Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties

Taylor Sorensen[1,2], **Liwei Jiang**[1,2], **Jena D. Hwang**[2], **Sydney Levine**[2], **Valentina Pyatkin**[1,2], **Peter West**[1,2], **Nouha Dziri**[2], **Ximing Lu**[1,2], **Kavel Rao**[1], **Chandra Bhagavatula**[2], **Maarten Sap**[3,2], **John Tasioulas**[4], **Yejin Choi**[1,2]

[1]Department of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
[3]Language Technologies Institute, Carnegie Mellon University
[4]Department of Philosophy, University of Oxford
{tsor13,yejin}@cs.washington.edu

Abstract

Human values are crucial to human decision-making. Value pluralism is the view that multiple correct values may be held in tension with one another (e.g., when considering lying to a friend to protect their feelings, how does one balance honesty with friendship?). As statistical learners, AI systems fit to averages by default, washing out these potentially irreducible value conflicts. To improve AI systems to better reflect value pluralism, the first-order challenge is to explore the extent to which AI systems can model pluralistic human values, rights, and duties as well as their interaction.

We introduce ValuePrism, a large-scale dataset of 218k values, rights, and duties connected to 31k human-written situations. ValuePrism's contextualized values are generated by GPT-4 and deemed high-quality by human annotators 91% of the time. We conduct a large-scale study with annotators across diverse social and demographic backgrounds to try to understand whose values are represented.

With ValuePrism, we build Value Kaleidoscope (or Kaleido), an open, light-weight, and structured language-based multi-task model that generates, explains, and assesses the relevance and valence (i.e., support or oppose) of human values, rights, and duties within a specific context. Humans prefer the sets of values output by our system over the teacher GPT-4, finding them more accurate and with broader coverage. In addition, we demonstrate that Kaleido can help explain variability in human decision-making by outputting contrasting values. Finally, we show that Kaleido's representations transfer to other philosophical frameworks and datasets, confirming the benefit of an explicit, modular, and interpretable approach to value pluralism. We hope that our work will serve as a step to making more explicit the implicit values behind human decision-making and to steering AI systems to make decisions that are more in accordance with them.

1 Introduction

When people confront difficult decisions (whether or not to break a promise, what degree program to enroll in, how to spend a Sunday afternoon), their options reflect their values (friendship, knowledge, freedom, saving money, spending time in nature). Two people in the same situation may make opposing decisions if they value different things or the same things but to varying extents (Figure 1). The notion that different human values can lead to distinct—though potentially equally valid—decisions is called *value pluralism* (Páez et al. 2020; Komppula et al. 2018; Brosch and Sander 2013; Keeney 1992; Griffiths 2021; Liscio et al. 2023).

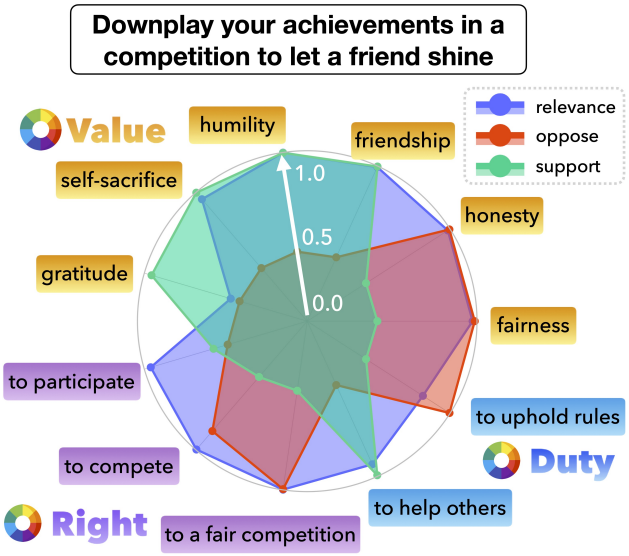

Figure 1: Different human values relate, support, or oppose everyday situations to varying degrees. KALEIDO is designed to generate, explain, and assess how the pluralistic human values, rights, and duties may shape human judgments.

Various fields have focused on this concept. Philosophers distinguish value *pluralism* (different views cannot be reduced into an ultimate "supervalue" (Williams 1985; Larmore 1987; Kekes 1993; Stocker 1990; Chang 1997; Dancy 2004)) from *monism* (there exists a single core value (Kant 1785/2002; Driver 2022)). Sociologists recognize cultural, social, and ideological differences that drive societal clashes, movements, and changes (Archive 2011). Psychologists empirically confirm that ethical experiences involve weighing

pluralistic values (Gill and Nichols 2008) and the dissonance that arises from misaligned values and beliefs (Festinger 1962).

