos

Mélanie Gornet, Simon Delarue, Maria Boritchev, Tiphaine Viard

Mapping AI Ethics: a meso-scale analysis of its charters and manifestos

MÉLANIE GORNET[§], SIMON DELARUE[*], MARIA BORITCHEV[*], and TIPHAINE VIARD[§],

LTCI, Télécom Paris, Institut Polytechnique de Paris i3, SES, Télécom Paris, Institut Polytechnique de Paris, France

The recent years have seen a surge of initiatives with the goal of defining what "ethical" artificial intelligence would or should entail, resulting in the publication of various charters and manifestos discussing AI ethics; these documents originate from academia, AI industry companies, non-profits, regulatory institutions, and the civil society. The contents of such documents vary wildly, from short, vague position statements to verbatims of democratic debates or impact assessment studies. As such, they are a marker of the social world of artificial intelligence, outlining the tenets of different actors, the consensus and dissensus on important goals, and so on.

Multiple meta-analyses have focused on qualitatively identifying recurring themes in these documents, highlighting the high polysemy of themes such as *transparency* or *trust*, among others. The broad term of "AI ethics" and its guiding principles hide multiple disparities, shaped by our collective imaginations, economic and regulatory incentives, and the pre-existing social and structural power asymmetries; through quantitative analyses, we validate and infirm previous qualitative results.

In this paper, we create and present a corpus of charters and manifestos discussing AI ethics through the process of collection and its quantitative analysis using text analysis to shed light on common and distinct vocabularies. Through frequency analysis, hierarchical topic clustering and semantic graph modelling, we show that the charters and manifestos discuss AI ethics along three broad axes: technical documents, regulatory ones, and innovation and business ones. We use our quantitative analysis to back up and nuance previous qualitative results, showing how some themes remain specific while others have fully permeated the space of AI ethics. We document and release our corpus, comprising of 436 documents, charters and manifestos discussing AI ethics. We release the corpus, its datasheet and our analysis, to open the way to further studies and discussions around vocabulary, principles and their evolution, as well as interactions among actors of AI ethics, in order to foster further studies on the topic.

CCS Concepts: • **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Applied computing** → **Sociology**.

Additional Key Words and Phrases: ai ethics, ai ethics manifestos, mesoscale analysis, mesosociology, social worlds

ACM Reference Format:

Mélanie Gornet, Simon Delarue, Maria Boritchev, and Tiphaine Viard. 2024. Mapping AI Ethics: a meso-scale analysis of its charters and manifestos. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3630106.3658545

1 INTRODUCTION

The proliferation of documents around the ethics of Artificial Intelligence has been such that several hundred documents have emerged since the early 2010s. These initiatives to guide AI ethics have been lauded around the world for contributing to opening up the dialogue between different stakeholders on AI benefits and risks, and providing tools to measure the ethical outcome of a decisions. They are seen as a stepping stone to developing AI regulation and binding norms [23]. However, they are also widely criticized for a variety of reasons: their opacity [7], their Western-centrism and claim to universality [30], and their polysemy, that oversimplifies complex ethical debates [21, 34]. These criticisms are partly captured in the following quote: *"Who could be against beneficence? However, problems immediately arise*

Manuscript submitted to ACM

when we start to define what beneficence means." [25]. Together, they contribute to outlining AI's social world [4], and understanding it helps shed light on the way knowledge is constructed in AI.

This trend has attracted a lot of attention and has led to numerous meta-analyses [1, 13, 18, 21, 22, 37, 41], in order to identify common themes and tenets. Both individuals and institutions took hold of this growing space, making it inherently sub-political [3], *i.e.* a space where regulations and societal orientations are decided largely outside of democratic spaces. Describing and understanding these spaces, where the actors and institutions are intertwined with competing interests and multilateral interdependencies, is of crucial importance to understand the social processes and disciplines that span them. This knowledge is key in order for citizens to evaluate the legitimacy of the acting structures and their propositions.

Our main goal in this paper is twofold: using a quantitative lens, we assess and map out the currents that shape the discussions and tension points around AI ethics [1]; we also provide a structured corpus to foster further analyses, and to unify previous works under a common methodology. The core contribution of this paper is the release of our corpus, containing 436 documents, their contents and some metadata. We provide a mesoscale analysis of the social world [4] of artificial intelligence, while comparing ourselves to previous meta-analyses on the topic. This is, to the best of our knowledge, the first publicly available corpus of this kind, and the second-largest existing database on the topic.

The remainder of this paper is as follows: we start by discussing related works in Section 2, and follow directly by describing our corpus's structure and contents in Section 3. We then proceed onto a quantitative analysis of this corpus, exploring term frequencies and topic modeling in Section 4, and explore the areas of consensus and controversy with semantic graphs in Section 5. We finally expose the limitations of our work in Section 6 and conclude in Section 7.

2 RELATED WORKS

Several studies have already analyzed AI ethics charters in search of common principles for AI. The most well-known of these meta-analyses is [21], which investigates more than 80 documents published through 2019. They found that five principles were present in more than half of the documents: *transparency*, *justice & fairness*, *non-maleficence*, *responsibility* and *privacy*.

Since then, several other works have explored a similar corpora of texts to identify common topics related to AI ethics [1, 13, 18, 22, 37, 41]. Some names may differ, but scholars seem to agree at least partially on the major themes present in the texts. Recurring themes that are present in all the meta-analysis are, in no particular order: *privacy*, *transparency*, *fairness*[2], *accountability*[3], and *safety*[4]. Other themes are less common, like *well-being*, *human oversight*, *solidarity*, *explainability*, *collaboration*... However, studies do not always agree on the principles most present in the texts. *Transparency* is the number one principle in some studies [21, 37], while for others, it is *privacy* that prevails [13, 18].

Instead of identifying these principles in the texts, some studies begin by establishing what they consider to be the best set of what constitutes "ethical AI". Notably, [14] builds a set of common principles around the four core principles commonly used in bioethics: *beneficence*, *non-maleficence*, *autonomy*, and *justice*, to which they add a new one, specific to AI ethics: *explicability*. Additionally, [13] offers an overview of the distribution of these themes among

[1]We recognize that the term "AI ethics" is loaded, notably because it shifts discussions towards making AI ethical, rather than its actors and institutions; furthermore, it assumes that AI *can* be made ethical, by ruling out the alternative of not using or sustaining AI. We use it in this article *because* it is the most common term, rather than out of endorsement.
[2]The principle of *fairness* is also referred to as *justice* or *non-discrimination*.
[3]The principle of *accountability* is also referred to as *responsibility*, even though the two notions have different meanings. We will consider here that they belong to the same broad theme, since we are trying to group together rather than separate.
[4]The principle of *safety* is contained in the principle of *non-maleficence* in [21], and sometimes also grouped with *security*.

the documents according to their sector: civil society, government, private sector, government, intergovernmental organization, and multi-stakeholders. However, they do not provide a quantitative analysis of these results. For its part, [41] gives an analysis of the frequency of topics mentioned across sectors. For instance, the principles of *privacy* and *security* are mostly cited by governments while *humanity*[5] and *accountability* are mostly cited by academia. The work by Zeng et al. [41] is the closest to a quantitative analysis. [37] conducts a similar analysis according to the documents' countries of origin. They note that *transparency* is widely cited by all countries around the world, to which can be added *confidentiality* in North America, *fairness* and *security* in Europe and *accountability* in Asia. In another study, [33] looks for key terms in the documents to identify missing themes and show the under-representation of populations from the global south.

Yet, simply looking at the principles does not prevent polysemy, accross countries and contexts. For instance, *privacy*, or *fairness* may be understood differently in the EU or in China [15]. To address this, one can look at the text as a whole, beyond the principles, and see if the vocabulary used differs by sector or country of origin. To investigate these differentials, [31] studies the frequency of identified keywords. For instance, Google or the UK government widely mention *"bias"* and *"fairness"*, but not *"diversity"* unlike the European Commission. However, [31] defines the keywords manually and only displays results by document, not by sector or country. To our knowledge, no temporal analysis of these documents, to see if certain principles are mostly cited in older texts and if some have emerged in recent ones, has been done yet.

