Codebook for the Survey on Modelling Morality for Text Analysis

anonymous
Version 1.0
15.02.2024

Abstract

This document describes the categories and variables used in the Survey on Modelling Morality for Text Analysis.

Contents

1 Introduction

This codebook describes the variables coded in the survey form that we used for our anonymous submission to ACL2024: A Survey on Modelling Morality for Text Analysis. The names in the left margin correspond to the variables defined in the survey form. Reviewers relied on this codebook to ensure consistent reviews for all papers.

2 Survey Categories

2.1 General Information

BibKey — Bibtex key.

Title — The title of the paper.

Author — The authors of the paper.

Year — Publication year.

2.1.1 Paper type

TypeResource — This paper presents new resources for morality detection (e.g., a new annotated corpus or dictionary)

TypeExperiment — This paper presents experiments that aim at detecting/predicting morality in text.

TypeApplicationAI — This paper investigates how morality is represented in LMs / how we can integrate morality in AI applicatiations.

TypeAnalysis — This paper uses NLP methods to analyse morality in text (e.g., studies from political / social science or related fields).

TypeNotRelevant — This paper is not relevant for the survey on morality in NLP (topic not related to morality / NLP).

TypeNotRelevantText — If true, please specify why the paper is not relevant.

2.1.2 Paper length

PaperContentLength — The content length of the paper (paper lenght without references / appendix / supplementary material).

PaperTotalLength — The total length of the paper (including references and appendix, but excluding supplementary materials that have been published as separate documents).

> Consistency check: PaperTotalLength has to be equal to (for extended abstracts) or longer than PaperContentLength.

2.1.3 Overview of the paper content

IncludesDictionary — Paper includes the creation of a dictionary for morality detection in text.

includesOntology — Paper includes the creation of an ontology/vocabulary/other resources in the context of linked data.

IncludesAnnotation — Paper includes the creation of new annotated corpora/resources.

IncludesRuleBased	Paper includes rule-based classification of morality/moral values (e.g., based on dictionaries).
includesLogicBased	Paper includes logic-based approaches to the classification of morality/moral values (e.g., based on PLS, Markov Logic Networks, etc).
IncludesUnsupervised	Paper includes unsupervised or semi-supervised classification of morality/moral values.
IncludesSupervised	Paper includes supervised classification of morality/moral values.
IncludesLLMPrompt	Paper includes prompt-tuning/instruction learning for morality/moral value detection in text (without fine-tuning). **Please note**: Experiments using LLMs *with* fine-tuning are classified as IncludesSupervised.
IncludesProbing	Paper includes probing/analysis of biases in LMs.
IncludesAsFeat	Paper uses moral value prediction to solve some other task (e.g., using moral values as features in another classification task). **Please note**: We do not use this variable when we want to analyse whether one variable can be used to predict a second variable (e.g., whether we can predict retweet counts, based on the moral values addressed in the tweets), but only in cases where the first variable is used as a feature to solve another NLP task (but when we are not interested in learning more about the first variable).
IncludesApplication	Paper applies the detection/prediction of morality for analysis in computational social science / cultural analytics.

2.1.4 Main points of the paper

The next three variables encode the main points of the paper in terms of motivation, contribution and results. Often (but not always) this information can be extracted from the abstract. The inserted text should be no longer than one sentence/bullet point.

PaperMotivationText	The motivation / main purpose of the paper.
PaperContributionText	The main contribution of the paper.
PaperResultsText	The main results of the paper.

2.2 Conceptual modelling of morality

2.2.1 Theory of Morality used in the paper

TheoryMFT	This paper uses the Moral Foundations Theory (MFT) [Graham et al., 2013]. If this applies, then we further encode which version of the MFT has been used (see variables NumMF* below). Most papers additionally add a category *non-moral* which we do not consider in our coding scheme.
NumMF5	Paper uses the 5 moral foundations *Care-Harm, Fairness-Cheating, Loyalty-Betrayal, Purity-Degradation, Authority-Subordination.*
NumMF10	Paper uses the 5 moral foundations *Care-Harm, Fairness-Cheating, Loyalty-Betrayal, Purity-Degradation, Authority-Subordination* are further split into vice/virtue (e.g., the dimension *Care-Harm* is split into the 2 categories *Care* and *Harm*).

NumMF6Liberty — Same categories as NumMF5 but with the additional Moral Foundation *Liberty-Opression*.

NumMF12Liberty — Same categories as NumMF10 but with the additional Moral Foundations *Liberty, Opression*.

