A Perspectivist Corpus of Numbers in Social Judgements

Marlon May, Lucie Flek, Charles Welch
Conversational AI and Social Analytics (CAISA) Lab, University of Bonn
{maymar, flek, cfwelch}@bit.uni-bonn.de

Abstract

With growing interest in the use of large language models, it is becoming increasingly important to understand whose views they express. These models tend to generate output that conforms to majority opinion and are not representative of diverse views. As a step toward building models that can take differing views into consideration, we build a novel corpus of social judgements. We crowdsourced annotations of a subset of the Commonsense Norm Bank that contained numbers in the situation descriptions and asked annotators to replace the number with a range defined by a start and end value that, in their view, correspond to the given verdict. Our corpus contains unaggregated annotations and annotator demographics. We describe our annotation process for social judgements and will release our dataset to support future work on numerical reasoning and perspectivist approaches to natural language processing.

Keywords: social norms, numerical reasoning, perspectivism

1. Introduction

Language models are increasingly being used in a wide array of applications, from education (Kasneci et al., 2023), to empathic conversation (Ma et al., 2020), to moral reasoning (Jiang et al., 2021b). An underlying assumption of most of these models is that there is a single ground truth or correct answer. This tends to lead to models that only capture the majority and silences minority voices (Fleisig et al., 2023). Prescriptive approaches emphasize the importance of multiple perspectives (Rottger et al., 2022), which coincides with recent work on pluralistic alignment, which has advocated new benchmarks and for models that can express ranges of opinion (Sorensen et al., 2024). They provide an example where, when asked a question, a model responds saying "Many think it's not okay ... while others deem it acceptable." Instead of asserting a single opinion, models can provide pluralistic responses like this, where multiple viewpoints are represented. Similarly, understanding the variation in opinions can aid in tackling the challenging problem of developing models that can express uncertainty (Jiang et al., 2021c; Lin et al., 2022).

Differing perspectives of who acted appropriately can be seen in judgements of conflict situations using Reddit data from previous works (Forbes et al., 2020; Plepi et al., 2022; Welch et al., 2022). These datasets provide valuable insight into what a persons point of view on an issue is, but not about the greater set of (un)acceptable behaviors. In order to get a better picture of these differences, we collected a dataset of social judgement ranges along with annotator demographics. An example is shown in Figure 1. One annotator says that you should not spend any money on jewelry, while the other says you should not spend over 5k. Similarly, one finds it acceptable to purchase 1-3 bus tickets if you desire to sit alone, while the other disagrees, stating that no number makes this behavior appropriate.

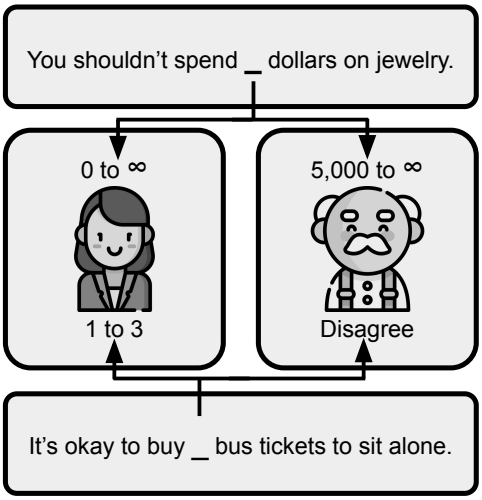

Figure 1: Example annotations of judgements. Each annotator provides a number range for the two questions, unless they disagree with any possible answer.

We extracted statements containing numerical values and asked crowd workers to replace a given number in the statement with a range of values that did not change the judgement. The corpus contains 3k annotations from 30 annotators with different backgrounds and can be used to study different people's perspectives on conflict situations and aid in the construction of models that can communicate varying or pluralistic points of view. Additionally, we believe that this corpus will be valuable for addressing shortcomings in numerical reasoning with language models, especially as it pertains to moral and social judgements (Geva et al., 2020).

NLPerspectives 2024@LREC-COLING 2024, pages 42–48

Filling in numbers allows us to easily extend an existing corpus of judgements from the Commonsense Norm Bank (Jiang et al., 2021a). This corpus contains judgements of social actions and in some cases moral reasoning as well. Moral reasoning about conflicts involves intuition, emotions, and a form of practical reasoning (Bucciarelli et al., 2008; Richardson, 2018). Recent work has defined clearer distinctions between moral and social judgements (i.e. convention), with the latter (e.g. wearing pajamas to school) having less to do with justice, rights, or welfare, and more to do with what is socially acceptable in a given community (Doherty and Kurz, 1996; Turiel, 2002). Both forms of judgement reveal people's beliefs, values, and shed light on their behaviors.

