Mapping AI Ethics: Datasheet

Original Datasheet from: 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. Available at https://arxiv.org/abs/1803.09010

LaTeX template from: Christian Garbin, Datasheet for dataset template. Available at https://www.overleaf.com/latex/templates/datasheet-for-dataset-template/jgqyyzyprxth

In this document, we use the words *dataset* and *corpus* interchangeably.

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created because of an absence of a common format for AI ethics charters. It was created initially for textual and structural analysis.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Dataset created by authors from Télécom Paris, Institut Polytechnique de Paris.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

No specific funding was provided to create the dataset. (Name of author removed for peer review) is funded by (Name of the grant removed for peer review) and (Name of the author removed for peer review) is funded by (Name of the institution removed for peer review). Mélanie Gornet is funded by a public research project of the french national research agency (ANR), Simon Delarue is funded by a public institution grant, Maria Boritchev and Tiphaine Viard are funded by Télécom Paris, a french public institution.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are Charters and Manifestos of AI ethics, along with 8 attributes:

- Title: a string, the title of the document;
- Institution: a list, the names of the organisations that issued the document;
- Authors: a list, the names of the persons that wrote the document;
- Sector: a list, categorical variables stating if the document emanates from the industry, academia, national authorities, international organisations, or the civil society;
- Country: a list, the countries of origin of the document;
- Date: a string, format yyyy-mm-dd, the data at which the document was made available (when known);
- Status: a string, "included" or "not included", according to the results of the annotation, corresponding to the inclusion status of the document in the current version of the dataset;
- Label: a string, correspoding to either the reason why the document was not included in the current version of the dataset, or specifying the type of document (ex: "SPI" for "study, policy, or impact assessment") when the document is included in the current version of the dataset;
- Annotator: a string, the first name of the person who annotated the document.

How many instances are there in total (of each type, if appropriate)?

There are 730 instances in total, among which 436 are included in the current version of the dataset (v2.0).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances, following a list of inclusion and exclusion criteria. The dataset is extracted from a larger set of documents discussing AI ethics. The dataset is intended to be representative of all sectors. The representativeness of the dataset is assured by human verification, following a manual annotation process.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Raw data consist of documents related to AI ethics found online and processed into an HTML file, then collected into a JSON file.

Is there a label or target associated with each instance? If so, please provide a description.

The Status and Label attributes can be used and seen as labels/targets.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
The publication dates of the documents are not always available/applicable.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
There is no explicit link between the instances other than the fact that they are all documents that discuss AI ethics. However, several instances can share some attributes like Country, Sector, Date or Institution.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
No specific data split is recommended. In Mapping AI Ethics, we use the entire dataset as the only tasks performed are exploratory analysis and topic modeling, which is an unsupervised method.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
Not that we know of.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
See preprocessing.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.
No instance is confidential as we only selected documents available online. However, some of the original documents are protected by licences; that is why we only release a preprocessed version of the dataset containing a list of words by alphabetical order. It is therefore impossible to reconstruct the complete structured content.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Not that we know of. But the common characteristic of the data set being to talk about ethics, we hope not.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The released dataset does not relate to people directly. However, some people were involved in the creation of the original documents, such as the institutions and the authors; we collected this information when it was public, but do not include it in the final dataset. It is also possible that authors' names are quoted in the content of the document, in which case they will appear in the content field with the rest of the words. But we did not look for them and we do not know which ones are present or not.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
The URL linking to the original documents can be found in the dataset. It is thus possible to trace the authors or institutions by going back to the original documents.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
The dataset does not contain sensitive data.