Documents related to AI ethics have become so numerous that there are works dedicated to compiling them [6]. These repositories contain all sorts of AI related documents: charters, regulations and laws, technical standards, tools, algorithmic assessments, checklists and other pieces of documents[7]. Others have specialized in compiling a certain type of documents[8]. These various types of AI-related documents are more and more studied through meta-analyses[9]. The authors of [5] take a step back, using expert knowledge to highlight four normative arenas that shape discourses around AI ethics.

In published meta-analyses, common topics and principles are manually found in the texts [13, 18, 21, 37, 41] or are directly defined before being searched for in the documents [1, 14]. Very few studies look at the text as a whole and, to the best of our knowledge, none has applied text analysis to AI ethics charters. However, such approaches have been applied to other types of documents. In relation to the ethics of AI, text analysis has been used to analyze documents related to sustainable AI in energy [35], engineering ethics education [26] or even national AI strategies [27, 28].

3 CORPUS COLLECTION AND OVERVIEW

We now detail our first contribution, the curation of a corpus of documents discussing "AI ethics". We detail our collection process, the formatting of the data, the preprocessing that was uniformly applied to the whole corpus, and finally its availability and ways of future contribution.

[5]The *humanity* principle defined in [41] encompasses, among other things, *human rights*, *dignity*, *freedom* and *well-being*.
[6]See for instance the Council of Europe initiative to compile every documents related to artificial intelligence: https://www.coe.int/en/web/artificial-intelligence/national-initiatives; the Algorithm Watch inventory of AI Ethics Guidelines: https://inventory.algorithmwatch.org/; the AI Ethics Lab's "Toolbox: Dynamics of AI Principles" https://aiethicslab.com/big-picture/; Alan Winfield's blogpost which list texts with their corresponding principles: https://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html [40]; the EthicalML GitHub that points to various AI guidelines and documents: https://github.com/EthicalML/awesome-artificial-intelligence-guidelines
[7]It is the case for the Council of Europe initiative, Ibid; and the EthicalML GitHub, Ibid.
[8]See for instance, the OECD AI Policy Observatory, specialized in policy papers and national strategies: https://oecd.ai/en/; or the Fast.ai initiative that points to academics and institutes to follow: https://www.fast.ai/posts/2018-09-24-ai-ethics-resources.html
[9]See [24] for an analysis of a whole variety of documents or [11] which studies various national AI strategies.

3.1 Collection

To choose which documents to collect, we referred to several existing repositories and meta-analyses. Table 1 shows the overlap between our corpus and previous works, showing that our database is the second-largest, behind the one compiled by the Council of Europe. We obtain a list of documents that were cited at least once in one of the previous works. Since our goal is to provide a quantitative outsight on previous papers, we refrained from adding documents that have never been considered in previous studies, though they do exist. In total, we annotated 730 documents and filtered them using the following list of inclusion criteria:

(1) The document must be freely accessible: we discard any document that we cannot find, that is behind a paywall, or that requires subscription to access;
(2) The document must be written in English, and not be in a draft state: we do not consider documents in another language, or unofficial translations;
(3) The document must discuss artificial intelligence and AI ethics;
(4) The document must be prescriptive : we do not include binding documents, standards, purely technical documents, or any purely descriptive documents. In the case of a largely descriptive document with a few prescriptive recommendations, we include the document and label it "SPI" (Study, Policy or Impact assessment).

We summarize our process as well as the number of documents filtered out at each step in Figure 1. Our rationale for selecting documents is guided by the desire to have a quality analysis of the documents. Having documents of the same nature allows for a more relevant comparison of the vocabulary used. This guides each of our inclusion criteria.

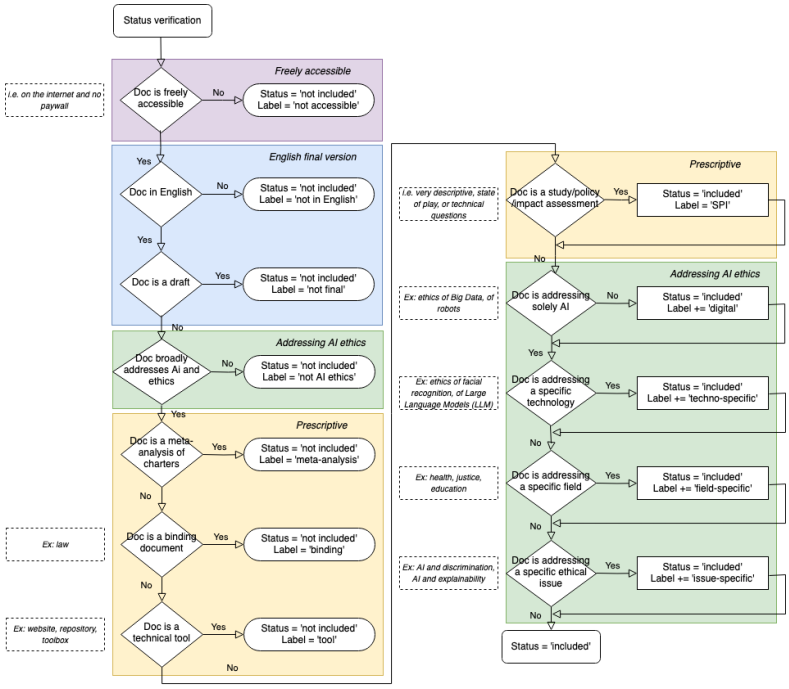
First, we remove non accessible documents. Not accessible might refer to paywalled documents, not found documents, or documents that we cannot automatically scrape (for example, multiple web pages).

We exclude non-official translations to avoid misunderstandings when we cannot ensure the quality of the translation, or when the translation itself imposes an unchecked western bias. For instance, in the document titled "Advisory Board on Artificial Intelligence and Human Society", an initiative of the "Minister of State for Science and Technology Policy" included in [21], the Chinese term usually translated as *harmony* in English, which comes with moral and social preconceptions that are closer to the translator than the original intent; for a concrete example, the interested reader can read the work of Werbach on the Chinese social credit systems [39]. Keeping only the latest versions and official releases allows us to respect the authors' words and to discard obsolete statements.

Selecting only prescriptive documents permits us to discuss how AI should be. On the contrary, more binding documents usually restrain their scope to what is possible or desirable with other constraints (economic, social or technological ones) and thus rather discuss how AI could be. Similarly, study on the state of AI ethics in the world rather discuss how AI is today or will be in the future.

We apply our annotation process to the 730 potential documents, and we include 436 of them in our corpus, only including documents that have been cited in at least one previous meta-analysis. Each document was assigned for review to one of the authors, and so we have 4 annotators. To ensure consistency between annotators, we collegially annotated 10 documents, and then selected 10% of the original base to be blindly annotated a second time by three of the four annotators. We measure inter-annotator agreement with Fleiss's κ, which takes its values between -1 (perfect disagreement) and 1 (perfect agreement), a value of 0 indicating a chance assignment. We obtain $\kappa = .712$ ($95\%CI, .577$ to $.847, p < .001$), indicating high agreement between the annotators [10]. We break down in Table 1 the overlap in

[10] https://statistics.laerd.com/spss-tutorials/fleiss-kappa-in-spss-statistics.php



(a) Flowchart documenting our annotation process. Each rhombus corresponds to a choice (of inclusion or annotation). The left column corresponds strictly to exclusion criteria, such as language, accessibility, prescriptivity, while the right column all consider included documents with specific, cumulative comments, such as documents focused on specific fields (*e.g.* healthcare).

Step	# of documents
Considered documents	730
Accessible	677
In English, is final	608
Addresses AI and AI ethics	510
Is prescriptive	436
Included	436

(b) Table summarizing the number of documents discarded at each step. Reading key: $730 - 677 = 53$ documents were discarded because we could not access them.