2. Related Work

Recent work in the field of natural language processing has acknowledged that many tasks do not have a single ground truth, including those that have previously been thought to have been objective (Basile et al., 2021, 2020). Not having a single ground truth is viewed as a positive, rather than negative (Aroyo and Welty, 2015). Others have suggested we move toward a data perspectivist approach, where people are encouraged to release unaggregated data and models are built to take multiple different people's perspectives into account instead of prescribing a single answer for any given task (Cabitza et al., 2023). We believe this is the most promising way forward to computationally modeling these judgements. Language models can be conceived in many forms, including that of a search engine (Ziems et al., 2023), in which case we would expect it to provide diverse answers to a question of what is right or wrong (or who acted inappropriately) that reflect a range of human views.

Jiang et al. (2021a) fine-tuned a T5 model (Raffel et al., 2020) for providing moral decisions on the Rainbow dataset (Lourie et al., 2021), a question and answer dataset containing commonsense knowledge. To further fine-tune their model, Delphi, on moral values they created the Commonsense Norm Bank. This is a corpus comprised of four datasets; Social Chemistry (Forbes et al., 2020), the commonsense section of ETHICS (Hendrycks et al., 2021), Moral Stories (Emelin et al., 2021), and the Social Bias Inference Corpus (Sap et al., 2020). The data contains judgements of everyday situations annotated by crowd workers who assigned a label, or verdict, such as "it's okay" or "you shouldn't". They are referred to as commonsense, as they ask crowd workers to use their commonsense judgement, rather than to assign a label based on a particular ethical theory. As shown by Fraser et al. (2022), Delphi inherits the moral values from annotators, which they note as following liberal Western values, neglecting other viewpoints. They note that the model generally follows the positive core principle of utilitarianism by treating the well being of all individuals equally, but does not accept the principle of instrumental harm.

Moral judgement and decision making are separate processes and though the former likely informs the latter, the decision made is affected by dispositional traits and attributes of the dilemma (Nasello et al., 2023). Such situational and personal differences are not taken into account in current models that assign moral judgements. The evolution of what is a social or moral judgement changes over time (Turiel, 2002). Moral judgements are concerned with "justice, rights, and welfare" (Turiel, 1983), while social judgements are about what is socially acceptable. Both tell us about people's values and beliefs as individuals and collectively as a culture.

Using large language models is associated with significant risks and societal harms (Wallach and Allen, 2009). It has been widely suggested that such models should not be used for automated decision making, but that humans should be part of the decision making process (Talat et al., 2021) and that computer scientists should not try to "reinvent ethics from scratch" (Hendrycks et al., 2021). A variety of safety concerns with such models have been identified, such as the *Tay Effect*, or the parroting of harmful information. Moral decision models also suffer from the *Eliza Effect*, where a model may agree with harmful content, e.g. responding "it's okay" to questions of causing harm (Dinan et al., 2021). We do not advocate for the use of any model for automated decision making. Instead we suggest that our corpus could be beneficial for the construction of models that can relay information about the variance in human beliefs rather than definitive judgement. These models could expose people to other points of view and would have a clearer positionality, which would allow for models to be more transparent about where the views they communicate originate from (Santy et al., 2023).

Another area where our corpus may help is with numerical reasoning. There are many ways to represent numbers, with performance varying by task (Thawani et al., 2021). Due to the human understanding of numbers it is likely that a logarithmic scale approach is the best choice for representing numbers in moral statements (Dehaene, 2011). Number ranges that do not change a person's view are informative for understanding the magnitude of an effect or boundaries a person might have and future models could be trained to sample from ranges or to encode the boundaries themselves.

3. Data Selection and Annotation

As foundation for the new dataset, we used the Commonsense Norm Bank (Jiang et al., 2021a). Our goal was to find statements containing numbers and to ask people to replace the number with a range, such that any number in the range satisfies the given judgement. Due to the enormous size of the Commonsense Norm Bank, it covers a large variety of situations, many of which contain numbers. We used spaCy [1] to extract situations containing numbers, but there are three problems with the extracted statements. First, the majority of the sentences only contain *one*, but not in numerical sense. For instance, "The best way to perfect one's talent is to practice often." Therefore, all sentences only containing *one* are removed from the dataset, to minimize the non-modifiable sentences. Second, ordinal numbers are removed, as they often cannot be replaced in a way that changes the judgement of a statement. Finally, numbers which refer to a date or have a special meaning are also not considered, e.g. 911, 24/7, and 50/50. In total, there are 37,746 statements that contain numbers, adhering to the specified criteria.