Meanwhile, in AI, there is a growing interest in developing human-centered AI that emphasizes participation from stakeholders. This approach necessitates the inclusion and exploration of pluralistic voices and values (Tasioulas 2022; Gordon et al. 2022). Yet, contemporary supervised AI systems primarily wash out variation by aggregating opinions or preferences with majority votes (Plank 2022; Talat et al. 2022; Casper et al. 2023; Davani, Díaz, and Prabhakaran 2022). As real-world AI applications are used to assist increasing and more diverse audiences, it is crucial to investigate and better model the values that are accessible and used by current AI systems.

In this work, we make the first large-scale attempt at investigating large language models' (LLMs') potential to model *pluralistic human values, rights, and duties*. Our effort is twofold: (1) we introduce VALUEPRISM, a large-scale dataset of pluralistic human values; (2) we build VALUE KALEIDOSCOPE (KALEIDO), an open and flexible value-pluralistic model.

The dataset: VALUEPRISM contains 218k contextualized values, rights, and duties distilled from GPT-4 connected to 31k human-written real-life situations.[1] While GPT-4 and its like have been shown to match human crowdworker annotation performance in some domains (Gilardi, Alizadeh, and Kubli 2023; Ziems et al. 2023; Rytting et al. 2023), we exercise caution and do not assume that GPT-4's outputs are necessarily correct or representative. To this end, we conduct large-scale human studies and find that humans rate the outputs as high-quality 91% of the time and have difficulty coming up with considerations that the model has missed, detecting missing values >1% of the time. We also conduct a comprehensive study with diverse annotators across diverse social and demographic groups to evaluate whose voices are represented in the values GPT-4 produces. Additionally, a growing line of work demonstrates that the large-scale with which data can be produced with LLMs can make up for the potential noise that is introduced, leading to student models which often surpass the teacher (West et al. 2022b; Kim et al. 2023; Jung et al. 2023).

The model: VALUE KALEIDOSCOPE (KALEIDO) is a value-pluralistic model based on VALUEPRISM that *generates*, *explains*, and assesses the *relevance* and *valence* (i.e., support or oppose) of contextualized pluralistic human values, rights, and duties. On top of the model, we build a flexible system KALEIDO^{SYS} leveraging KALEIDO's generation and relevance prediction modes to create a diverse, high quality set of relevant values for a situation (See Fig. 2). In human studies, people rate our system's outputs as more correct and complete than the teacher's (GPT-4). Annotators also find that our largest model matches the teacher's performance at rationalizing and predicting valence. Additionally, we show that KALEIDO can help explain ambiguity and variability underlying human decision-making in nuanced situations by generating contrasting values. We also demonstrate that KALEIDO can be adapted to various philosophical frameworks without explicit training.

Overall, our work represents the first comprehensive attempt to articulate decision-making into fine-grained, pluralistic components of human values employing large language models. The resulting dataset and model[2] serve as a large-scale resource explicitly supporting value pluralism, shedding light on future AI development that accommodates a rich and inclusive tapestry of value alternatives.

2 Value-pluralistic Framework: Values, Rights and Duties

2.1 Why Are Pluralistic Human Values Critical?

Machine learning methods are generally designed to model averages, but can miss nuance and in-group variation unless explicitly accounted for (Gordon et al. 2022; Davani, Díaz, and Prabhakaran 2022). To go beyond this, we take inspiration from philosophical value pluralism, the stance that there are many different normative values (Mason 2006), as opposed to one super-value that all other values can be reduced to. This is distinct both from political pluralism, which posits that diversity is beneficial to democratic society and supports the distribution of power among diverse groups (Britannica Editors 2002; Martí 2017; Landemore 2013); and from relativism, which holds that no moral system is more correct than another (Gowans 2021).

Without taking a hard stance on these positions, we seek to better model humans' plural values to make explicit the implicit values in human decision-making. Our hope is that, if pluralistic values can be adequately (though imperfectly) modeled, we can take a step towards ensuring that automated decision-makers act in accordance with them.

2.2 Framework Motivation and Definition

In this work, we model human-centered plural values to make explicit implicit values in human decision-making. We settle on *values* (Mason 2006), *rights* (Prabhakaran et al. 2022; Wenar 2023), and *duties* (Alexander and Moore 2021) as our three core concepts. We propose a commonsense framework for reasoning about them, and outline it below.