Fig. 1. Flowchart and numeric breakdown of the inclusion criteria for the collection of the MapAIE corpus.

included documents with previous papers. It shows that we included documents used in a variety of studies. However, we could not include all of them, as many did not meet our inclusion criteria.

3.2 Formatting

We list all documents, their title and institution of origin, the URL address at which we reach them, and our annotations in a tabular file. All documents are either in PDF or HTML format. We automatically download each document, and extract its contents using Python scripts. In the case of PDF files, we use the Python library `PyPDF2`[14]. In the case of HTML files, the situation is more complex, as just downloading the page includes a lot of boilerplate content (menus, headers, links to other pages, etc.). We design an algorithm to extract the main content of the page, by finding the deepest element in the HTML structure tree (DOM) that contains the largest content.

[14] https://pypi.org/project/PyPDF2/

		MapAIE	**[21]**	**[13]**	**[37]**	**[18]**	**[14]**	**[41]**	**[1]**	**[12]**	**AW**	**CoE**	**[40]**	**EthML**
MapAIE (our paper)	(100.00%)	436	73	32	15	20	6	65	35	9	114	360	20	12
Jobin et al. [21]	(87.95%)	73	83	20	3	18	6	32	22	7	74	69	18	6
Fjeld et al. [13]	(86.49%)	32	20	37	2	12	5	21	31	3	27	31	11	6
Tidjon et al. [37]	(51.72%)	15	3	2	29	3	2	4	3	2	4	8	2	5
Hagendorff [18]	(95.24%)	20	18	12	3	21	4	14	16	3	18	17	6	4
Floridi et al. [14]	(100.00%)	6	6	5	2	4	6	6	5	2	6	6	4	2
Zeng et al. [41]	(78.31%)	65	32	21	4	14	6	83	24	6	50	54	13	6
Attard-Frost et al. [1]	(76.09%)	35	22	31	3	16	5	24	46	4	32	36	10	7
Eur. Parliament [12]	(75.00%)	9	7	3	2	3	2	6	4	12	7	7	6	1
Algorithm Watch[11]	(71.70%)	114	74	27	4	18	6	50	32	7	159	122	19	8
Council of Europe[12]	(60.50%)	360	69	31	8	17	6	54	36	7	122	595	18	10
Winfield [40]	(83.33%)	20	18	11	2	6	4	13	10	6	19	18	24	4
EthicalML GitHub[13]	(80.00%)	12	6	6	5	4	2	6	7	1	8	10	4	15

Table 1. The matrix of documents in our dataset (MapAIE), compared to previous works. Reading key: 87.95% of documents in Jobin et al. [21] are in MapAIE, and 73 documents are included both in MapAIE and Jobin et al. [21].

Theme	Keywords
fairness	fairness, algorithmic fairness, bias
xai	xai, lime, shap
regulation	personal, right, law, harm, gdpr, discrimination, article, biometric, regulation
agi	agi, artificial general intelligence

(a) Our themes and keywords.

	Fairness	XAI	Regulation	AGI
Fairness	0.48	0.04	0.37	0.05
XAI	0.04	0.04	0.03	0.01
Regulation	0.37	0.03	0.51	0.05
AGI	0.05	0.01	0.05	0.07

(b) Co-occurrences of themes in our corpus.

3.3 Pre-processing

We automatize preprocessing for text fields. All text is processed using the python libraries `BeautifulSoup`[15] and `NLTK`[16]. `BeautifulSoup` is designed to manipulate HTML structures and extract textual contents; the *Natural Language Toolkit* (`NLTK`) provides tools for working with human language data, for the text itself. We systematically remove numbers, URLs, and stop words [17] present in the NLTK `english stopwords` corpus, and put all text in lowercase. Then, we retrieve all the lemmas appearing in the text, i.e all the canonical forms corresponding to the words[18] composing the text; for example, the lemma "train" corresponds both to the words "training" and "trained". Finally, we remove all lemmas that contain less than 3 characters.

Our final corpus comprises of 436 documents. We release online [19] the tabular file listing all documents (included or not), the corpus itself, as well as its datasheet [16] and the parsing and preprocessing code. Due to intellectual property limitations, we cannot publicly release the scraped contents as is. Instead, we release the code required to download and build the corpus in a single command. All materials are available publicly, on academic storage (provided by our

[15] https://pypi.org/project/beautifulsoup4/
[16] https://www.nltk.org/
[17] Words that are very commonly used in a language, such as "the", "is", etc. in English.
[18] We recognise that the term "word" is not the one generally used in linguistics to describe a textual content. For the sake of simplicity, we use it in this article to stand for "token" or "word form", or "lexeme".
[19] http://mapaie.telecom-paris.fr

institution), as well as on GitHub. In order to ensure reproducibility and open the way to new analyses, we publicly document our process, allowing individuals to include new documents so that anyone can contribute to enlarging the corpus, provided they follow our annotation guidelines.

3.4 Creating thematic corpora

From the initial corpus, we build several thematic corpora along guiding themes identified in previous meta-analyses. These corpora do not form a partition of the corpus: a document can belong to multiple corpora. We specifically discuss analysis and results of these subcorpora along themes we identified (in Section 4.2.1) and themes identified by Jobin et al. [21] (in Section 4.2.2). We display the themes we identify and the associated keywords in Table 2a, and show the co-occurrences of themes in Figure 2b.

4 ANALYSIS AND RESULTS

4.1 Exploratory analysis

Let us start by examining a few generalities about the corpus. First of all, a comment on the length of the documents. We nuance the common preconception that AI ethics charters are short and of little practical use [19]: we notice instead a difference between purely positional statements and more fleshed-out documents, with roughly 20% of documents exceeding 10000 words (around 20 pages of text).

Word	Frequency
data	43412
systems	16852
use	16663
intelligence	16242
artificial	14702
human	14334
also	13583
public	12126
rights	11759
system	11757
research	11485
may	11234
development	10195
digital	10186
new	9907

(a) Lemmas with the highest term frequency across all documents with their total word counts.

Word	# Documents
use	331
data	331
public	319
information	318
development	318
also	317
intelligence	314
human	314
systems	313
new	311
research	306
make	305
society	305
privacy	305
social	304

(b) Lemmas with the highest document frequency with their total document occurrence.

Words	Frequency
artificial intelligence	43669
data protection	39668
personal data	36742
machine learning	34862
human rights	33757
ai system	33031
data use	31289
data protection regulation	30926
data collection	30179
public sector	30149
european commission	30130
impact assessment	29891
member states	29118
general data protection regulation	28411
best practices	27333

(c) n-grams with the highest co-occurrence frequency.

Fig. 3. Term frequency and document frequency of lemmas in the corpus. Reading key: the lemma "data" appears 43412 times in the whole corpus; it is used in 331 documents among the 436 that constitute our corpus and is the second most used lemma. It occurs in a bigram with the lemma "protection" 39668 times, in a trigram with the lemmas "protection regulation", and in a 4-gram.

