13

ANALYSIS OF THE PRINCIPLED AI FRAMEWORK'S CONSTRAINTS IN BECOMING A METHODOLOGICAL REFERENCE FOR TRUSTWORTHY AI DESIGN

Daniel Varona and Juan Luis Suárez

Introduction

The approaches that seek to solve social problems rooted in the use of technology, specifically in the use of artificial intelligence through its machine learning core, are as dissimilar as the problems that they try to solve (Varona, 2018; Varona, Suarez, & Lizama-Mue, 2020). They also delineate an area of research that is increasingly attracting interest from the academic and professional communities. In a short period of time, the problems have evolved through different stages, such as issues of bias (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019), fairness (Mehrabi et al., 2019; Sahil & Rubin, 2018; Trewin, 2018), and principled artificial intelligence (AI) (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020; Mittelstadt, 2019), to mention the latest. The latter approach seeks to explore the feasibility of using international human rights law as the basis for developing what is denominated "trustworthy AI", recognizing that the variables associated with most social problems stemming from the use of artificial intelligence and machine learning are reflected in the aforementioned corpus of law.

So far, the principled AI approach, according to Fjeld et al. (2020), includes 35 documents published between 2016 and the last quarter of 2019. The authors added five other documents published in the first half of 2020 which share the same scope. All together, the principled AI international framework gathers authors and signatories such as government entities (15: 37.5%), inter-government entities (3: 7.5%), multistakeholders (8: 20%), civil society (5: 12.5%), private sector (8: 20%), and church (1: 2.5%). The documents included in the principled AI international framework can be categorized as action plan (1: 2.5%), commitment (1: 2.5%), considerations (2: 5%), general recommendations (1: 2.5%), guidelines for developers (1: 2.5%), policy principles (12: 30%), policy usage (1: 2.5%), principles (15: 37.5%), principles and recommendations (1: 2.5%), recommendations (4: 10%), and standardization recommendations (1: 2.5%). The selection criteria were mainly directed by gathering regulatory initiatives ruling the design and use of AI solutions at the country, region, and global levels whose authors

 DOI: 10.4324/9781003024583-15

have actual agency for it. Appendix I lists the documents included in the analyzed corpus for principled AI.

The principled AI framework described in the previous paragraph seeks to set the basis for trustworthy AI in form of a public policy framework; however – as the results in this chapter show – it must still mature to also become a methodology framework adopted by artificial intelligence developers. In the other hand, the current lack of international standards or any other auditing mechanism containing references to the framed AI principles to be used in the development of artificial intelligence solutions suggests that there exist certain difficulties in assimilating the principles proposed within the referenced framework and its adjustment into a useful methodological tool for the design of such AI solutions.

The objective of this chapter is to identify some of the causes that hinder the construction of a methodological tool based on the public policy framework of principled AI. To achieve this, we analyze the set of documents that constitute the international framework of public policies for principled AI, integrating holistic and computer techniques into a design thinking approach that helps us deconstruct the problem.

In this chapter, we adapted a six-step methodology for data science projects proposed by Lizama-Mué in (Lizama-Mue & Suarez s.f.) that involve: (1) problem study, (2) data collection, (3) data preparation, (4) modeling, (5) test and evaluation, and (6) communication. The method has been successfully applied in other research projects like (Suarez & Lizama-Mue, 2020; Monroig Vives, 2017; Segarra, 2018) and follows an approach to problem solving similar to the design thinking methodology, which is what we are aiming for in this chapter due to the characteristics of the data being processed.

This research also aligns with this volume's efforts to highlight the interdisciplinarity needed to tackle social problems arising from the use of technology, in particular the expansion of machine learning to many personal and social domains. In addition, the study uses holistic techniques complementing other computational techniques while identifying difficulties the current AI international regulatory framework based on the International Human Rights Law is facing to become a methodological tool for a trustworthy AI-aligned design.

Method

In order to identify the theoretical-methodological elements that may become difficulties in the construction of a mechanism that facilitates a fairness-aligned design for artificial intelligence, better defined as trustworthy AI based on the principled AI international framework determined by the studied corpus, we first established clusters taking into consideration datapoints like year, month, country, and city of publication; type of author; and type of document, which help us support other analyses exhibited in the "Results and Discussion" section; then we used the datapoints ITPractitionersAuthorDistribution and PolicyMakersAuthorDistribution to assess the balance in the background distribution among authors in IT and related sciences and non-IT authors.

The executed analytical methods included n-gram extraction, weighting word combinations while examining those with more value for the text based on their frequency, first over the documents to have a global view of the narrative used along the corpus; second over the principles and guidelines sections to narrow the analysis to the proposed principles and suggested methods for adapting the principles; and third on the principles' declarations, seeking the variables used to describe trustworthy AI that were absent up to that point in the examined documentation.

Then we applied lexical diversity analytics to the full corpus. This is a practical technique to evaluate the closeness of different body texts and speeches regarding their content when

different language structures are exploited. We did so with the intent to find patterns of ideas used by the different types of authors and in different types of documents that can further support the operationalization of variables in the text the AI designers need to comply with when the principled AI international framework becomes a methodological reference for AI design. Then verb extraction from both the principles and guidelines sections was executed to identify the verb taxonomy and group the suggested actions along the corpus. And last, topic modeling was performed over the principles' declarations, helping us to explore the apparent disconnection between the ideas behind the principles' declarations with the remaining text body on the corpus. The analyzed documents were gathered from each author's website in PDF format, then converted to plain text to facilitate data processing.