Values: These are the *intrinsic goods or ideals* that people pursue or cherish, such as happiness, well-being, justice, or freedom. Values are the desirable qualities that people may seek in their lives and in the world. They are often the guiding principles for individuals and societies, shaping goals, motivations, and preferences.

Duties: Duties are the *moral obligations or responsibilities* that individuals owe to others or to society at large. They are categorical reasons for doing or refraining from doing something, independent of whether we want to do or refrain from doing that thing. Duties can be weighty reasons, not easily overridden by competing concerns, and their violation

[1] Datasheet for Datasets (Gebru et al. 2018) documentation in App. N.

[2] Dataset: https://huggingface.co/datasets/allenai/ValuePrism
Model(s): https://huggingface.co/allenai/kaleido-xl (5 model sizes)
Code: https://github.com/tsor13/kaleido
Demo: https://kaleido.allen.ai/

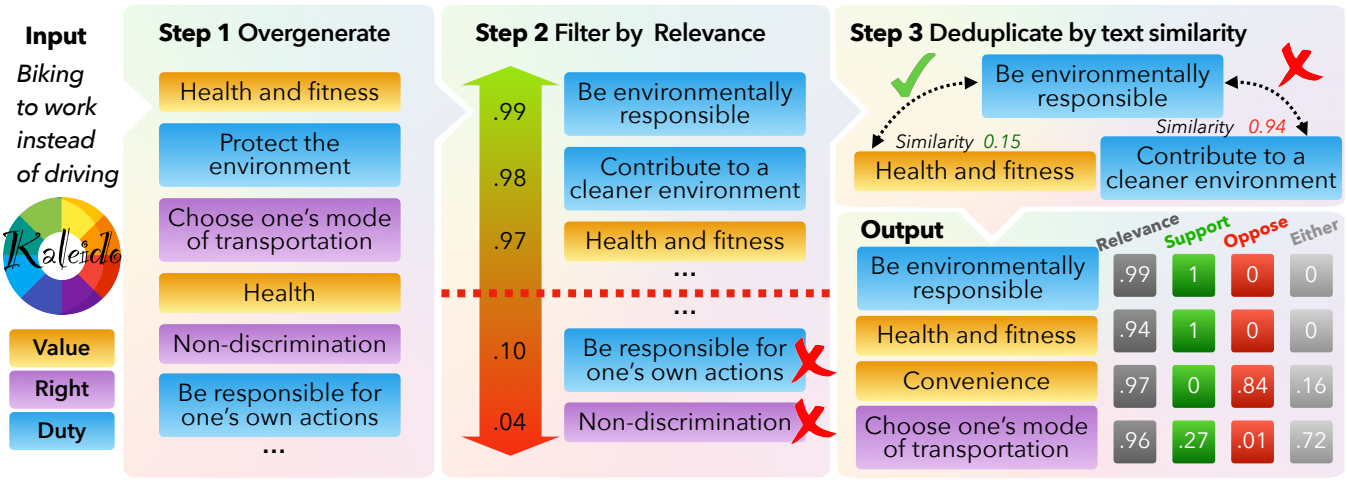

Figure 2: KALEIDO^{SYS} system workflow that includes 1) generating 100 values, rights and duties; 2) filtering by relevance as rated by KALEIDO; 3) removing repetitive items; and computing relevance and valence scores for each value, right, and duty.

may justify blame and self-blame (guilt). Duties can arise from relationships, social roles, or moral principles, and they guide our actions and decisions.

Rights: Rights are the *entitlements or claims* that individuals have against others or society, which are usually based on moral or legal grounds. These can be positive rights (e.g., the right to education, healthcare, or free speech) or negative rights (e.g., the right to not be harmed, enslaved, or discriminated against). Rights serve to protect the fundamental interests of individuals and establish certain boundaries that others must respect.

3 KALEIDO: Value-pluralistic Modeling

We introduce KALEIDO, a language-based multi-task system that *generates*, *explains*, and assesses the *relevance* and *valence* (i.e., support or oppose) of pluralistic human values, rights, and duties, grounded in real-world contexts.

3.1 Tasks

We develop four tasks for modeling values, rights, and duties, all grounded in a given context situation.

Generation (open-text) *What values, rights, and duties are relevant for a situation?* Generate a value, right, or duty that could be considered when reasoning about the action.

Relevance (2-way classification) *Is a value relevant for a situation?* Some values are more relevant than others.