Most frequent terms across the corpus are represented in Table 3a. Terms like "system" and "data" are over-represented, while other lemmas follow a rapid decay. Notably, "artificial intelligence" is much less used than the term data for

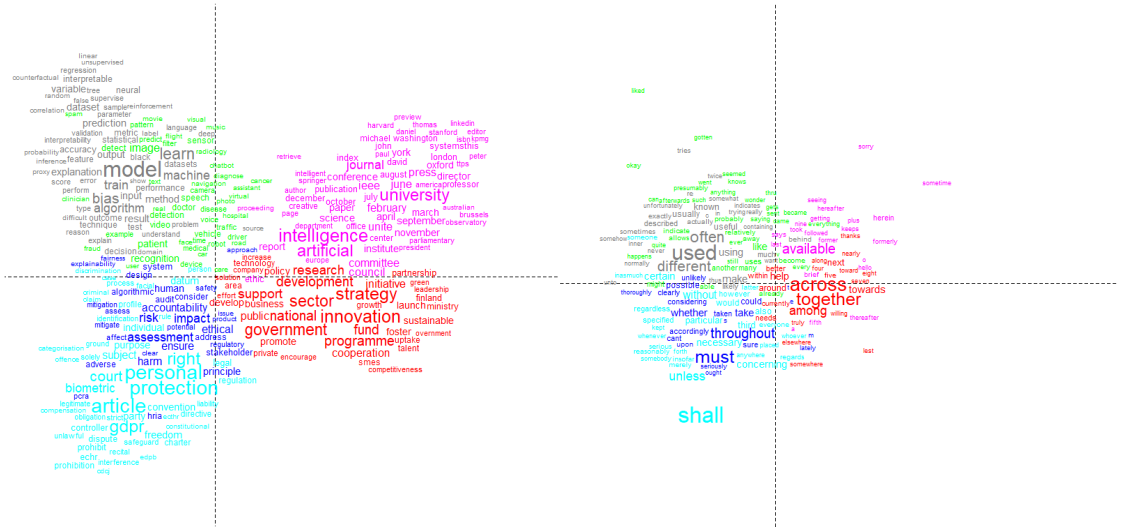


Fig. 4. Two-dimensional visualisation of the clusters obtained with hierarchical classification on our corpus, obtained with correspondence analysis. The size of words is proportional to their importance (in terms of number of occurrences) in the corpus, and distances are linear. Explained variance: 61.5%.

instance. Yet we need to keep in mind that the lemma "AI" is removed during preprocessing because it is too short and thus does not appear in this list. Document frequency, however, follows a much slower decay (Figure 3b): many terms are present in several documents. The term most common to documents is "use", followed by "data" and "public". "Artificial intelligence" only appears in 314 documents out of 436; the remaining 122 documents typically discuss AI in a narrower sense, *e.g.* "machine learning for face recognition", or use the word "AI" without explaining what it stands for, which we deemed fully in scope.

We show in Table 3c the most frequent n-grams, *i.e.* sequences of n words that frequently co-occur together (for example, "artificial intelligence" is a 2-gram, and "data protection regulation" is a 3-gram). n-grams give us more meaningful insights into the themes and discussions of the corpus, by capturing common turn of phrases. Unsurprisingly, *artificial intelligence, data protection, machine learning* and *human rights* come up as very frequent, with most of the top n-grams being related to legal and regulatory texts (*personal data, fundamental rights*, etc.). It also highlights the central role of European institutions as regulators of artificial intelligence as of the writing of this paper.

4.2 Understanding recurring themes and common topics

We continue our study by an analysis of common themes in our corpus. We first analyze the whole corpus in Figure 4, and discuss the main currents of thought we find. The clusters are built using the Reinert method [32], a hierarchical clustering method , and the results are visualised with a correspondence analysis [20]. Each text in the corpus is analyzed through the lens of co-occurrences of lemmas in fixed size text segments. We use segments of size 40, though we examined different segment sizes (between 2 and 200) to ensure the stability of the results. All analyses were made using the IRaMuTeQ software [20].

[20] http://www.iramuteq.org/

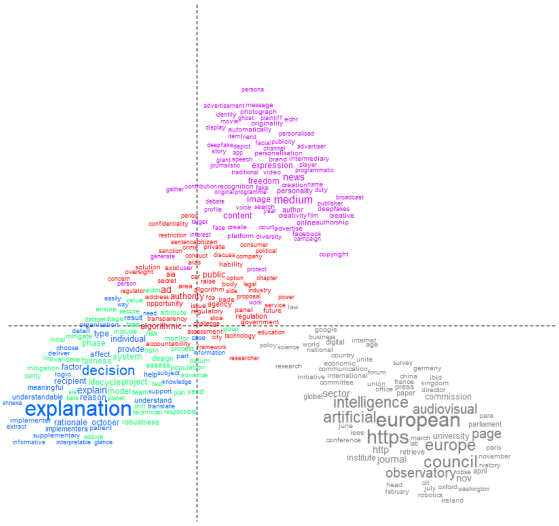
The general clustering in Figure 4 highlights 6 clusters, corresponding to different themes: two of them are technical (centered around models/techniques and applications, respectively), two are more regulatory (centered around laws and policies, respectively), and the last two correspond to a business-oriented and a very generic cluster, respectively. In Figure 4 right, we see how different common words are associated with each cluster: while the technical and applicative clusters use descriptive language ("used", "often"...), the regulatory cluster uses prescriptive one ("must", "shall"...).

To each main current (technical, regulatory, innovation) corresponds a different paradigm: the technical documents largely follow a model-driven paradigm, while the regulations and laws follow a data-driven paradigm; finally, documents discussing innovation largely frame it as strategies, programs and plans in order to keep a competitive edge. We note the absence of a *user-driven* paradigm, examining the role of human beings in relation to AI and its ethics. Though this is partly captured by regulation and law in the form of *data*, the correspondence between human and (personal) data is nothing but systematic, even though it is a common assumption of machine learning models [8]. Indeed, unlike humans, data is typically reduced to atoms of information and vector-based. This irreducibility, along with works around the ethnography of algorithms, studying how end-users react and use data algorithms, have shown effects of resistance and decoupling between institutional discourses and practical use [9, 10]. We also note the absence (at least at this scale) of strong discussions on social justice issues, even though sexism, racism [29, 42] and labour inequality [38] are well-documented problems in artificial intelligence models and datasets.

4.2.1 Analyzing themes in the corpus. We analyze themes that follow [21] in Section 4.2.2, but we have also decided to expand this analysis to themes that have emerged since 2018. Our hypothesis is that analyzing these themes brings a complementary perspective. The visualisations of these analyses are presented in Figure 5, each subfigure corresponding to one of the themes outlined in Section 3.4.

We notice that *explainable AI* (XAI, Figure 5a) remains a technic-dominated area, with very specific technical vocabulary (explanation, decision; bottom-left cluster), with another technical cluster on top-right more centered around applications of explainable AI, with the terms *deepfakes, content, diversity, fake, etc.* Well separated is a regulation cluster (bottom right), centered around the European Union, with few meaningful words. In the case of the *Artificial General Intelligence* (AGI, Figure 5b), a term that is commonly tied to the moral panic that AI systems will overcome human beings in the long-term, we see that the technical cluster completely disappears, while the regulatory one drastically shrinks: in other words, AGI is not a topic of interest from the technical point of view, and marginally so in the case of regulation. Instead, the terms mobilized focus on standardisation (bottom right), human and moral considerations (top), and medical and health considerations. Quite interestingly, the last two subcorpora, related to *fairness* (Figure 5c) and *regulation* (Figure 5d) are both similar to the global analysis of the corpus. We take away from this that (i) fairness has become a commonplace term, that is reproduced in all areas of "AI ethics" (though, possibly with polysemy), and (ii) that most documents in our corpus discuss regulation, indirectly or not.

4.2.2 Confronting with themes in the literature. The analysis presented in the previous section gives us an opportunity to confront the corpus against recurring themes identified in the literature. We filter our corpus using the keywords outlined in [21]; we then run the same preprocessing, clustering and correspondence analysis on each sub-corpora. When using the keywords and themes identified in [21], while the clusters' words change marginally, the gist of the results stay the same, with clusters separating the data along three lines: technical, regulatory and innovation/business, in addition to a more generic cluster. The reasons for this relatively small changes are multifactorial: firstly, the keywords listed in [21] are quite generic (listing, among others, "disclosure" and "showing" under the theme *Transparency*). This is not necessarily a problem in the original case, which focused on a qualitative analysis and where researchers can decide



(a) Explainable AI subcorpus. Explained variance: 67.88%.

(b) Artificial General Intelligence subcorpus. Explained variance: 60.2%.

(c) Fairness subcorpus. Explained variance: 64.1%.

(d) Regulation subcorpus. Explained variance: 66.6%.