The data was prepared using Python (Oliphant, 2007; Python Software Fundation, s.f.), a generic modern computing language widely used for text analytics. Python's development environment is enriched with libraries like Gensim (Rehurek & Sojka, 2010), which we used for topic modeling; SciPy (Virtanen et al., 2020) tools including Pandas (McKinney, 2010), used for structuring the data; Matplotlib (Hunter, 2007), used for data visualization; IPython (Pérez & Granger, 2007), used for interactive computing and programming; and NLTK (Loper & Edward., 2009), used for natural language processing. The entire corpus was filtered removing stop words like prepositions, articles, and pronouns, among others, and repeated headings, footers, and margin notes. Thus, we were able to work with a consolidated and semantically robust corpus.

Using NLTK, the corpus was analyzed searching for n-grams that helped to describe from a macro perspective the content of the documents in a primary stage and then to narrow the scope of the analysis into smaller sections, like the principles and the guidelines, which were manually extracted from the documents and stored separately. Also, the verbs from the principles and guidelines sections were extracted and analyzed in order to review the taxonomy of the proposed actions along the corpus. We extracted the parts of speech that matched with verbs, grouped them using the lexeme similarity criteria, and added the size of the lexeme to the lemma.[1]

Most of the modeling was done in topic extraction. In that regard, we applied the latent Dirichlet allocation (LDA), Gensim's implementation (Blei, Ng, & Jordan, 2003), to detect topics among the principles' declarations. LDA is a generative probabilistic model in which each document is considered as a finite mix over an underlying set of topics. Each topic is represented as a set of words and their probability, which means that we were able to rank the topics on the corpus and the keywords in each topic. As mentioned before, doing the topic extraction helped to corroborate some inferences resultant from previous parts of the analysis.

As for the test and evaluation, we centered our verification efforts on the topic extraction part of the analysis due to being, among the techniques used in the study, the one with an additional intrinsic uncertainty. If not conducted properly, the outcome of the topic extraction could lead to misleading conclusions. One of the most important issues related to topic modeling with LDA is to know the optimal number of topics (k) that should be examined. To overcome those issues, we built different LDA models with variable values of k, computed the coherence for each topic, and selected the model with the highest coherence value. Consequently, the best results were found with ten topics and ten keywords per topic, presented in the "Results and Discussion" section.

The authors want to note that applying the mentioned natural language processing (NLP) techniques allowed us to learn a set of specific traits from the texts, detailed in the "Results and Discussion" section, that otherwise would be more difficult and time consuming to identify using a classical/traditional approach. Although neither the amount of analyzed documents nor

their length demanded complex NLP procedures, we were able to take advantage of the simpler methods within the NLP domain for text processing.

Results and discussion

As mentioned in the introduction, there are currently numerous efforts by several entities – governments, non-governmental agencies, private sector, and so on – to achieve a fairer artificial intelligence. In this chapter, we focus on the efforts whose main emphasis seeks to standardize the responsible design and subsequent proper use of artificial intelligence.

In the authors' opinion, trustworthy AI is an emerging research interest whose starting point can be located around the last quarter of 2016 and rapidly reaches maturity just two years later, in 2018, likely due to the intensity of the geopolitical interests of several nations around AI (Suarez, 2018). In this respect, 2018 and 2019 are the years exhibiting a peak of publications of the documents forming the analyzed regulatory framework for principled AI, with an average of 15.5 documents each.

When exploring the most involved actors in producing or as signatories of these documents included in the regulatory international framework for principled AI, the United States occupies a clear first position among countries with the highest involvement (27.5%), followed by China (12.5%) and France (10%). The three of them can be related to half of the documents analyzed in this study. Figure 13.1a shows more information about the origin and authorship of the documents. The left wing of Figure 13.1a exhibits the distribution of documents in the analyzed corpus attending to their type, while the right wing displays the document distribution attending to their author type.

In Figure 13.1a, it can be noted that in the United States, China, and France, the documents are produced mainly by authors listed as multistakeholders and government entities, to reuse the catalogue proposed in Fjeld et al. (2020). There is a greater representation of the private sector in the United States when compared with the remaining countries. The fact that governments are active entities in the production of these documents shows their commitment to solving the ethical and social problems that the current degree of penetration of artificial intelligence in almost every aspect of daily life entails. The origins of these problems have been located in the use of artificial intelligence and in the early stages of its design, according to the themes highlighted by most of the principles proposed in the corpus. This is coherent with similar findings from previous studies (Varona, 2018; Varona et al., 2020).