Although some statements may have more than one modifiable number, we only ask annotators to replace one of the numbers to simplify the annotation process. The following provides an example for the complexity of the interdependence of numbers: "Am I expected to take legal action if someone is doing something that is clearly illegal, in the context of wanting to take legal action because my ex who is 15 is dating a 23 year old man?"

Before they start the survey, annotators are given a detailed description of the task and two examples. They are informed about the study and the possibility to opt-out, and that their results including demographics will be published while maintaining their anonymity. They are instructed that each statement will contain at least one number and to enter the start and end of a range that does not change the judgement. The instructions say that if they disagree with the text label or the number cannot be changed, they should set the start to -1 and end to -1. Otherwise, they should provide a number span. If they think there is no upper bound, they should set the end to -8 (positive infinity). The lower bound should be greater than or equal to zero except when using the special values -1 and -8. As we are dealing with real life situations, the numbers used correspond to the natural numbers.

Annotators were asked to provide their demographic information, including their gender identity, nationality, age, religion, political orientation, and level of education. For gender identity, 46.7% reported male, 53.3% female, and 0% non-binary. The majority of the participants were from the United States, totaling 87%, with 3% each from Georgia, Russia, India, and Germany. The ages ranged from 20 to 58 with a median of 34. Christianity was the highest reported religion at 73%, with 3% Muslim, 7% Hindu, and 17% unaffiliated. The political leaning uses a 5 point scale, with 21% far left, 14% left, 24% central, 10% right, and 31% far right. The level of education included 13% upper-secondary, 57% bachelor's or equivalent, and 30% masters or equivalent. Annotators were recruited using Amazon Mechanical Turk (AMT). Others have noted the skew of demographics of AMT workers and future work would benefit from capturing more diverse perspectives (Difallah et al., 2018). Surveys contained 100 questions each. On average an annotator took 47 minutes to complete the survey. This is roughly 30 seconds per question and 12.75 USD per hour.

We included an attention question in each survey that stated "This is an attention question, so the start should be set to X and the end should be set to Y." Annotators were rejected if they put the incorrect numbers. The surveys were then manually checked for quality, rejecting annotators who put the same answer for most or every question (commonly 0 to X) or who put answers that did not make sense for certain statements (e.g. putting a number for a time of day that is greater than the number of hours in the day).

4. Corpus Statistics

In total, we collected three to six annotations for each of 898 statements from 30 annotators. For 32% of the responses, annotators used infinity as the end point, meaning that any number above the start value aligns with the verdict. For 16% of responses, annotators disagreed with the verdict and did not provide a number range, though only 19 total questions contain no number ranges from any of the assigned annotators for that question. All other instances contained variation in the provided number ranges, indicating the highly subjective nature of the statements.

Only nine statements have entirely disjoint ranges, each of which have at least one non-modifiable answer. For 37 of the questions, at least two annotators gave disjoint ranges, but there is some overlap between the other answers. We calculated the inter annotator agreement between spans by treating each range as a set; disagreement / not modifiable is treated as an empty range. In ranges that end with infinity, we set the end value to the highest value of all finite start and end ranges. We treat ranges as inclusive and take the intersection divided by the union (Jaccard index). We find the average agreement across spans to be 0.39

[1] https://spacy.io/

Situation	Verdict	Annotator 1	Annotator 2
Is Mike lowballing the seller by $**1500** below the car's worth ok?	It's not okay	1,000 to 2,000	1,000 to infinity
Appreciating your youth, in the context of I'm glad I was born in **2000** and not earlier.	It's okay	1900 to 2023	2000 to 2020
Is I buy alcohol for my **25**-year-old little brother a good behavior?	It's ethical	18 to infinity	20 to infinity
In the context of Will's team won, making him **10** bucks richer, must I place friendly bets on sports games?	It's okay	1 to 1,000	0 to infinity
Is playing Russian Roulette okay, if I have played russian roulette on my birthday for the past **3** years?	It's wrong	Disagrees	3 to infinity

Table 1: Examples of situations, verdicts, and annotations from our corpus. Numbers that annotators were asked to replace are shown in bold. Each instance has three to six annotations in total.

with an inter-quartile range of 0.12 to 0.83. This high range is to be expected, as judgements are highly subjective and vary across individuals.