Fig. 5. Thematic analysis along our subcorpora, visualised with a correspondence analysis.

on a case-by-case basis if a word matches a theme; furthermore, the concepts have spread across actors and institutions since 2018, year of [21], and they are now sufficiently widespread that they are not markers of differenciation anymore. We list here the themes the authors identified, explaining the key changes they induce in terms of text analysis, *i.e.* how the four main clusters (technical, legal/regulation, innovation and generic) evolve and change; a cluster becoming smaller and more specific is typically due to less documents discussing this paradigm in the subcorpus.

Transparency (257 documents). There are no changes along this theme, showing how transparency has permeated discourses around AI ethics and is now used indiscriminately in technical, regulation and innovation documents.

Justice and fairness (78 documents). Another widely used theme. The business and innovation cluster shrinks in terms of size, while the legal and regulation one becomes larger; the technical cluster becomes more specific, explicitely citing algorithmics fairness related terms.
Beneficence, non-maleficence (83 documents). In this case, the technical and regulation clusters get closer and tighter, while the innovation and generic clusters remain mostly unchanged. This is due to both technical and regulation documents mentioning these topics, in extremely similar terms.
Responsibility (154 documents). There are no cluster changes along this theme, even though the technical cluster shrinks in size, and become slightly more specific.
Privacy (106 documents). The legal and technical clusters fuse into a single one, highlighting more specific applications (such as, for example, *"homomorphic cryptography"*, leaving the rest relatively unchanged.
Freedom and autonomy (25 documents). In this theme, the clusters become more specific, discussing jobs and work-related issues, specific technical terms such as "bias" or "model manipulation"; other clusters gather terms related to creativity and cooperation, along with a small regulation cluster focused on the implementation of legal texts.
Trust (279 documents). There are no specific changes, showing that the topic has permeated AI ethics.
Sustainability (159 documents). While the core results remain unchanged, the law/regulation and technical more separated, indicating less overlap in how these topics are discussed by regulatory and technical documents.
Dignity (124 documents). The main results do not change, apart from the legal and regulation cluster becoming much larger than the technical one. Indeed, *dignity* has a strong legal connotation and is routinely used in this context.
Solidarity (32 documents). The legal/regulation and innovation clusters remain stable. However, the technical cluster becomes more specific (citing terms around *interpretability, explanation, fairness...*), and the generic cluster is replaced by a more interesting one, centered around jobs, employment and economy.

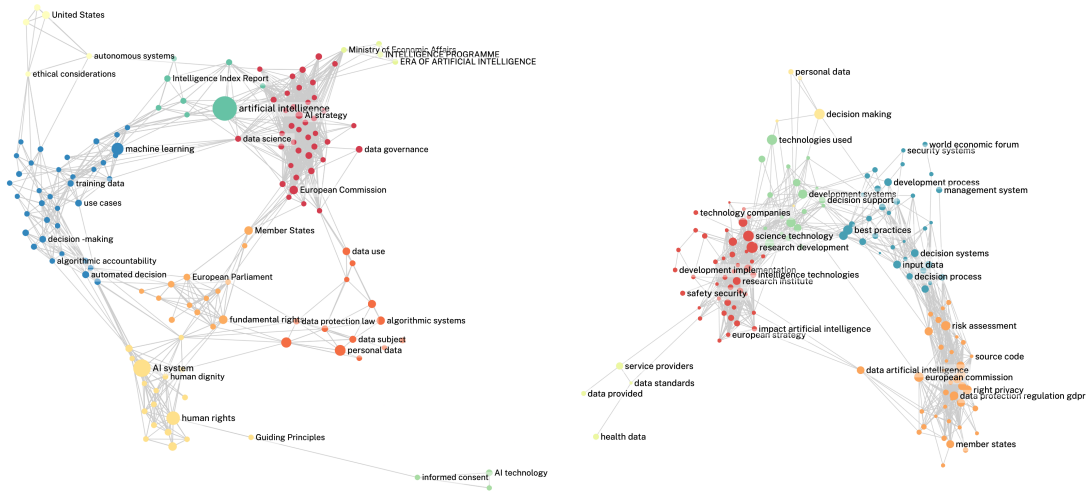
In conclusion, while some themes have been consistently picked-up on by the various actors and institution, this is not the case for all of them, especially the more specific ones. The number of documents associated to each theme sorts the themes in a different order than the one in [21], though we are not the first to notice this (see Section 2).

5 AREAS OF CONSENSUS AND OF CONFRONTATION

Delving into the specific vocabulary, and relative importance of the areas around which discourses are structured, this gives us the possibility to look into both consensual and confrontational areas. In this section, we use semantic graphs to identify some controversies inherent to modern artificial intelligence. These graphs, by showing us words that are at the frontier of clusters (*i.e.* typically linked to nodes of their own cluster as well as other ones, as for example "Member States" in Figure 6a), show us the themes where semantic and semiotic qualms happen.

5.1 Methodology

We build co-occurrences graphs. A graph is a tuple $G = (V, E)$, where V is a set of nodes ($\{u, v, w, \ldots\}$) and E a set of edges (*i.e.* pairs of nodes, $\{(u, v), (u, w), \ldots\}$). We will consider graphs to be undirected (*i.e.* $(u, v) = (v, u)$) and loopless (*i.e.* $u \neq v$). We build the graphs so that nodes are n-grams in the corpus (with $2 \leq n \leq 5$), and there is an edge between two nodes if the n-grams significantly co-occur in the corpus. Significance is tested via a chi-square (χ^2) test, which compares the observed and expected frequencies of the outcomes of variables. The size of the node in the visualization is propotional to its degree, *i.e.* the number of connections with other nodes: the higher the number of connections, the bigger the circle representing the node. Nodes are then colour-coded using the Louvain algorithm [6], a common graph



(a) Co-occurrence graph of *n*-grams.

(b) Co-occurrence graph of *n*-grams filtered using TF-IDF.

Fig. 6. Co-occurrence graphs. Reading key: the node corresponding to the bigram "artificial intelligence" is part of the green cluster in the left graph. It connects the blue cluster with the red cluster. It is represented with a wide circle as it is highly connected.

clustering algorithm that detects subsets of nodes that are more connected together than with the rest of the graph, by optimising an objective function. Notice that, due to its aggregative design, Louvain typically favours larger clusters.