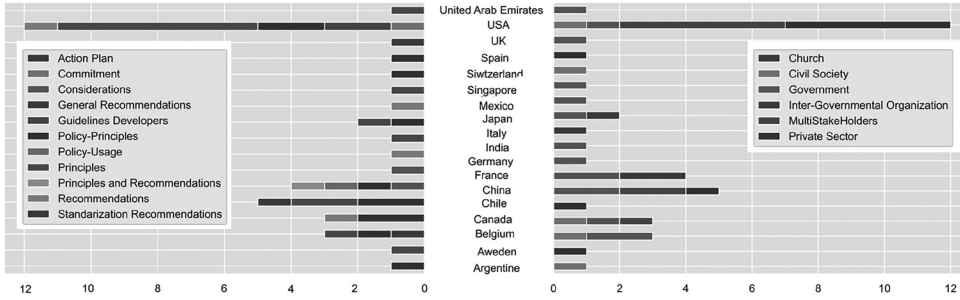

Figure 13.1a Document and author type distribution per country

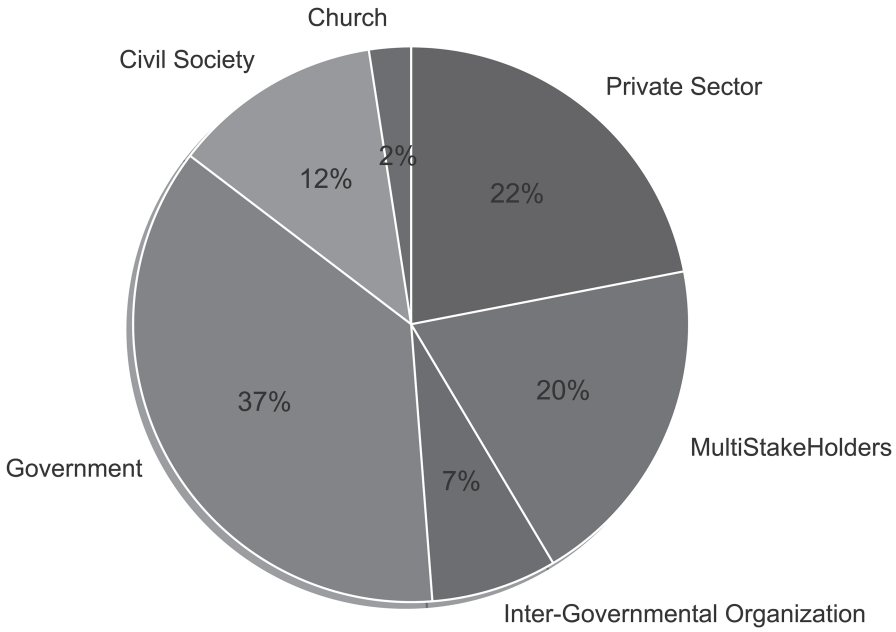

Figure 13.1b Author type general distribution

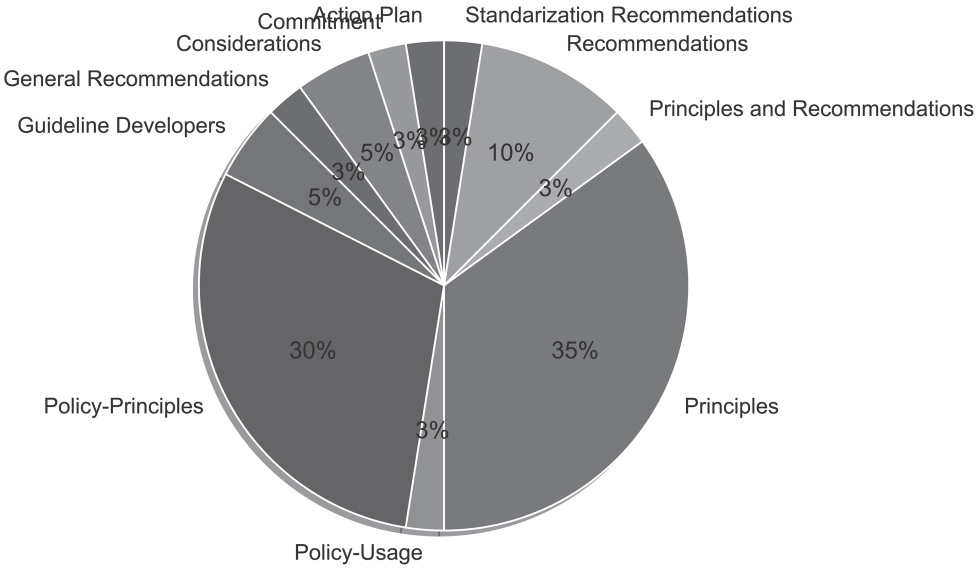

Figure 13.1c Document type general distribution

The authors recognize that the list of countries presented in Figure 13.1a represents developed countries with high technological drive; we believe it is necessary to stress that social problems as result of the use of technology are problems that transcend the digital gap between developed and developing countries. Hence, although the provisions of the studied documents should be considered equally valuable for and by all countries, it would be important to have a larger and more equitable presence of countries from different parts of the world so that a diversity of ethical and social problems can be studied in connection with the use of AI in specific contexts.

Figure 13.1b synthesizes the different types of documents. The leading role of governments in authoring the regulatory frameworks for the design and use of artificial intelligence becomes clear throughout the corpus. This government-type author has authorship in more than a third (37.5%) of the documents, followed by multistakeholders and private-sector type authors, with 20% each. In contrast, the values exhibited by the civil society author type denote the need for greater activism on their part, as these organizations would provide important input from many different stakeholders affected by the decisions made by using artificial intelligence systems.

The presence of intergovernmental organization and church type authors denotes that the efforts to devise a regulatory framework transcend geographical borders while connecting different entities, places, and people with a common idea which is then adapted to the particularities of each place.

When analyzing the types into which the documents can be classified according to their purpose, Figure 13.1c shows predominantly those whose objective is to propose principles (35%), and policy principles (30%), which, together with the recommendation-type documents (10%), represent 75% of the corpus.