We provide examples from our corpus in Table 1. We see the top two rows pertaining to selling a car and appreciating one's youth. Lowballing the seller of a car could be viewed through a moral lens, though some may consider different ways of negotiating as a social norm. The appreciation of youth in certain years but not others points to personal preferences about the state of the world. The middle example relates to the acceptable age to have alcohol with annotators having slightly different answers. The latter two examples in the table reveal differences in annotator preferences on more controversial issues, namely gambling and suicide.

We also notice that some annotators provide the largest possible range in response to the survey questions, using more X to infinity ranges than other annotators. For these annotators, a number outside of that range would receive a different verdict. For example, one question asks what amount of money is unacceptable to spend on pornography. One annotator indicates that any amount of money is unacceptable, while another provides the range 10 to 250 dollars. This does not imply that spending 251 dollars should be acceptable.

Additionally, it would be beneficial to make distinctions between the types of judgements. We could, for instance, ask annotators if their judgement for a given situation comes from moral reasoning, social norms, or personal preferences. Such annotations would further assist in modeling each independently and making distinctions between moral judgements and other types of judgements (Talat et al., 2021); a distinction recent work does not always make. Though our corpus may support numerical reasoning with number ranges, it would also be interesting to extend this work with fill-in-the-blank style annotations of non-number words.

5. Discussion and Future Work

As the financial resources for the survey were bound to Amazon MTurk, getting more samples with more diverse demographics was not possible. This leads to two limitations of work at hand. First, there is a strong bias in nationality. In future research, this bias could be reduced by using demographic prescreening or a more diverse platform to ensure a representative group of annotators. Second, this work does not have enough examples to provide a solid statistical analysis between judgement and demographics. Further work should consider a representative group of annotators as well as the collection of more annotations per example to support this analysis.

A more costly, but beneficial approach would be to require a justification of the judgement to get a deeper understanding and explanation of the annotators decision. By providing additional context to the scenario, some ambiguities might be eliminated, e.g. specifying the value of the car in the first example from Table 1, but may increase other effects such as the anchoring effect. The context might even change the judgement, as moral situations are often sensitive to small variations, see (Awad et al., 2020) for different scenarios of the trolley problem.

In future work, annotation could be expanded to specify that annotators should determine the type of range either hard or flexible transition and whether there are multiple ranges or only one. A hard range could be the minimum drinking age, where the annotator has a belief about an exact number. A flexible transition, e.g. for lowballing the car seller, would be where the number is approximate, but changing it slightly may not impact the annotators opinion. This could be done by yes-no questions or a textual justification of the range, as the current version does not explain the decision

making process. Questions with the age of people often end with the upper bound of a human lifespan; some annotator answer with ∞ others with 80 or 100. Clearly most of the statements are true for all humans older than X and do not exclude people who are 81 or older.

6. Conclusion

We constructed a corpus of social judgements that asks people to fill in number ranges that do not change a given judgement. Our corpus was crowdsourced from 30 annotators and contains 898 statements for a total of 3k annotations. This work adds to available social judgement data by providing ranges of (un)acceptable behaviors and accompanying annotator demographics. This work supports perspectivist and pluralistic approaches with a goal of creating models that can understand and express multiple points of view, whose point of view it is, and uncertainty about definitive answers. We will publicly release our corpus to promote future work on numerical reasoning, social norms, and perspectivist natural language processing.

7. Ethics Statement

In this paper, we studied different views of moral and social judgements. A potential misinterpretation of this paper's intent would be that we condone the idea of using LLMs to make ethical decisions.

- We do not condone the use of LLMs or any other models to automate moral or ethical decision making.
- We do not condone systems that could deceive a user into believing they are interacting with a human.
- We do not condone systems that in any manner indicate it is a substitute for professional assessment of specific situations requiring ethical consideration.

Having stated this, we believe there may be a place for researching how to create conversational systems that can relay or incorporate diverse human perspectives. LLMs currently present many risks in creating such systems and serious ethical challenges.