NumMF6 — Same categories as NumMF5 but with *Fairness* being further divided into the two new Moral Foundations Equality and Proportionality.

NumMF12 — Same categories as NumMF10 but with *Fairness* being further divided into the two new Moral Foundations Equality and Proportionality.

NumMF7 — Same categories as NumMF5 but with *Fairness* being further divided into the two new Moral Foundations Equality and Proportionality, plus the additional category *Liberty*.

NumMF14 — Same categories as NumMF10 but with *Fairness* being further divided into the two new Moral Foundations Equality and Proportionality, plus the additional categories *Liberty, Opression*.

NumMFOwn — MFT with with additional newly defined Moral Foundations.

NumMFNotSpecified — The specific schema for the encoding of Moral Foundations has not been specified in the paper/unclear what number of MFs has been used.

TheoryHumanval — This paper uses Schwartz' Human Values [Schwartz and Bilsky, 1987].

TheoryOther — This paper uses another theory of morality not listed above.

TheoryOtherText — If TheoryOther is true, then we also encode the name of the theory that has been used in the paper.

TheoryOwn — This paper presents a new/own theory for morality/moral values.

TheoryOwnText — If TheoryOwn is true, then we also encode the name of the new theory used in the paper.

TheoryNone — This paper does not refer to any theory / theoretical background as the basis for modelling morality.

Please note: We do not consider Rules of Thumb (RoT) as a theory as it lacks a theoretical foundation/validation. We therefore code the use of RoT as TheoryNone.

2.2.2 Definition of moral values

DefinitionYes — This paper includes a precise, theory-based definition for the concept of moral values/morality (more than just a reference to the theory).

DefinitionVague — This paper includes a vague description or a reference to some theory.

DefinitionNo — This paper includes no definition of morality/moral values at all.

2.2.3 Level of analysis

The level of analysis used for modelling morality in text:

UnitDocument — Morality is encoded/analysed on the document level.

UnitSegment — Morality is encoded/analysed on the text segment level. As a text segment, we consider subsections of a document or of a sentence (e.g.,

a paragraph or a sentence pair etc.). Tweets and other social media posts/comments/messages are considered to be documents (as they include the whole text/message and not just a subset of it).

UnitSentence — Morality is encoded/analysed on the sentence level.

UnitToken — Morality is encoded/analysed on the token level (e.g., for dictionary-based approaches). This also includes word stems, lemmas, multi-word units and regular expressions.

UnitFrame — Morality is encoded/analysed on the level of entities/semantic frames (e.g., approaches that not only encode the moral foudation/value but also the holder and target of the moral sentiment).

2.2.4 Main purpose for modelling morality

The next variables encode the main purpose or goal of the paper for modelling morality. Here we distinguishh between the following goals:

GoalFraming — The paper aims at investigating (political) framing/moral rhetoric in text.

GoalPerson — The paper aims at analysing the moral values of a person/group/society/culture.

GoalSentiment — The paper aims at analysing the moral sentiment/stance towards a specific target (a person/group/organisation etc.).

GoalComparison — The paper wants to compare moral values to other concepts (e.g., stance, emotions, etc.).

GoalTheoryMoral — The paper wants to test/evaluate/improve a theory on moral values (e.g., Moral Foundations Theory, Schwartz' Human Values).

GoalTheoryOther — The paper wants to test/evaluate/improve another theory not directly related to moral values (e.g., Mediatization Theory).

GoalAI — The (long-term) goal of the paper is to integrate moral values in AI systems/applications. This includes papers that aim at a better understanding how morality is represented in current LMs, what biases exist and how we can remove them.

2.3 Data

The next set of variables encodes general properties of the data used in the paper.

2.3.1 Language(s)

LangEn — The paper works with English data.

LangOther — The paper works with languages other than English.

If this is true, then we also encode which languages have been investigated (variable LangOtherText).

LangOtherText — The language(s) studied in the paper (comma-separated).

LangNotSpecified — The paper does not specify which language(s) are studied.

Please note: In some cases, the language has not been mentioned but can be inferred from context. One example are the TV duells of presidential candidates in the US, as it is clear that these have been in English. However, this does not apply to a collection of tweets that address events that

took place in the US, as many US citizens speak more than one language and many tweets written in the US do not use English. In cases were the paper does not include information about the language of the data but refers to a known Engish dataset, we also use the variable LangEn instead of LangNotSpecified.

2.3.2 Data type

This set of variables encodes the data type that has been used in the paper:

dataSM — The paper studies social media data.