It should be clarified that the "Action Plan" document type reflects the document titled "AI for Europe" authors' criteria, which presents it as an action plan; however, its scope is limited to recommending principles for the design and use of artificial intelligence and its future implementation into a mechanism in service of designers and consumers. The same occurs for the document typified as "Guidelines Developers".

Regarding the ratio of authors according to their specialization in technical or non-technical backgrounds, we believe it is relevant to recommend, as part of the set of good practices during the documentation future regulatory/standardization documents like the ones being analyzed, including education background and the empirical experience of contributors. In most cases, the analyzed documents lack data related to their authors' training. In other cases, it is stated that prior to their approval that documents were subject to consultation, without going into further details. Only seven documents vaguely declare information related to their authors' background. Based on this information, it can be said that among the documents produced by the private sector and multistakeholders (universities, organizations responsible for methodological standardization, etc.), there is a larger presence of authors with technical training. The opposite occurs for those documents authored by government organizations.

Registering the authors' background information for the documents would certainly help improve our understanding of the difficulties these documents face when they are used to establish a practical tool that designers and consumers of AI can use. We also wanted to stress the need for both regulators (policy makers) and regulated (developers and consumers) to collaborate in establishing a common language that would help the transfers of information and knowledge across domains.

A general analysis of the documents in the corpus is presented in the following.

- Analysis of documents as a whole

A total of 40 documents addressing principled AI with a mean length of 10,246 words were analyzed. In addition, the space devoted to the enunciation and description of the principles proposed in these documents exhibits an average length of 359 words, while for guidelines to help implement these principles, the average length was of 136 words. This represents a relative average length of 0.0350 and 0.0133 units, respectively. That is surprising, given that the portion of the documents dedicated to achieving their goals, which is the proposal of principles and guidelines/recommendations for implementing them, is limited to a significantly small share of the text body.

Differences of relative length for the distinct parts of the body texts evidence that more efforts are dedicated to defining context and justifying the need for principles in most documents than to fully addressing the principles and explaining their implementation. Hence, it can be said that the analyzed documents are more tuned in to providing a description of the problem they hope to solve than to deepening the solution they are supposed to outline.

The corpus was also tested to assess its lexical diversity on each document. This analysis did not yield unforeseen patterns. In general, the documents exhibit low values of the lexical diversity metric, with 0.76 the maximum value achieved, which is also a peak within the dataset. The mean value obtained by the documents in the corpus was 0.38, with a standard deviation of 0.17 units. The pattern that was identified shows that lexical diversity values are inversely proportional to the length of the corresponding document. In that sense, these results could denote a saturated language using certain discursive currents across the corpus, but it could also mean a high specialization in a specific area of argumentation, if possible, aligned with the purpose of the documents, which we feel more inclined to believe after reading the texts.

In order to explore the corpus' content beyond the metadata, we present the analysis of the language used. Table 13.1 shows the ten most used n-grams in the corpus.

As expected, the top five positions of the unigrams frame the context of the texts' argument. Interestingly, the term "human" occupies position seven in the same column; perhaps this demonstrates the human approach that the regulatory framework intended to be defined by these documents. More details on this specific aspect are provided in the following. Another interesting finding to highlight is that the scope of the regulatory framework delimited by the documents in the corpus can be distinguished in positions eight "use", nine "development", and ten "technology". It should be stated that the table presented is an excerpt from a larger analysis, where the unigram terms that interest us, "ethics" and "ethical", are ranked in the 13th and 25th positions, respectively, with relative frequencies of 0.30 and 0.23.

If examined closely, it can be seen in the bigrams column how the unigrams gain context; consequently, the focus on "human" and "use" takes on a new nuance in human rights and the use of artificial intelligence. Another element suggested through the bigrams is that that focus will be influenced by an ethical approach to autonomous or intelligent systems (row 9), with special consideration to personal data (row 10). The authors believe it is worth mentioning that from the extended analysis of the 50 most relevant N-grams, in the bigrams column, terms like "data protection" at position 25 and "trustworthy ai", in position 38, add support to the previous statement about the corpus' ethical approach to artificial and intelligent systems through how they handle personal data.

It can also be noted from the trigrams that the effort to achieve national strategies for an ethically aligned design is orchestrated in the context of artificial intelligence, with support in rows 2, 4, 5, 7, and 8 from Table 13.1. The previous idea gains strength by including trigrams like "ethically aligned design" (row 23), "ethical matters raised" (row 26), "matters raised algorithms" (row 27), and "intelligent systems law" (row 28), as well as the terms "humans keep

Table 13.1 Ten most frequently used n-grams across all documents

	Unigrams	*Absolute freq.*	*Relative freq.*	*Bigrams*	*Absolute freq.*	*Relative freq.*	*Trigrams*	*Absolute freq.*	*Relative freq.*
1	ai	7476	1.94	artificial intelligence	1693	0.44	autonomous intelligent systems	416	0.11
2	data	4156	1.08	ai systems	653	0.17	ieee global initiative	351	0.09
3	systems	2502	0.65	machine learning	603	0.16	ethics autonomous intelligent	318	0.08
4	intelligence	2011	0.52	intelligent systems	446	0.12	global initiative ethics	314	0.08
5	artificial	1798	0.47	autonomous intelligent	429	0.11	initiative ethics autonomous	314	0.08
6	research	1728	0.45	human rights	368	0.10	algorithms artificial intelligence	126	0.03
7	human	1670	0.43	use ai	326	0.08	strategy artificial intelligence	112	0.03
8	use	1595	0.41	computer science	321	0.08	national strategy artificial	111	0.03
9	development	1515	0.39	ethics autonomous	319	0.08	discussion paper national	109	0.03
10	technology	1506	0.39	personal data	298	0.08	paper national design	109	0.03

upper" (row 29), "keep upper hand" (row 29), and "upper hand ethically" (row 30) from the extended analysis.