Regarding our data collection, participants were informed about the purpose of the study, the nature of their involvement, and their freedom to withdraw at any point. As the Commonsense Norm Bank itself contains offensive material, annotators were warned that the questions can contain offensive content. As discussed in our related work, there are risks associated with the use of LLMs and others have advised against their use in automated decision making (Talat et al., 2021). Additionally, language models trained on huge amounts of data will parrot hegemonic and discriminatory world views (Bender et al., 2021). Fine-tuning a model may alter it's behavior but does not remove these harmful biases, which will surface unpredictably and can even be exploited via adversarial attacks (Zou et al., 2023).

8. Availability

We provide a Hugging Face repository with the dataset.[2] This dataset is available under the CC BY-NC-SA 4.0 licence.[3]

Acknowledgements

This work has been supported by the Federal Ministry of Education and Research of Germany (BMBF) as a part of the Junior AI Scientists program under the reference 01-S20060, and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning. Any opinions, findings, conclusions, or recommendations in this material are those of the authors and do not necessarily reflect the views of the BMBF or Lamarr Institute. We appreciate the anonymous reviewers for their detailed and constructive feedback.

9. Bibliographical References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1).

Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. 2020. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5):2332–2337.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *ArXiv preprint*, abs/2109.04270.

Valerio Basile et al. 2020. It's the end of the gold standard as we know it. on the impact of preaggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

[2] https://huggingface.co/datasets/Marlon154/moral-number-corpus

[3] https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Monica Bucciarelli, Sangeet Khemlani, and Philip N Johnson-Laird. 2008. The psychology of moral reasoning. *Judgment and Decision making*, 3(2):121–139.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Stanislas Dehaene. 2011. *The number sense: how the mind creates mathematics*, rev. and updated ed edition. Oxford university press, New York.

Djellel Eddine Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 135–143. ACM.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *ArXiv preprint*, abs/2107.03451.

Michael E Doherty and Elke M Kurz. 1996. Social judgement theory. *Thinking & Reasoning*, 2(2-3):109–140.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.

Kathleen C. Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? probing delphi's moral philosophy. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 26–42, Seattle, U.S.A.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021a. Can machines learn morality? the delphi experiment. *ArXiv preprint*, abs/2110.07574.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021b. Delphi: Towards machine ethics and norms. *ArXiv preprint*, abs/2110.07574.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021c. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Julian A Nasello, Benoit Dardenne, Michel Hansenne, Adélaïde Blavier, and Jean-Marc Triffaux. 2023. Moral decision-making in trolley problems and variants: how do participants' perspectives, borderline personality traits, and empathy predict choices? *The Journal of Psychology*, pages 1–21.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Henry S. Richardson. 2018. Moral Reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2018 edition. Metaphysics Research Lab, Stanford University.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *ArXiv preprint*, abs/2402.05070.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. A word on machine ethics: A response to jiang et al.(2021). *ArXiv preprint*, abs/2111.04158.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online.

Elliot Turiel. 1983. *The development of social knowledge: Morality and convention*. Cambridge University Press.

Elliot Turiel. 2002. *The culture of morality: Social development, context, and conflict*. Cambridge University Press.

Wendell Wallach and Colin Allen. 2009. *Moral machines: teaching robots right from wrong*. Oxford University Press, Oxford ; New York.

Charles Welch, Joan Plepi, Béla Neuendorf, and Lucie Flek. 2022. Understanding interpersonal conflict types and their impact on perception classification. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 79–88, Abu Dhabi, UAE.

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. Large language models are built-in autoregressive search engines. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2666–2678, Toronto, Canada. Association for Computational Linguistics.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *ArXiv preprint*, abs/2307.15043.

10. Language Resource References

Emelin, Denis and Le Bras, Ronan and Hwang, Jena D. and Forbes, Maxwell and Choi, Yejin. 2021. *Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences*. Association for Computational Linguistics.

Forbes, Maxwell and Hwang, Jena D. and Shwartz, Vered and Sap, Maarten and Choi, Yejin. 2020. *Social Chemistry 101: Learning to Reason about Social and Moral Norms*. Association for Computational Linguistics.

Dan Hendrycks and Collin Burns and Steven Basart and Andrew Critch and Jerry Li and Dawn Song and Jacob Steinhardt. 2021. *Aligning AI With Shared Human Values*. 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.

Nicholas Lourie and Ronan Le Bras and Chandra Bhagavatula and Yejin Choi. 2021. *UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark*. AAAI Press.

Sap, Maarten and Gabriel, Saadia and Qin, Lianhui and Jurafsky, Dan and Smith, Noah A. and Choi, Yejin. 2020. *Social Bias Frames: Reasoning about Social and Power Implications of Language*. Association for Computational Linguistics.