Valence (3-way classification) *Does the value support or oppose the action, or might it depend on context?* Disentangling the valence is critical for understanding how plural considerations may interact with a decision.

Explanation (open-text) *How does the value relate to the action?* Generate a post-hoc rationale for why a value consideration may relate to a situation.

The generation task depends only on a situation while the other tasks evaluate a given value, right, or duty w.r.t. a situation. For examples of each task, see Table 1 and App. A.2.[3]

[3]Appendix may be referenced in the arxiv version: https://arxiv.org/abs/2309.00779

Situation: Telling a lie to protect a friend's feelings		
Task	**Input**	**Output**
Generation	{situation}	Value: Honesty
Generation	{s}	Value: Friend's well-being
Relevance	{s}, Value: Honesty	Yes
Relevance	{s}, Value: Economic well-being	No
Valence	{s}, Value: Honesty	Opposes
Valence	{s}, Value: Friend's well-being	Supports
Explanation	{s}, Value: Honesty	If you value honesty, it may be better to tell the truth even if it hurts feelings.

Table 1: Illustrative examples of each task, with {situation}/{s} standing in for the example situation.

3.2 Dataset: VALUEPRISM

We leverage the symbolic knowledge distillation (West et al. 2022a) pipeline to distill high-quality knowledge from powerful generative models like GPT-4, which have been shown to compare favorably to human annotations on quality, coverage, and diversity (West et al. 2022a; Gilardi, Alizadeh, and Kubli 2023; Ziems et al. 2023). Importantly, based on our preliminary exploration, GPT-4 excels at enumerating a *wide* range of value alternatives compared to average human annotations.

We verify the dataset's quality with human annotators and show that 91% of the distilled data is deemed high quality,

surpassing typical quality of human generated data (West et al. 2022a; Hwang et al. 2021; Zhou et al. 2023). Details on dataset statistics and splits are provided in App. F.1 and examples from VALUEPRISM can be found in App. A.

Situations We obtain a set of 31k situations for deriving pluralistic considerations by carefully filtering out ill-formatted, irrelevant, and low-quality instances from a set of 1.3M human-written base situations. [4] To balance out an outsize proportion of toxic, NSFW, and sexually explicit content, we down-sample these situations to 5% of all data, leading to an increase in the overall diversity of the dataset, as measured by the normalized count of unique n-grams (dist-2: .23→.36, dist-3: .54→.67, details in App. F.1). We filter using a Flan-T5 (Chung et al. 2022) few-shot classifier.

Values, Rights, and Duties Generation For each of the 31K situations, we prompt GPT-4 to generate a batch of relevant values, rights, and duties (Table 2) with open-text rationales. GPT-4 also attributes whether the corresponding value, right, or duty supports (justifies), opposes (condemns), or whether the valence might depend on the context or interpretation. Details of data generation and prompting are in Appendices F.1 and M. The resulting dataset is rated as high-quality by 3/3 annotators 91% of the time (§4.1).

Type	Total	Unique	Avg. Per Situation
Situations	31.0k	31.0k	1
Values	97.7k	4.2k	3.15
Rights	49.0k	4.6k	1.58
Duties	71.6k	12.8k	2.31

Table 2: VALUEPRISM Dataset Statistics. The total, unique, and average per situation statistics of generated values, rights, and duties are shown.

Multi-task Setup We convert VALUEPRISM into a sequence-to-sequence format for multi-task training (Table 1). The relevance task data is sampled contrastively, where positive examples are the values generated for the situation and negative examples are drawn from values mapped to other situations, while the data for the remaining tasks are taken directly from the GPT-4 generated outputs. We split the data (by situation) into 80% train/10% validation/10% test (See Table 3).

	Relev.	Valence	Gen.	Expl.	Mixture
Train	349k	175k	175k	175k	874k
Val	44k	22k	22k	22k	109k
Test	44k	22k	22k	22k	109k
Total	437k	219k	219k	219k	1.1M

Table 3: VALUEPRISM subtask statistics by split. Mixture is composed of the relevanec, valence, generation, and explanation tasks combined.

[4] Sourced from user queries submitted to the Delphi demo.