Having presented an analysis of the terms in which the documents from the corpus are presented and considering the small relative length dedicated to the approach of principles for a trustworthy AI, we now turn to focus the analysis specifically on the principles section and evaluate its correspondence with the rest of the document.

- Analysis of proposed principles

Table 13.2 shows the results of the ten most frequent n-grams in the section dedicated to the declaration and argumentation of the principles proposed by each document in the corpus.

When we delimit the analysis to the proposed principles argumentation section, as can be seen in Table 13.2, the context in which the object of argument is demarcated remains the same, which is in complete consistency with the rest of the sections of the documents, as expected. However, verbs such as "must" in row 4 and "ensure" in row 8 arise that convey some subjectivity to the principles. We resume this idea by presenting an analysis of the verbs used in the principle's descriptions.

Table 13.2 Ten most frequently used n-grams across principles

	Unigrams	*Absolute freq.*	*Relative freq.*	*Bigrams*	*Absolute freq.*	*Relative freq.*	*Trigrams*	*Absolute freq.*	*Relative freq.*
1	ai	532	3.91	ai systems	111	0.82	artificial intelligent systems	11	0.08
2	data	208	1.53	artificial intelligence	59	0.43	aida driven decisions	9	0.07
3	systems	154	1.13	use ai	25	0.18	context consistent state	8	0.06
4	must	131	0.96	ai system	24	0.18	consistent state art	8	0.06
5	use	112	0.82	personal data	18	0.13	ai systems must	8	0.06
6	human	97	0.71	ai technologies	18	0.13	states parties present	8	0.06
7	development	94	0.69	ai must	18	0.13	parties present covenant	8	0.06
8	ensure	92	0.68	ai development	18	0.13	ai system lifecycle	6	0.04
9	government	82	0.60	ai research	17	0.12	appropriate context consistent	6	0.04
10	people	74	0.54	machine learning	16	0.12	present covenant recognize	6	0.04

The bigram "ais must" in row 7 supports the idea mentioned in the previous paragraph on subjectivity associated with the skills that should be attributed to the design and consumption of artificial intelligence systems. Other bigrams strengthen the context of the principles, recognizing their range of action from the academy (row 8), and industry (row 9), as well as the emphasis on personal data (row 5). The column that exhibits the trigrams in Table 13.2 provides no new or relevant information rather than highlighting the intrinsic subjectivity in proposals such as "ai systems must" (row 5) and recognizing the work throughout the life cycle of artificial intelligence systems (row 8).

Interestingly, unigrams such as "rights", "privacy", "security", "transparency", and "fairness" exhibit a relative frequency of 0.40, 0.39, 0.30, 0.26, and 0.25 units respectively, occupying more distant positions within the extended analysis (50 n-grams). They are historically among the terms that describe the ethical dilemmas rooted in the use of artificial intelligence solutions. The same goes for trigrams "standards best practices", "privacy data protection", "equality diversity fairness", "diversity fairness social", and "fairness social justice", with relative frequencies that hold values of 0.04 units for the first and 0.02 for the rest.

At this point, the authors are convinced that, although it appears that there is a common agreement among the involved actors on which issues need work to achieve a trustworthy AI, they address those issues differently. The absence of a common criterion could, among other difficulties, affect the definition of a methodological mechanism for software developers. Pursuing proof of whether such agreement over the fundamental issues, expressed in the form of principles, really exists, and despite the apparent disconnection with their description, we narrowed the analysis to the principle statements.

Table 13.3 shows the most common ten n-grams used in the enunciations of the principles proposed in each document across the corpus. Note that the description of the principles is

Table 13.3 Ten most frequently used n-grams across the principles' declarations

	Unigrams	*Absolute freq.*	*Relative freq.*	*Bigrams*	*Absolute freq.*	*Relative freq.*	*Trigrams*	*Absolute freq.*	*Relative freq.*
1	ai	57	4.29	ai systems	9	0.68	ai systems deployed	3	0.23
2	principle	32	241	artificial intelligence	9	0.68	inclusive growth sustainable	2	0.15
3	privacy	20	150	ensure bot	5	0.38	growth sustainable development	2	0.15
4	data	19	143	non discrimination	4	0.30	sustainable development well	2	0.15
5	transparency	18	1.35	accountability transparency	4	0.30	development well human	2	0.15
6	fairness	16	1.20	u government	4	0.30	values fairness transparency	2	0.15
7	human	16	1.20	principle respect	3	0.23	fairness transparency explainability	2	0.15
8	rights	15	1.13	sustainable development	3	0.23	transparency explainability robustness	2	0.15
9	ensure	15	1.13	transparency explainability	3	0.23	explainability robustness security	2	0.15
10	systems	15	1.13	privacy security	3	0.23	robustness security safety	2	0.15

excluded from the analysis. As can be seen in the table, the n-grams change completely compared to Table 13.2.