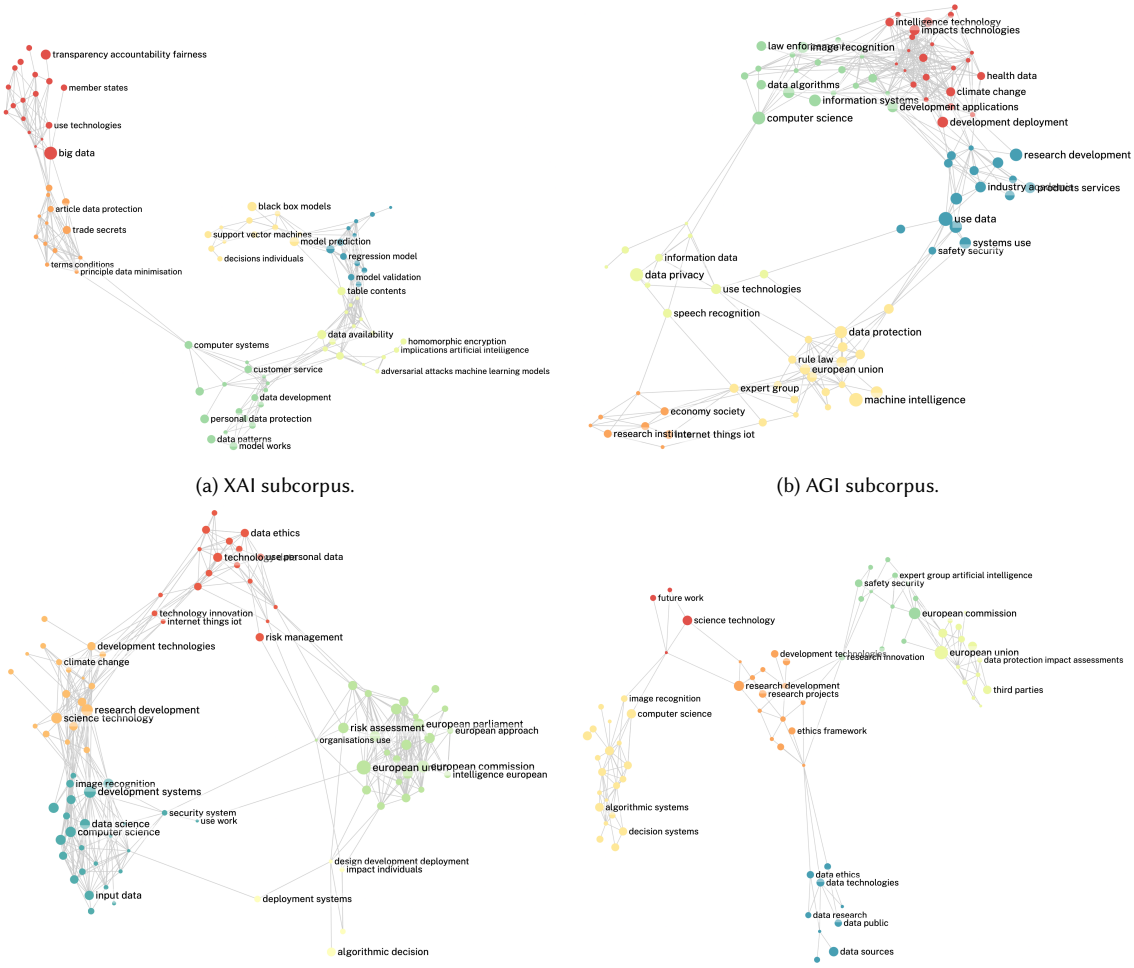
5.2 Results

We show in Figure 6 graphs built from our corpus [21]. The left co-occurrences graph displays relationship between *n*-grams. The right co-occurences graph displays the same relationship with *n*-grams filtered on important words, as per the *Term Frequency - Inverse Document Frequency* (TF-IDF) metric, which measures the extent to which a word appears a lot in a document (*Term Frequency*), but seldom in most documents of the corpus (*Inverse Document Frequency*).

From Figure 6a, we notice the predominant position of the term "artifical intelligence", connecting two major communities related to AI techniques (blue) and governance (red). We observe that "machine learning" belongs to technical usage, while business actors and impact assessment writers tend to focus more on "data science". Interestingly, while the European Parliament and Council are together in a cluster related to fundamental rights (orange), they are separated from the European Commission, which is closer to governance topics (red). This outlines the role of the European Commission as a provider of expertise, rather than a regulatory or legislative instance.

Filtering *n*-grams on the most important terms allows to avoid the influence of generic terms such as "artifical intelligence". In Figure 6b, clusters are slightly modified and we find four major communities: (i) Research & Development (red); (ii) technological systems (green); (iii) management and process (blue); and (iv) protection and regulation (orange). Interestingly, the importance of individual rights related nodes is lowered after considering TF-IDF; "human rights" or "human dignity" disappear to the benefits of themes such as "right privacy" or "data protection". Moreover, we notice the absence of terms such as *fairness, ethics* or *explainability*, as they appear widely through the corpus: the terms "ethic[s|al]" appears in 81.6%, "fair[ness]" in 72.05%, "explain[able|ability|ation]" in 67.4% of documents. Overall, we

[21] Interactive graphs are available at (a) and (b).



(a) XAI subcorpus.

(b) AGI subcorpus.

(c) Fairness subcorpus.

(d) Law subcorpus.

Fig. 7. Thematic graph analysis, along our subcorpora. A link between two terms means that they co-occur significantly in the subcorpus.

observe strong semantic proximity between technically-oriented clusters (red and green), but highlight how distant such considerations can remain from operational and economical aspects (blue) as well as from regulation vocabulary (orange).

We further analyze thematic co-occurrences graphs by filtering our corpus using the keywords in [21] [22]. We observe in Figure 7c that filtering the data using commonly used terms such as "Fairness" only induces minor change in the co-occurrences graph; the different clusters and their relationships remain stable. Similarly, the "Artificial General Intelligence" (AGI) graph, in which all communities are kept in their original proportions, suggests that the term is broadly used by all categories of actors in the AI world. On the other hand, by focusing on documents containing "XAI",

[22]Interactive thematic subgraphs are available at (a), (b), (c) and (d).

we exhibit a highly technical graph where regulatory considerations are almost not present at all. At the other side of the spectrum, the "Regulation" graph in Figure 7d evokes several aspects of AI regulation in addition to the technical references. However, we observe how business oriented terms are absent from this perspective. These two examples suggest a strong semantic boundary between these two worlds.

6 LIMITATIONS

Let us outline some limitations of our work. The most obvious limitation is related to restraining our search to documents in English. Indeed, we made this choice to be able to compare texts on the same semantic level; but it leaves out multiple documents that have been written in other languages. We refrained from making any conclusions about the geographical origin of documents discussing AI ethics, even though we collected the data: we do know that our corpus is heavily biased in that regard. This bias stems notably from our country of origin, the language inclusion criteria, and the fact that we prioritised documents that were already mentioned in previous meta-analyses that exhibit such bias themselves.

Other limitations concern the methods used for our quantitative analysis. To begin with, the exploratory analysis is based entirely on word occurrences. However, this depends a lot on how the words are counted, which is influenced by our preprocessing method. For example, "AI" was filtered out by our preprocessing, so "artificial intelligence" has a lower word count than it would have if both versions of that term were counted together. Furthermore, both analysis methods we use are good at capturing common themes, rather than themes corresponding to less frequent terms or terms specific to one document. For instance, the theme of power struggles is not completely absent in the corpus but, because it is not statistically central, it is dismissed by the model. Lastly, for intellectual property reasons, we cannot publicly release the textual contents of the corpus, only make them downloadable. This means that documents becoming unavailable in the future will not be downloaded.

7 CONCLUSION

In this paper, we collected and created the first public corpus of AI ethics related documents with their contents, rather than a list of documents matched to a reading grid. We showed that our corpus covers significant portions of most well-known previous studies, and we use it to confirm past results. In addition to a pre-trained model, it can be used to measure and quantify word embedding bias in such documents, using current debiasing methods [17, 36]. After shortly describing the corpus and the term frequencies, we quantitatively analyzed it along two axes: we use textual analysis to highlight the main areas being discussed, and semantic graph analysis to identify points of controversy. We analyse both the main corpus and four subcorpus, as well as compare our results to previous works.

Let us now detail a few perspectives this work opens. The most straightforward one is linked to the corpus: adding new documents to the corpus is made easy, and since all our code is available, makes reproducing our work with more data accessible. Another interesting perspective would entail setting up a data visualisation platform, to search, visualize and explore the corpus's documents, making our corpus a valuable tool for a wider audience. An interesting perspective is to study the temporality of these documents and concepts, in particular to outline arbitrations that durably shaped AI ethics. We would also like to explore the polysemy of words used in the AI field, by applying more advanced natural language processing methods to analyse the corpus' semantic contents. Using the Abstract Meaning Representation (AMR [2]) framework, we can extract semantic graphs from each of the documents in the corpus, and then apply methods from graph studies to the obtained semantic graphs in order to identify the underlying structures.

Acknowledgments. The authors would like to thank Matthieu Labeau for taking the time to share his expertise in Natural Language Processing, and Valérie Beaudouin for her comments and insights.

REFERENCES

[1] Blair Attard-Frost, Andrés De los Ríos, and Deneille R Walters. 2022. The ethics of AI business practices: a review of 47 AI ethics guidelines. *AI and Ethics* (2022), 1–18.

[2] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, Sofia, Bulgaria, 178–186.

[3] Ulrich Beck. 1992. Modern society as a risk society. *The Culture and Power of Knowledge. Inquiries into Contemporary Societies* (1992), 199–214.

[4] Howard S Becker. 1976. Art worlds and social types. *American behavioral scientist* 19, 6 (1976), 703–718.