If this is true, then we additionally encode which type of social media has been used:

dataSMTwitter — The paper uses Twitter microtext.

dataSMReddit — The paper uses Reddit data.

dataSMFacebook — The paper uses data from Facebook.

dataSMOther — The paper studies another type of social media data.

dataNews — The paper studies newswire data.

dataOther — The paper uses another type of data.

If this applies, then we additionally encode the data type (variable DataOtherText):

DataOtherText — Specifies which other type of data has been used in the paper.

2.3.3 Topic domain of the data

The next set of variables captures the topical domain of the data. The categories used do not present a well-defined schema of topic domains but only aim at encoding the most frequently used topic domains in the papers.

dataDomainPolitics — The data used in the paper is from the political domain.

dataDomainScience — The data used in the paper is scientific data.

dataDomainLaw — The data used in the paper is from the legal domain.

dataDomainReligion — The data used in the paper is religious data.

dataDomainFiction — The data used in the paper is fictional data (e.g., novels, movie subtitles etc.), usually covering different topical domains.

dataDomainOther — The paper uses data from another (or multiple other) topical domain(s).

If this applies, then we additionally encode the domain(s) of the data (variable dataDomainOtherText).

dataDomainOtherText — Specifies the topical domain(s) of the data used in the paper (comma-separated). If the data comes from one or more benchmark datasets, then we insert the keyword "benchmark" in the text field and retrieve the exact topical domain via the name(s) of the benchmark dataset(s).

2.3.4 Corpora/resources used in the paper

The next set of variables encodes which resources have been used in the data. This includes dictionaries, corpora, benchmarking datasets and analytical frameworks (e.g., SemAxis, MoralDirections, etc.).

Please note: We do *not* include resources that are only used as a baseline (e.g., a dictionary baseline).

RESOURCESDICT The paper uses a dictionary (e.g., the Moral Foundations Dictionary).

RESOURCESDICTVALUES The paper uses the Values Lexicon [Wilson et al., 2018].

RESOURCESMFTC The paper uses the Moral Foundations Twitter Corpus [Hoover et al., 2020].

RESOURCESMFRC The paper uses the Moral Foundations Reddit Corpus [Trager et al., 2022].

RESOURCESMFQUESTIONNAIRE

The paper uses a version of the Moral Foundations Questionnaire [Graham et al., 2011].

RESOURCESSOCIALCHEMISTRY

The paper uses the Social Chemistry 101 dataset [Forbes et al., 2020].

RESOURCESSCRUPLES The paper uses the Scruples dataset [Lourie et al., 2021].

RESOURCESMORALSTORIES The paper uses the Moral Stories dataset [Emelin et al., 2021].

RESOURCESETHICS The paper uses the Ethics dataset [Hendrycks et al., 2021].

RESOURCESMCM The paper uses the Moral Choice Machine (MCM) corpus [Schramowski et al., 2020].

RESOURCESMORALSTRENGTH The paper uses the MoralStrength dataset [Araque et al., 2020].

RESOURCESSEMAXIS The paper uses SemAxis [An et al., 2018].

RESOURCESHHH The paper uses the Helpful, Honest, & Harmless (HHH) dataset [Askell et al., 2021].

RESOURCESTRUSTFULQA The paper uses the Trustul QA dataset [Lin et al., 2022].

RESOURCESOTHER The paper uses some other resource(s).

If this is true, then we additionally encode the name(s) of the used resource(s) (variable RESOURCESOTHERTEXT).

RESOURCESOTHERTEXT Specifies which other resource(s) has/have been used in the paper. Here we not only include resources that focus on moral/human values but also include other resources that have been used in the paper (e.g., RealToxicityPrompts).

2.4 Resource papers

The next set of variables only apply to papers that create/present a new resource for modelling morality in text.

ANNOTSIZETEXT

Specifies the unit and size of the created resurce. We code this variable, using a text field where we insert the number of instances and the unit of analysis, e.g.: "100 documents *or* 20,000 sentences *or* 5788 frames.

We use the following unit definitions specified below:

- Document: the unit of annotation is the whole document (e.g., a news article, a tweet, ...).

- Sentence: the unit of annotation is a sentence.
- Segment: the unit of annotation is a text segment that is part of a larger unit (e.g., a document, a sentence, ...). We also use "Segment" for sentence pairs or Question-Answer pairs etc.
- Token: the unit of annotation is a token or multi-word expression (e.g., for dictionary-based approaches).
- Frame: the unit of annotation is a semantic frame. This is used for annotations that not only capture a moral value but also the participants (such as the holder and target of that value).