The unigrams present in the principle declarations include variables like "privacy", "transparency", and "fairness", while bigrams include "non discrimination", "accountability transparency", "transparency explainability", and "privacy security". Last up, trigrams narrow the idea of inclusive growth and sustainable development with human values, among which "robustness" is added to those already mentioned. It seems that in this case, the triad is more intertwined. In the extended analysis (50 most frequently used n-grams), it can be noted how the variables related to the context and scope of the corpus are pushed to more distant positions on the list in favor of those variables linked to the objectives the proposed principles seek to achieve.

When comparing Tables 13.2 and 13.3, the disconnection between the principle declarations and their descriptions becomes clear. This is a common problem among the analyzed documents, and it could be an element that makes it difficult to build a standardization mechanism that serves as a methodological reference in the design of trustworthy AI. The description of the proposed principles has a more direct link to the documents in general than with its statement, whereas a more logical link between these two parts should be established through the statements of principle in a "justification-enunciation-argumentation" outline.

The general overview of the documents referenced their focus on humans and human rights as the agreed-upon basis for principled AI, while the scope of the documents mainly

encompasses the use and development of technologies from an ethical perspective, with a human rights-based foundation. In general, the documents seek to support the concept of trustworthy AI from variables such as data protection and, all together, to arrive at the conception of an artificial intelligence ethically aligned design.

It is not until the analysis is narrowed to the principle statements that the variables associated with trustworthy AI are extended with terms such as privacy, transparency, fairness, non-discrimination, accountability, explainability, and security. This denotes that the value of principles lies in their statements in addition to pointing to the existing disconnection of the principle statements with their description and with the remaining body text.

In an attempt to arrange a "justification-enunciation-argumentation" outline, the authors point to the need to operationalize the variables involved with the idea behind trustworthy AI. Our hypothesis is that having the network of concepts coexist with the trustworthy AI concept can influence a common understanding of it as the object of study and therefore impact the effectiveness and efficiency of the efforts dedicated to its design. We are motivated by the premise that understanding a problem well makes up 50% of its solution.

A section that also has been reserved a place, although small, in most documents is dedicated to proposing a set of guidelines and recommendations supporting principle implementation. The following section shows an analysis of the guidelines and their alignment with the principles.

- Analysis of guidelines to implement the principles

When considering the guidelines as the set of actions projected to support the implementation of the principles, they are expected to represent concrete actions attached to construction like taxonomies such as applying and creating. In this regard, it is curious that verbs with a presence among the 50 n-grams most frequently used are "consider", with a relative frequency of 0.77 units; "ensure", with 0.47; and "must", with 0.44. It should be clarified that the verbs "use", "design", and "research" are also within this set of n-grams but just exercising an explanatory function, for example, in sentences such as: ". . . operator organizations use . . ."; ". . . ai design . . ."; and ". . . ai research . . .", to give some of examples also present among the bigrams and trigrams identified as the 50 most frequent.

In an effort to expand upon the actions proposed for the implementation of the principles, the authors extracted the verbs and submitted them to a lemmatizing process to consolidate them into a summarized list that helps in better understanding the skills behind the proposed actions. The summarized list consists of 30 verbs. It can be declared that while there are verbs associated with taxonomies such as apply (17.2%) and create (14.8%), which represent approximately one-third of the list, the set of verbs mostly includes actions commonly related to other taxonomies, such as understand (20.6%), analyze (32.1%), and evaluate (15.3%).

That second group of taxonomies encompasses a set of skills that could be perceived as passive skills in a practical context such as software design, particularly when designing artificial intelligence solutions. The authors identify this inclination for passive skills in the language used to describe the proposed actions for the implementation of the principles as another element elevating the difficulties for principle assimilation by AI designers. Consequently, this could prevent the framework from becoming a methodological reference during the AI project lifecycle. At the same time, we understand that this inclination can respond to an interest in maintaining the proposal as a general framework that can then be adapted to each context as needed, safeguarding its global character.

The same is true when performing this analysis on the principles, where there is also an inclination toward verbs representing passive abilities such as the aforementioned. This may be normal given the practical context in which the actions within the corpus are framed. A balance between effectiveness and generality of the proposals based on the cost and benefit linked to the use of certain language remains to be achieved in these types of documents.