Event	**Driving at the maximum speed of the road**	
Value	Efficiency	support
	Safety	oppose
	Rule-following	oppose
	Law-abiding behavior	oppose
Right	Right to freedom of movement	either
	Right to safety	oppose
	Right to life and physical integrity	oppose
Duty	Duty to follow traffic laws	oppose
Event	**Gambling with my boss**	
Value	Professionalism	oppose
	Loyalty	oppose
	Personal enjoyment	support
	Financial well-being	oppose
Right	Right to work in a safe and professional environment	oppose
	Autonomy	either
Duty	Duty to maintain a professional relationship with your boss	oppose
	Duty to uphold company policies	oppose
Event	**someone cuts me off. i honk ten times**	
Value	Courtesy	oppose
	Respect for others	oppose
	Safety	support
	Harmony	oppose
Right	Right to peaceful enjoyment of the road	oppose
	Right to not be subjected to harassment	oppose
	Right to safety	oppose
Duty	Duty to express displeasure	support
	Duty to be a considerate driver	oppose

Table 4: Example outputs from KALEIDO^{SYS}-3B.

3.3 Model Training

We use Flan-T5 3B (Chung et al. 2022) as the base model for KALEIDO. After fitting hyperparameters with a grid search, we train on a mixture of all four tasks with Huggingface's Trainer (Wolf et al. 2020) for 4 epochs with early stopping. Training takes 19 hours on two A100 GPUs. Training details are in App. G.

3.4 A System of Diverse Values: KALEIDO^{SYS}

We use KALEIDO to generate a diverse set of values, rights, and duties by overgenerating (top 100 beams) and removing low-quality and repetitive outputs via the relevance score and text similarity respectively. We use Rouge-score (Lin 2004) for n-gram similarity and a Transformers (Wolf et al. 2020) mpnet model[5] for sentence embeddings. See Fig. 2

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

for an illustration of the system and App. H/Algorithm 1 for more details. We tune the system parameters (relevance score threshold, similarity thresholds) using Gibbs sampling (Casella and George 1992) to maximize RougeL-Sum F1 score on the validation set. Ablation experiments in §5.1 provide insights on each system component, and example system outputs can be found in Table 4 and App. B.

4 Data Analysis

4.1 VALUEPRISM Is High-Quality

We conduct human validation on a subset (10%) of VALUEPRISM to assess its quality on the Mechanical Turk platform[6]. Given the generated situation and values, rights, and duties and their explanations, we ask the annotators to assess the relevance and quality of the generations. The results show that annotators find the great majority of the data as high quality. 91% of the values/rights/duties were marked as good by all three annotators and 87% of the valences were marked as correct by all three annotators.

In an attempt to find any values that may have been missed, we also prompt crowdworkers to fill in any missing values, rights, or duties. Crowdworkers did not seem to find it easy to come up with missing values as we get suggestions 0.35% of the time. Full annotation details for this and other studies are in App. I.

4.2 Evaluation by Diverse Annotators

Prior research has reported unjust biases in LLMs against marginalized groups (Sap et al. 2019; Feng et al. 2023). We evaluate VALUEPRISM by recruiting a diverse population of 613 annotators[7] through CloudResearch (Litman, Robinson, and Abberbock 2017) targeting those marginalized groups to the extent possible.[8] We collect 31k annotations across 683 values, rights, and duties in the context of 100 situations, along with demographic information across eight categories. The annotators mark 1) if they agree with each value, right, or duty listed for a given situation and 2) if they spot any missing perspective. We do not find notable statistical significance, and do not reject the null hypothesis that there is no difference between groups. Additional group statistics, p-values, and qualitative analyses are in App. E.

4.3 Diversity of VALUEPRISM

We analyze the diversity of the situations, and values, rights, and duties from three perspectives: *lexical diversity* that calculates unique n-grams, *topical diversity* that assesses semantic diversity via topic analysis[9], and *clustering*. Both the situations and the values cover diverse and distinct concepts with high lexical variations indicating a diverse variety of events and values captured by VALUEPRISM (Table 9). The topic word cloud (Fig. 8) shows that VALUEPRISM covers a broad spectrum of common topics like "save", "kill", and "helping" for situations and "respect", "care", and "promote" for values. Clusturing shows that the corpus encompasses a wide variety of themes, reflecting the diversity and richness of situations and values, rights, and duties. For more data analysis, see App. C.

5 Experiments

5.1 Our System Against the Teacher

Generating correct and complete sets of values Central to our research is the capability to model pluralistic values, rights, and duties. Ideally, these values should be correct, have high coverage, and be aligned with human preferences. We recruit crowdworkers to evaluate KALEIDO^{SYS} directly against GPT-4 across these three dimensions.