2.4.1 Annotation setup

The next variable encodes the annotation setup (for papers that include manual annotation). Possible values are listed below:

annotCrowd	Annotations are done in a crowdsourcing setup.
annotTrained	Annotations are done by trained coders.
annotMixed	Paper includes both, crowdsourcing and annotations by trained coders.
annotNoInfo	Paper does not provide any specific information on the annotation setup.
annotNoAnnot	Paper does not include any annotations / not relevant.

2.4.2 Diversity of annotators

The next variable encodes whether information on the annotators' background has been encoded (only relevant for crowdsourcing setups/setups with many coders). The possible options are:

annotViewsYes	Yes, the paper provides information about the annotators' demographic background and/or moral values.
annotViewsNo	No, the paper does not encode this information.
annotViewsNotRelevant	This question item is not relevant for the paper.

2.4.3 Annotation guidelines / task descriptions

The next variable encodes whether the annotation guidelines / task instructions are available.

AnnotSchemaYes	Yes, the paper provides (a link to) the annotation guidelines or task instructions.

If this applies, then we also encode the length of the guidelines as follows:

AnnotSchemaLen1	Length of annotation guidelines: less than 2 pages.
AnnotSchemaLen2	Length of annotation guidelines: 2 to 3 pages.
AnnotSchemaLen4	Length of annotation guidelines: 4 to 5 pages.
AnnotSchemaLen6	Length of annotation guidelines: 6 to 10 pages.
AnnotSchemaLen10	Length of annotation guidelines: more than 10 pages.

AnnotSchemaNo	The paper does not provide information about the annotation guidelines / task instructions.

AnnotSchemaNotRelevant

This item is not relevant for the paper (e.g., annotation guidelines are not relevant for the created resource).

2.4.4 Inter-Annotator Agreement

The next set of variables encode whether the paper reports Inter-Annotator Agreement (IAA).

IAAYes The paper reports Inter-Annotator Agreement.

If this is true, then we additionally encode the type of annotations, the IAA score and the mesure used for computing IAA.

annotIAATypeText Specifies the type of annotation (i.e., what has been annotated; e.g.: Moral Foundations, MF roles, Schwartz' Human Values, ...). For more than one type of annotation, insert all values separated by commas.

annotIAAScoreText Reports the IAA score. If more than one score is reported, insert all values separated by commas.

annotIAAMetricText Specifies the metric used for computing IAA (e.g., Cohen's κ, Inter-coder correlation, ...). For more than one measure, insert all values separated by commas).

IAANo The paper reports no Inter-Annotator Agreement.

IAANotRelevant This item is not relevant for the paper (e.g., paper does not include human annotations).

2.4.5 Analysis of disagreements

The next variable encodes whether the paper includes an analysis of the disagreements for the human annotations.

AnnotErrAnalysisYes Yes, the paper provides a detailed and informative analysis of the disagreements.

AnnotErrAnalysisRudimentary

The paper provides only a superficial, rudimentary analysis of the disagreements.

AnnotErrAnalysisNo The paper does not include an analysis of the disagreements between coders.

AnnotErrAnalysisNotRelevant

The variable is not relevant for the paper (e.g., paper does not include human annotations).

2.4.6 Availability of the resource

The next variable encodes whether the resource presented in the paper is available for the research community.

AnnotResourceAvailableYes

Yes, the resource is (freely) available.

If this is true, then we also encode the resource URL and provide an optional field for comments.

AnnotResourceAvailableYesURL

The URL where the data can be downloaded.

AnnotResourceAvailableYesComment

Optional field for comments.

AnnotResourceAvailablePartly

The resource is partly available (e.g., the annotations are available but not the data).

If this is true, then we also encode the resource URL and provide an optional field for comments.

AnnotResourceAvailablePartlyURL

The URL where the data can be downloaded.

AnnotResourceAvailablePartlyComment

Optional field for comments.

AnnotResourceAvailableNo

No, the data is not available.

2.5 Papers with experiments

The next set of variables only apply to papers that present experiments on modelling morality in text. This also applies to **resource papers that include baseline experiments** for the new resource. It does not include analysis papers where dictionaries, classifiers or other methods have been used to extract the categories of interest, without any comparison or evaluation of the approaches.

2.5.1 Method / approach

The next variable encodes which method(s) or approach(es) has/have been used in the paper.