Any other comments?
None.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
Each of the retrieved documents has a different format and structure, with most of them either in HTML of PDF format. The documents are retrieved in two formats, TXT and PDF. Content is then preprocessed following the procedure described in the "Preprocessing/cleaning/labeling" section. Contents and attributes are compiled in a JSON format.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
After the instances (Charters and Manifestos of AI ethics) are annotated through manual human annotation (status = "included/not included" and corresponding explanatory label), all documents labeled as "included" are automatically scrapped from the corresponding URLs. The code for

the scrapping is publically available, under license, and has been validated through compilation and usage from scratch on different computers by the authors of the paper.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

To list documents to collect, we referred to several existing repositories and meta-analyses. We selected documents from this list using the following criteria:

1. The document must be freely available online to everyone: we do not collect documents that were behind a paywall, or require a subscription.
2. The document must be written in English: we do not collect documents in another language, or unofficial translations;
3. The document must be a final version: we do not collect drafts;
4. The document must discuss artificial intelligence and ethics: we did not collect documents that are too specific, for example focused on facial recognition only, or too general, for example discussing technology and business ethics;
5. The document must be prescriptive : we do not include meta-analysis of charters; we do not include any binding documents, standards, or purely technical documents that take no stance, or any purely descriptive documents.

The complete list of criterions can be found in the paper and in the descriptive flowchart that was used in the annotation process.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Only the authors were involved in the collection process.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The first collection of instances was done between May 2022 and January 2023. The dataset contains documents that are several years old. The inclusion annotation process and the retrieving of content have been done between October 2023 and January 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No ethical review was conducted, but a legal and ethical discussion took place before the creation of the corpus, resulting in design and release choices presented in the current document.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The released dataset does not relate to people directly. However, some people were involved in the creation of the original documents, such as the institutions and authors; we collected this information when it was public, but do not include it in the final dataset. It is also possible that authors' names are quoted in the content of the document, in which case they will appear in the content field with the rest of the words. But we did not look for them and we do not know which ones are present or not.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data was collected via online sources.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Individuals were not notified.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Individuals did not consent.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
No analysis performed.

Any other comments?
None.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
Before releasing the dataset, we preprocessed its content to ensure that the structure of the document cannot be reconstructed. The text was processed using the python libraries `BeautifulSoup` (to manipulate the HTML structure and extract the textual contents) and `NLTK` (for the text itself). The words of the titles have been grouped in a specific field and the words of the whole text in another one. We systematically removed the stop words present in the corresponding NLTK corpus, and put all text in lowercase. We also computed the bigrams and tri-grams (*e.g.* "artificial intelligence" is a bi-gram), adding to our

vocabulary n-grams that appear more than 70 times in our data. Finally, we preprocess all the text through a standard lemmatizer without part-of-speech tagging. For topic modeling, additional thresholds were set on term and document frequency. To see the value of these thresholds and why we chose them, please refer to Section 4 of our paper.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
"Raw" data corresponds to original documents that are available online. They can be found at the URL links provided for each instance.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
The pipeline for preprocessing the data can be found on our website: (Link to the website removed for peer review).

Any other comments?
None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
The dataset has been used for exploratory analysis and topic modeling. These approaches are described in our paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No other papers use our dataset since we release them alongside each other.

What (other) tasks could the dataset be used for?
The datset could be used for many Natural Language Processing (NLP) tasks: word embeddings, clustering, graph analysis, further exploratory analysis based on search for keywords, etc.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
Not that we can think of.

Are there tasks for which the dataset should not be used? If so, please provide a description.
Not that we can think of.

Any other comments?
None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be available online on our website: (Link to the website removed for peer review). As well as on our GitHub: (Link to the github removed for peer review).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
The dataset will be downloadable directly from our website, or available on GitHub. It does not have a DOI.

When will the dataset be distributed?
The dataset will be distributed as soon as the article is published.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The dataset is copyrighted, under licence Creative Commons CC-BY-SA, which lets others remix, tweak, and build upon your work even for commercial purposes, as long as they credit you and licence their new creations under the identical terms. The code is licensed under GNU GPLv3, which allows any uses, modifications and redistributions as long as the code is open-source and with attribution.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

Any other comments?
None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The dataset will be hosted on our institutional website as well as on GitHub, and maintained by the authors.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Contact info can be found on our website or GitHub. Additionally, we provide this information here: (email address removed for peer review)

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
The dataset will be updated by the authors to correct errors or add new instances in later version of the dataset.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
Not applicable.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
All dataset versions will be available on the GitHub. Only the last version of the dataset will be available on our website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
Interested parties can submit a document they deem fit to our criteria and that was preprocessed through our pipeline, so that anyone can contribute to enlarging the dataset. These contributions will be validated by the authors before including them in a new version of the dataset.

Any other comments?
None.