[5] Bilel Benbouzid, Yannick Meneceur, Nathalie Alisa Smuha, and Liz Carey-Libbrecht. 2022. Four shades of AI regulation. *Reseaux* 232233, 2 (2022), 29–64.

[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[7] Jean-Christophe Bélisle-Pipon, Erica Monteferrante, Marie-Christine Roy, and Vincent Couture. 2022. Artificial intelligence ethics has a black box problem. *AI & SOCIETY* (2022).

[8] Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, and Liz Carey-Libbrecht. 2018. Neurons spike back. *Réseaux* 211, 5 (2018), 173–220.

[9] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017), 2053951717718855.

[10] Angèle Christin. 2020. The ethnographer and the algorithm: beyond the black box. *Theory and Society* 49, 5-6 (2020), 897–918.

[11] Murat Durmus. 2021. Overview of National AI-Strategies.

[12] European Parliament. Directorate General for Parliamentary Research Services. 2020. *The ethics of artificial intelligence: issues and initiatives*. Technical Report. Publications Office.

[13] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020).

[14] Luciano Floridi and Josh Cowls. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* (2019).

[15] Pascale Fung and Hubert Etienne. 2022. Confucius, cyberpunk and Mr. Science: comparing AI ethics principles between China and the EU. *AI and Ethics* (2022).

[16] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[17] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).

[18] Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and machines* 30, 1 (2020), 99–120.

[19] Anne Henriksen, Simon Enni, and Anja Bechmann. 2021. Situated accountability: Ethical principles, certification standards, and explanation methods in applied AI. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 574–585.

[20] Hermann O Hirschfeld. 1935. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 31. Cambridge University Press, 520–524.

[21] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[22] Arif Ali Khan, Sher Badshah, Peng Liang, Muhammad Waseem, Bilal Khan, Aakash Ahmad, Mahdi Fahmideh, Mahmood Niazi, and Muhammad Azeem Akbar. 2022. Ethics of AI: A Systematic Literature Review of Principles and Challenges. In *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022 (EASE '22)*. Association for Computing Machinery, 383–392.

[23] Lyse Langlois and Catherine Régis. 2021. Analyzing the Contribution of Ethical Charters to Building the Future of Artificial Intelligence Governance. In *Reflections on Artificial Intelligence for Humanity*, Bertrand Braunschweig and Malik Ghallab (Eds.). Springer International Publishing, 150–170.

[24] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics* 26, 4 (2020), 2141–2168.

[25] Luke Munn. 2022. The uselessness of AI ethics. *AI and Ethics* (2022).

[26] Osama Nasir, Saamia Muntaha, Rana Tallal Javed, and Junaid Qadir. 2021. Work in Progress: Pedagogy of Engineering Ethics: A Bibliometric and Curricular Analysis. In *2021 IEEE Global Engineering Education Conference (EDUCON)*. 1553–1557.

[27] Theodoros Papadopoulos and Yannis Charalabidis. 2020. What do governments plan in the field of artificial intelligence?: Analysing national AI strategies using NLP. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*. Association for Computing Machinery, 100–111.

[28] Gleb Papyshev and Masaru Yarime. 2023. The state's role in governing artificial intelligence: development, control, and promotion through national strategies. *Policy Design and Practice* (2023), 1–24.

[29] Edmund S Phelps. 1972. The statistical theory of racism and sexism. *The american economic review* 62, 4 (1972), 659–661.

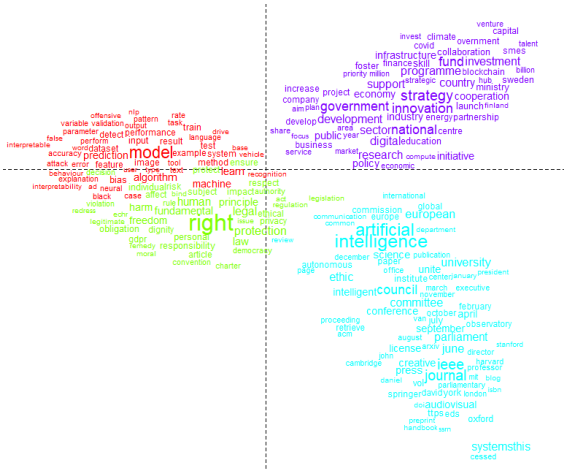
[30] Emmanuel R. Goffi and Aco Momcilovic. 2022. Respecting cultural diversity in ethics applied to AI : a new approach for a multicultural governance. *Misión Jurídica* 15, 23 (2022), 111–122.

[31] Connor Rees and Berndt Müller. 2022. All that glitters is not gold: trustworthy and ethical AI principles. *AI and Ethics* (2022).

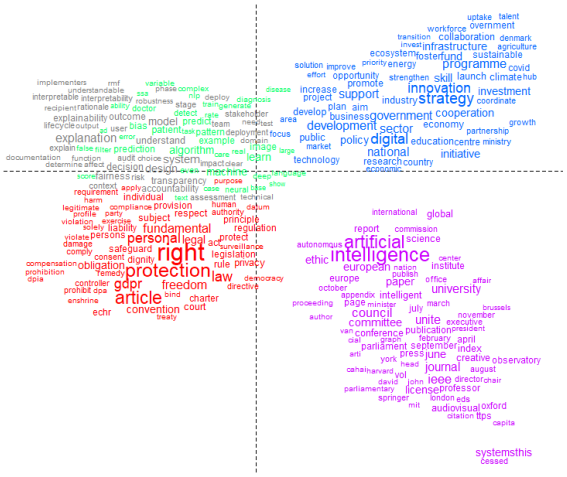
[32] Max Reinert. 1990. Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application. *Les cahiers de l'analyse des données* 15, 1 (1990), 21–36.
[33] Cathy Roche, P. J. Wall, and Dave Lewis. 2022. Ethics and diversity in artificial intelligence policies, strategies and initiatives. *AI and Ethics* (2022).
[34] Mark Ryan and Bernd Carsten Stahl. 2020. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society* 19, 1 (2020), 61–86.
[35] Tahereh Saheb, Mohamad Dehghani, and Tayebeh Saheb. 2022. Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis. *Sustainable Computing: Informatics and Systems* 35 (2022).
[36] Sarah Schröder, Alexander Schulz, Philip Kenneweg, Robert Feldhans, Fabian Hinder, and Barbara Hammer. 2021. Evaluating Metrics for Bias in Word Embeddings. *arXiv preprint arXiv:2111.07864* (2021).
[37] Lionel Nganyewou Tidjon and Foutse Khomh. 2022. The different faces of ai ethics across the world: a principle-implementation gap analysis. *arXiv preprint arXiv:2206.03225* (2022).
[38] Paola Tubaro, Antonio A Casilli, and Marion Coville. 2020. The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society* 7, 1 (2020), 2053951720919776.
[39] Kevin Werbach. 2022. Orwell That Ends Well? Social Credit as Regulation for the Algorithmic Age. *U. Ill. L. Rev.* (2022), 1417.
[40] Alan Winfield. 2019. An Updated Round Up of Ethical Principles of Robotics and AI.
[41] Yi Zeng, Enmeng Lu, and Cunqing Huangfu. 2018. Linking artificial intelligence principles. *arXiv preprint arXiv:1812.04814* (2018).
[42] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist—it's time to make it fair.

A CONFRONTING WITH THEMES IN THE LITERATURE: SUPPLEMENTARY PLOTS

We add in this section the plots we used for interpretation in Section 4.2.2 *Confronting with themes in the literature.*

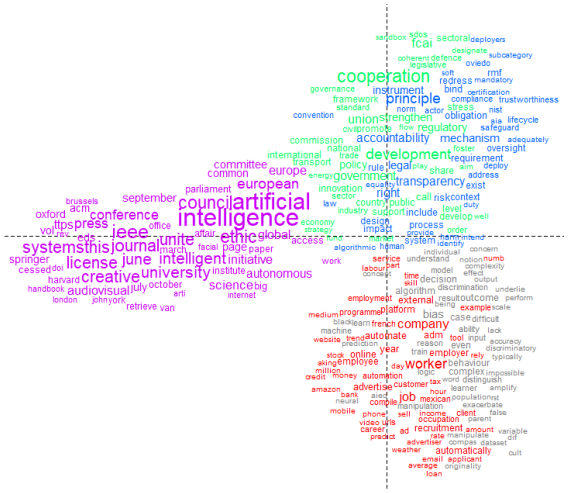


(a) Beneficence and non-maleficence subcorpus. Explained variance: 74.24%.