Please note: This does not include the baseline experiments. We only encode the main method(s) presented in the paper.

ruleML This paper applies rule-based methods (e.g., dictionaries).

featML This paper applies feature-based machine learning algorithms (e.g., SVM, Naive Bayes, Decision Trees etc.).

logicML This paper applies logic-based methods (e.g., Statistical/Deep Relational Learning, Markov Logic Networks, ...)

expTransformers This paper applies finetuned Transformers.

If this is true, then we additionally encode which model(s) have been used.

expTransformerText List of models used in the experiments (comma-separated).

expReinforcement This paper applies Reinforcement Learning.

expLLM This paper uses LLMs without finetuning (zero-/few-shot; instruction learning etc.).

Please note: Finetuned models are coded as expTransformers.

If this is true, then we additionally encode which approach(es) have been used.

expLLMText Transformer-based model(s) used in the experiments (excluding the baseline systems).

semiML This paper applies semi-supervised ML methods.

If this is true, then we additionally encode which approach(es) have been used.

semiMLText List of approaches used in the experiments (comma-separated; e.g., active learning, weak labelling, etc.).

unsuperML This paper applies unsupervised ML methods.

If this is true, then we additionally encode which approach(es) have been used.

unsuperMLText List of approaches used in the experiments (comma-separated).

expOther This paper applies another method not listed above.

If this is true, then we additionally encode which method(s) have been used.

expOtherText List of methods used in the experiments (comma-separated).

2.5.2 Error analysis

This variable encodes whether the paper presents an error analysis for the experiments.

Please note: We also consider ablation studies as (a type of) error analysis.

ExpErrAnalysisYes The paper provides a detailed and informative error analysis.

ExpErrAnalysisRudimentary

The paper provides a rudimentary error analysis.

ExpErrAnalysisNo The paper does not provide an error analysis.

ExpErrAnalysisNotRelevant

This question item is not relevant.

2.5.3 Replicability of train/test splits

The next set of variables focusses on the replicability of the experiments.

replicTrainTestYes The paper clearly specifies which data points have been used for training/development/testing.

replicTrainTestAmbig The paper includes some information about the data splits but there is some ambiguity.

replicTrainTestNo The paper does not provide information on how the data has been split

into train/dev/test sets.

replicTrainTestNotRelevant

This question item is not relevant.

2.5.4 Replicability of the "ground truth"

The next variable is relevant for datasets that include multiple labels assigned by different annotators (but not in a multilabel setup but where the annotators disagreed / the annotations provide different perspectives on the data). The variable encodes whether the "ground truth" labels that have been used in the experiments can be easily retrieved without ambiguity.

replicGoldClear The paper clearly specifies which labels have been used as ground truth in the experiments.

replicGoldAmbig The paper describes the process of retrieving the gold labels but there is some ambiguity.

replicGoldUnclear The paper does not provide information on how the gold labels have been determined.

replicGoldNotRelevant

This question item is not relevant for this paper.

2.6 Papers with analyses

The next set of variables only apply to papers that present an analysis focussing on research questions in the fields of political / social science, cultural analytics, DH, etc.

2.6.1 Background

The next variable describes the research field / background of the paper.

analysisFieldPolitics The research background of the paper is political/social sciences.

analysisFieldMedia The research background of the paper is media & communication studies.

analysisFieldPsychology The research background of the paper is psychology.

analysisFieldOther The paper has another research background not listed above.

If this applies, then we additionally encode the research background (variable analysisFieldOtherText).

analysisFieldOtherText Specifies the research background of the paper.

2.6.2 Type of analysis

The next variable encodes what type of analysis is presented in the paper.

analysisExplore The paper presents an exploration / visualisation of the data.

analysisRQ The paper formulates and investigates one or more clearly specified research questions.

analysisHypo The paper presents evidence for/against one or more research hypotheses (using significance tests).

analysisNotRelevant The question item is not relevant for this paper.

2.7 Replicability & validation

The next set of variables focusses on the replicability and validation´and applied to all paper types.

2.7.1 Availability of the data

The next variable does not apply to newly created resources like dictionaries, annotated corpora etc. For those we use the the variables AnnotResourceAvailableYes/AnnotResourceAvailablePartly described above (see §2.4). It also does not apply to benchmark datasets/resources that have been used in the experiments presented in the paper (for those, see §2.3.4).

Instead, this variable encodes whether the (raw) data that has been used in the paper (e.g., to conduct an analysis, test a hypothesis etc.) is available.

dataAvailYes The data used in the paper is freely available.