- Principle-related topics analysis

Topic extraction is used, in this case, to triangulate some of the observations presented earlier. The 10 most represented topics in the text are listed in the following; it should be clarified that the topics are extracted based on the sections dedicated to principle declaration only:[2]

```
Topic 1.    "system"(0.000) + "ai"(0.000) + "right"(0.000) +
"must"(0.000) + "technology"(0.000) + "shall"(0.000) +
"human"(0.000) + "people"(0.000) + "research"(0.000) +
"decision"(0.000)

Topic 2.    "agency"(0.036) + "assessment"(0.017) +
"even"(0.017) + "system"(0.017) + "accountability"(0.016) +
"obligation"(0.015) + "automate"(0.015) + "assess"(0.014) +
"individual"(0.013) + "decision"(0.013)

Topic 3.    "system"(0.029) + "ai"(0.014) + "value"(0.013)
+ "human"(0.010) + "rapidly"(0.009) + "automation"(0.009)
+ "grow"(0.009) + "power"(0.008) + "people"(0.008) +
"share"(0.008)

Topic 4.    "right"(0.064) + "shall"(0.043) + "law"(0.017) +
"freedom"(0.013) + "education"(0.012) + "protection"(0.011) +
"public"(0.011) + "include"(0.008) + "religion"(0.008) +
"family"(0.008)

Topic 5.    "wide"(0.020) + "solution"(0.017) + "defi-
nition"(0.014) + "seek"(0.012) + "practice"(0.012) +
"way"(0.009) + "dialogue"(0.008) + "algorithm"(0.008) +
"explore"(0.008) + "research"(0.007)

Topic 6.    "remedy"(0.009) + "diversity"(0.004) + "pro-
mote"(0.004) + "inclusion"(0.004) + "effective"(0.003) +
"equality"(0.011) + "non"(0.002) + "discrimination"(0.001) +
"right"(0.000) + "system"(0.000)

Topic 7.    "must"(0.040) + "life"(0.013) + "develop-
ment"(0.013) + "public"(0.012) + "ais"(0.012) + "peo-
ple"(0.012) + "individual"(0.011) + "decision"(0.009) +
"human"(0.009) + "personal"(0.009)

Topic 8.    "constraint"(0.004) + "educate"(0.004) +
"oppose"(0.004) + "maximize"(0.004) + "openness"(0.004) +
"scientist"(0.004) + "listen"(0.004) + "interpretable"(0.004) +
"engineering"(0.004) + "socially"(0.001)
```

```
Topic 9.    "system"(0.028) + "ai"(0.026) + "datum"(0.020) +
"ensure"(0.012) + "human"(0.012) + "technology"(0.010) +
"design"(0.009) + "development"(0.007) + "must"(0.007) +
"people"(0.007)

Topic 10.   "government"(0.039) + "ai"(0.015) +
"public"(0.013) + "policy"(0.012) + "ensure"(0.011) +
"research"(0.011) + "sector"(0.010) + "must"(0.010) +
"recommend"(0.010) + "take"(0.010)
```

Among the topics listed previously, topic 9 is identified as the most relevant, since it is the most representative topic for the principles section for 25 documents, and it represents 62.5% of the documents in the corpus. The representativeness of topic 9 is followed by topic 10, which is dominant in 10% of the documents, and topic 2, dominant in 5%. In contrast, topic 1 does not exhibit any dominance in any of the documents in the corpus. The rest of the topics happened to dominate the distribution of representativeness in one document each.

It is interesting to see what these topics that have become dominant in the great majority of documents reflect: first, the objects of discourse – say "systems", "ai", and "technology", to name examples; second, the action field influenced by these objects – "human", "people", and "decision", among others; and third, the subjective methodological approach that we have already criticized in previous sections, expressed in terms such as "must", "ensure", and "assess", for example.

This supports the idea that there is a clear notion of the problem being addressed with the regulatory framework for AI on the basis of the variables that are affected, but the consensus on the methodological approach to be followed has yet to mature since all author types throughout the period and space covered by the analyzed documents face difficulties in providing a tool that can be used in practice, without intrinsic ambiguities embedded within the passive skills already argued by AI solution designers.

Finally, it is necessary to highlight the vague significance of the term "right" among the principles. The authors hoped that, while the proposed principles were based on an approach that seeks to use Human Rights International Law as a reference to achieve a trustworthy AI, the term would exhibit greater representation or at least a greater representation of the concepts that, in the context at hand – "privacy", "equity", and so on – constitute the term's coexistence network. However, the observed results are not consistent with that estimation.

At the time of this research, no International Standardization Office (ISO) standards were found, nor any published by the Institute of Electrical and Electronics Engineers (IEEE) standards association, although it is known that the latter institution is taking into consideration some of the documents of the corpus in the design of standards related to the subject at hand. Thus, a contrast with the terminology used in the standards could not be established.

Conclusions

The results demonstrate the digital gap between developed and developing countries where there exists an overrepresentation of the former and the need for more representation of the latter in harnessing the ethical and social problems with origins in the design and consumption of artificial intelligence.

The chapter identifies the need for policy makers and designers of AI solutions to join efforts while addressing trustworthy AI as a common goal and the need to record the authors'

background to explore the use of language from a technical and a nontechnical perspective, studying the effects on the resulting guidelines for the assimilation of the principles.

The declaration and description of the proposed principles show a degree of subjectivity from the verbs used. This resulting ambiguity of action was reinforced when analyzing the guidelines for principle implementation, obstructing these guidelines' adoption as a methodological reference for AI design

The pursuit of trustworthy AI through the principled AI framework it is still a process in transformation before it can be properly used as a methodological reference by developers, and it is evident that further intermediate layers of interpretation towards principle adoption are needed as a methodological reference for the design of artificial intelligence solutions.

Notes

1 In English, for example, run, runs, ran, and running are forms of the same lexeme, with run as the lemma by which they are indexed. Lexeme, in this context, refers to the set of all the forms that have the same meaning, and lemma refers to the particular form that is chosen by convention to represent the lexeme.