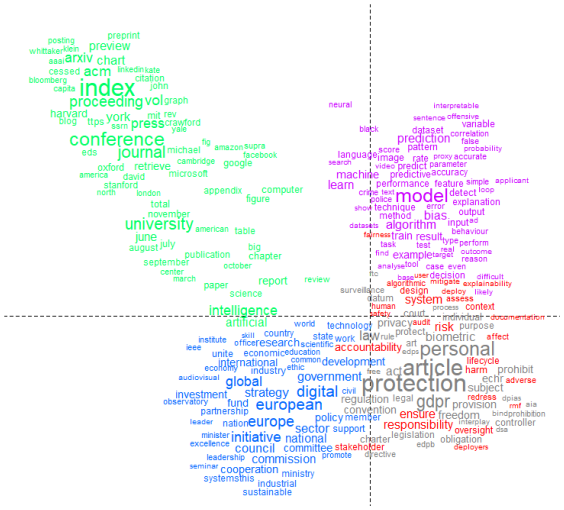


(b) Dignity subcorpus. Explained variance: 64.91%.

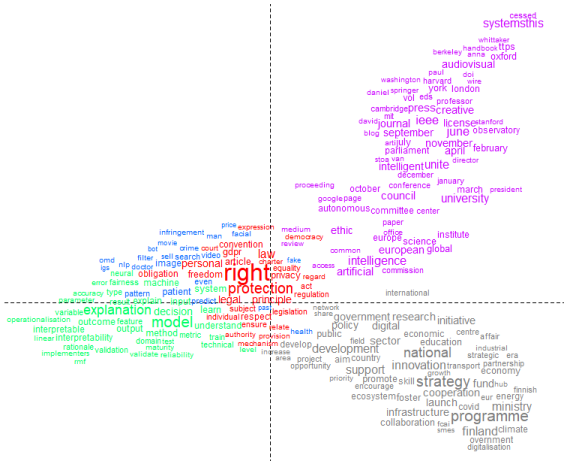
Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009



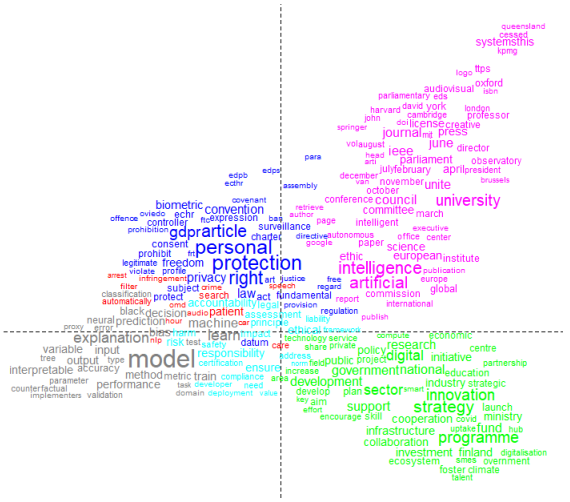
(c) Freedom and autonomy subcorpus. Explained variance: 68.62%.



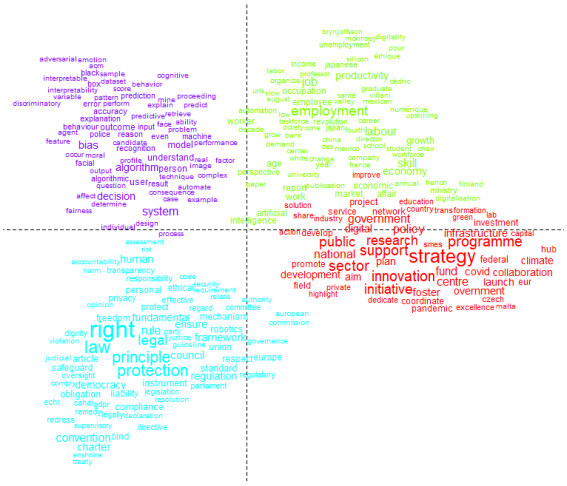
(d) Justice and fairness subcorpus. Explained variance: 69.73%.



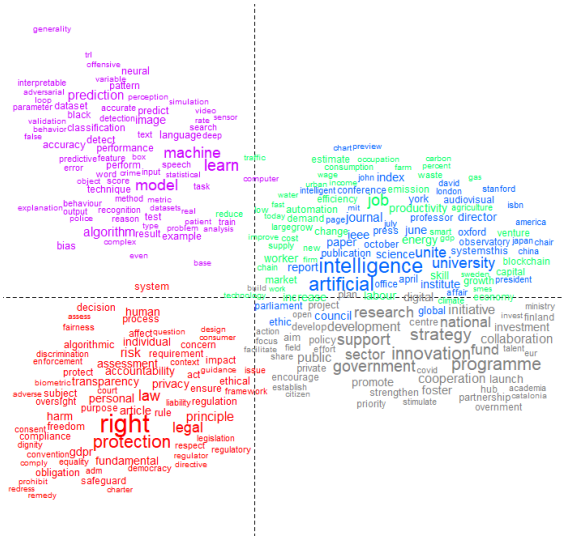
(e) Privacy subcorpus. Explained variance: 63.58%.



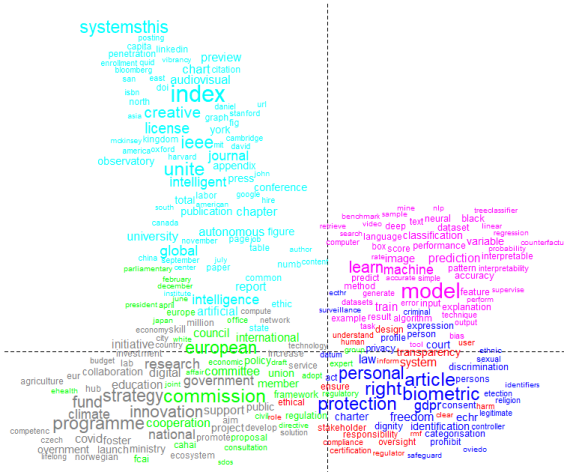
(f) Responsibility subcorpus. Explained variance: 57.71%.



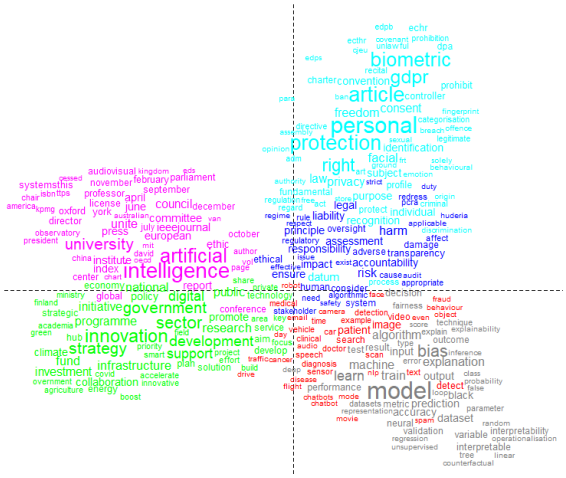
(g) Solidarity subcorpus. Explained variance: 80.88%.



(h) Sustainability subcorpus. Explained variance: 65.31%.



(i) Transparency subcorpus. Explained variance: 58.82%.



(j) Trust subcorpus. Explained variance: 57.39%.

Fig. 8. Thematic analysis along subcorpora extracted from themes in the literature, visualised with a correspondence analysis.