If yes, then we also encode the URL where the data is available and provide a field for comments (e.g., whether a specific license applies etc.)

dataYesUrlText The URL where the data or license can be found/downloaded.

dataYesCommentText A text field for additional comments (e.g., concerning the license, costs or other restrictions).

dataAvailPartly The data used in the paper is available under some license terms.

If this applies, then we also encode the URL where the data can be retrieved and provide a field for comments.

dataPartlyUrlText The URL where the data or license can be found/downloaded.

dataPartlyCommentText A text field for additional comments.

dataAvailNo

The data used in the paper is not available.

dataAvailNoInfo

The paper does not provide any information on the availability of the data.

dataAvailNotRelevant

This question item is not relevant (e.g., no additional data was used).

2.7.2 Preprocessing

The next variable focusses on preprocessing.

replicPreprocClear The paper includes a clear description (or the code) for how the data has been preprocessed.

replicPreprocAmbig The paper includes some information on preprocessing but there is some ambiguity so that it is difficult to replicate the results.

replicPreprocUnclear The paper does not describe what has been done for preprocessing the

data.

replicPreprocNotRelevant

This question item is not relevant.

2.7.3 Availability of the code

replicCodeYes — The code used in the paper is made available.

If yes, then we also encode the URL where the code is available (usually a link to a github repository).

replicCodeText — Specifies the URL where the code can be found.

replicCodeNo — The code used in the paper is made available.

replicCodeNotRelevant — This question item is not relevant for the paper.

2.7.4 Validation

The next variable focusses on the validation of the results presented in the paper.

validationHypothesis — The paper validates the results, using hypothesis testing.

Please note: this does not include the comparison of two or more systems, using significance tests. We only use this variable if the paper clearly states a hypothesis and uses the significance tests to validate it.

When using significance tests to verify that the difference in performance between two systems on a test set is above chance, we use the variable validationAnnotation.

validationAnnotation — The paper validates the results by comparing them to human annotations/a gold standard.

validationCorrelation — The paper validates the results in a correlation study.

validationTriangulation — The paper uses triangulation with some survey data/questionnaires/etc.

validationOther — The paper presents some other method for validation.

If this applies, then we additionally encode which validation method has been applied.

validationOtherText — Specifies the validation approach used in the paper.

validationNone — The paper does not present a validation of the results.

References

[An et al., 2018] An, J., Kwak, H., and Ahn, Y.-Y. (2018). SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.

[Araque et al., 2020] Araque, O., Gatti, L., and Kalimeri, K. (2020). Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:105184.

[Askell et al., 2021] Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., and Kaplan, J. (2021). A general language assistant as a laboratory for alignment.

[Emelin et al., 2021] Emelin, D., Le Bras, R., Hwang, J. D., Forbes, M., and Choi, Y. (2021). Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Forbes et al., 2020] Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., and Choi, Y. (2020). Social chemistry 101: Learning to reason about social and moral norms. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

[Graham et al., 2013] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013). Moral Foundations Theory. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier.

[Graham et al., 2011] Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385.

[Hendrycks et al., 2021] Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. (2021). Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

[Hoover et al., 2020] Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T. E., Chin, J., Leong, C., Leung, J. Y., Mirinjian, A., and Dehghani, M. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

[Lin et al., 2022] Lin, S., Hilton, J., and Evans, O. (2022). Truthfulqa: Measuring how models mimic human falsehoods.

[Lourie et al., 2021] Lourie, N., Bras, R. L., and Choi, Y. (2021). Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13470–13479.

[Schramowski et al., 2020] Schramowski, P., Turan, C., Jentzsch, S., Rothkopf, C., and Kersting, K. (2020). The moral choice machine. *Frontiers in Artificial Intelligence*, 3(36).

[Schwartz and Bilsky, 1987] Schwartz, S. H. and Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*, 53(3):550–562.

[Trager et al., 2022] Trager, J., Ziabari, A. S., Davani, A. M., Golazizian, P., Karimi-Malekabadi, F., Omrani, A., Li, Z., Kennedy, B., Reimer, N. K., Reyes, M., Cheng, K., Wei, M., Merrifield, C., Khosravi, A., Alvarez, E., and Dehghani, M. (2022). The moral foundations reddit corpus.

[Wilson et al., 2018] Wilson, S. R., Shen, Y., and Mihalcea, R. (2018). Building and validating hierarchical lexicons with a case study on personal values. In *Social Informatics: 10th International Conference, SocInfo 2018, St. Petersburg, Russia, September 25-28, 2018, Proceedings, Part I 10*, pages 455–470. Springer.