2 The analysis of the topics for the documents containing the principles, the topics on the basis of the proposed guidelines for the implementation of the principles, and contrasts between them and the topics drawn from the section dedicated to the proposal of principles are exhibited in another publication; in order to maintain the focus of this chapter on the exploration of principles and in correspondence with the communication strategy of the research project.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center for Internet & Society.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*, 90–95. doi:10.1109/MCSE.2007.55

Lizama-Mue, Y., & Suarez, J. L. (s.f.). *Data science methodology for projects in digital humanities and social sciences*. Work in progress.

Loper, S. B., & Edward., E. K. (2009). *Natural language processing with Python*. O'Reilly Media.

McKinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). *A survey on bias and fairness in machine learning*. arXiv:1908.09635v2 [cs.LG]

Mittelstadt, B. (2019). *Principles alone cannot guarantee ethical AI*. Oxford Internet Institute, University of Oxford.

Monroig Vives, R. (2017). *Social networks, political discourse and polarization during the 2017 Catalan elections* (Master Thesis). Western University, London, Canada.

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, *9*(3), 10–20.

Pérez, F., & Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science & Engineering*, *9*, 21–29. doi:10.1109/MCSE.2007.53

Python Software Foundation. (s.f.). *Python. (Python Org) Recuperado el February de 2020, de*. Retrieved from www.python.org

Rehurek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora*. ELRA.

Sahil, V., & Rubin, J. (2018). *Fairness definitions explained*. ACM/IEEE International Workshop on Software Fairness. Gothenburg, Sweden.

Segarra, A. G. (2018). *A data-driven analysis of video game culture and the role of let's plays in YouTube* (Master Thesis). Western University, London, Canada.

Suarez, J. L. (2018, Julio-Agosto). La nacionalización de la estrategia en torno a la inteligencia artificial Estado, política y futuro. *Revista de Occidente*, 5–18, 446–447.

Suarez, J. L., & Lizama-Mue, Y. (2020). Victims of language: Language as a pre-condition of transitional justice in Colombia's peace agreement. En *Transitional justice in comparative perspective, preconditions for success* (pp. 97–127). Palgrave Macmillan.

Trewin, S. (2018). *AI fairness for people with disabilities: Point of view*. Cornell University. arXiv:1811.10670.

Varona, D. (2018, Julio-Agosto). La responsabilidad ética del diseñador de sistemas en inteligencia artificial. *Revista de occidente*, 104–114, 446–447.

Varona, D., Suarez, J. L., & Lizama-Mue, Y. (2020). Machine learning's limitations in avoiding automation of bias. *AI & Society*. https://doi.org/10.1007/s00146-020-00996-y

Virtanen, P., Gommers, R., Oliphan, T. E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Appendix
LIST OF DOCUMENTS INCLUDED IN THE ANALYZED CORPUS

Year	*Author*	*Document Title*
2016	Partnership on AI	Tenets
2016	U.S. National Science and Technology Council	Preparing for the Future of AI
2017	UNI Global Union	Top 10 Principles for Ethical AI
2017	Future of life	Asilomar AI Principles
2017	Tencent Institute	Six Principles of AI
2017	ITI	AI Policy Principles
2017	The French Data Protection Authority (CNIL)	How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence
2018	Council of Europe: European Commission for the Efficiency of Justice CEPEJ	European Ethical Charter on the Use of AI in Judicial Systems and their environment
2018	Amnesty International, AI Now	Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems
2018	T20: Think20	Future of work and education for the digital age
2018	The public voice coalition	Universal Guidelines for AI
2018	Access Now	Human Rights in the age of AI
2018	University of Montreal	Montreal Declaration for responsible AI
2018	Microsoft	Microsoft AI Principles
2018	Google	AI at Google: Our Principles
2018	Telefónica	AI Principles of Telefónica
2018	Microsoft	Responsible bots: 10 guidelines for developers of conversational AI
2018	Standards Administrations of China	White paper on AI Standardization
2018	Mission Assigned by the French Minister	For a Meaningful AI
2018	UK House of Lords	AI in the UK

(*Continued*)

Year	*Author*	*Document Title*
2018	Niti Aayog	National Strategy for AI
2018	British Embassy in Mexico City	Towards an AI Strategy in Mexico: Harnessing the AI Revolution
2018	German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs	AI Strategy
2018	Treasury Board of Canada Secretariat	Responsible Artificial Intelligence in the Government of Canada. Digital Disruption White Paper Series
2019	Organisation for Economic Co-operation and Development OECD	OECD Principles on AI
2019	G20	G20 Principles on AI
2019	IEEE Standard Association	Ethically Aligned Design
2019	New York Times	Seeking Ground Rules for AI
2019	Beijing Academy of AI	Beijing AI Principles
2019	AI Industry Alliance	AI Industry Code of Conduct
2019	Telia Company	Guiding Principles on trusted AI Ethics
2019	IA Latam	Declaration of the Ethical Principles for AI
2019	IBM	IBM Everyday Ethics for AI
2019	Smart Dubai	AI Principles and Ethics
2019	Monetary Authority of Singapore	Principles to promote FEAT AI in the Financial Sector
2019	Government of Japan, Cabinet Office, Council for Science, Technology, and Innovation	Social Principles of Human-Centered AI
2019	European High-Level Expert Group on AI	Ethics Guidelines for Trustworthy-AI
2019	Chinese National Governance Committee for AI	Governance Principles for a New Generation of AI
2019	IEEE Standard Association	IEC White Paper Artificial intelligence across industries
2020	Vatican	Rome Call for AI Ethics
2020	European Commission	AI for Europe