We run several variations of KALEIDO^{SYS}: all five model sizes (60M–11B); 3B version without the relevance or text similarity components (*-relevance*, *-text similarity*); and 3B with modified system parameters to output more or fewer values, rights, and duties[10] (*verbose*, *concise*). To understand the added benefit of the system, we also train a baseline seq2seq 3B model on the same data that predicts a batch of values, rights, and duties in one generation pass, as opposed to generating 100 candidates with beam search and filtering down with the relevance/deduplication components as in KALEIDO^{SYS}. We test each version against GPT-4 on a set of 200 test situations, evaluated by 2 annotators each.

From Table 5, we make several observations. The three largest versions of our system outperform GPT-4 on all evaluated dimensions, with the largest variant (11B) being the most favored overall. Moreover, the models generating a higher number of values ($>$11) are preferred by humans for coverage and accuracy.[11] KALEIDO^{SYS} also shows an advantage over the direct output seq2seq model trained on the same data, demonstrating the added benefit of our inference system. Furthermore, removing relevance leads to a drop in the overall preference, which is not observed in *verbose* with the same number of outputs. This suggests relevance is indeed a contributing factor to the generation quality. Finally, humans show lower preference for outputs without deduplication with text similarity.

While it may seem unintuitive that our student model surpasses the teacher, we suspect a few possible explanations for this: student models are still of significant size, able to generalize from the large distilled dataset to become a strong specialist; and the relevance score serves as a critic, improving performance. Additionally, there is a growing body of re-

[6]For this and other human studies, we have acquired the opinion of our institutions's Internal Review Board. The opinion finds our project exempt from a full review process and we have acquired a letter of exception. We hash crowdworker IDs so annotations cannot be back-traced to individual workers.

[7]E.g., Race: 168 white, 115 Black, 61 asian, 34 hispanic/latinx; Sexual orientation: 390 straight, 68 LGBQ+. Gender: 258 male, 201 female, 9 non-binary or other; Full details are in App. E.

[8]We chose CloudResearch specifically because of its ability to target by demographic. One limitation of this study, however, is that all of our respondents are U.S.-based (where CloudResearch operates). Prior work has shown that value representation can vary across nationality as well (Santy et al. 2023), and we hope to extend this study internationally in the future.

[9]Via BERTopic https://maartengr.github.io/BERTopic

[10]To better understand how changing the parameters can affect the output/precision/recall, see Figure 4.

[11]This is in line with prior work showing that humans prefer longer outputs with more unique n-grams (Wang et al. 2023b)

Model	Overall	Cover.	Acc.	Avg. #
KALEIDO^{SYS} 3B	55.5	65.1	58.9	8.2
-relevance	51.9	81.4	64.3	11.2
-text similarity	50.0	60.5	52.9	8.2
verbose	**58.0**	**86.1**	**69.0**	11.1
concise	39.0	27.4	32.4	5.0
KAL^{SYS} 11B	**58.3**	71.1	62.5	8.3
KAL^{SYS} 770M	**57.9**	67.3	60.8	8.2
KAL^{SYS} 220M	44.9	59.0	50.8	8.1
KAL^{SYS} 60M	32.0	53.0	37.1	8.5
Direct Output	42.5	37.9	40.0	6.8
GPT-4	50.0	50.0	50.0	7.0
GPT-3.5-turbo	39.5	49.0	39.8	8.0

Table 5: The overall, coverage and accuracy win rate percentage against GPT-4 by human evaluators along with the average number of generated values, rights, and duties. (Here and throughout, best results within 1% are bolded.)

Model	Explanation	Valence	Rel. corr.
KALEIDO 3B	92.6	92.0	**0.30**
KAL 11B	**94.8**	**92.6**	0.25
KAL 770M	90.3	90.3	**0.31**
KAL 220M	86.9	86.3	**0.30**
KAL 60M	75.9	72.3	0.28
GPT-4	**94.7**	**93.1**	-

Table 6: Human Evaluation. Explanation and Valence scores are correctness rates of the output, while Relevance is the correlation of relevance score with the percentage of people who marked a value as relevant.

cent work where specialized student models surpass teacher models (Hsieh et al. 2023; West et al. 2023; Jung et al. 2023).

Explanation and Valence Label Quality We also evaluated the explanation generation and valence labeling abilities of each model using 700 values, rights, and duties from the test split of VALUEPRISM. Crowdworkers were tasked with evaluating the quality of explanations, their effectiveness in linking values to actions, and agreement with valence labels. As depicted in Table 6, the 11B model's performance closely aligns with that of GPT-4. The 11B model achieved Valence accuracy within a 1% difference from GPT-4 and slightly outperformed it in terms of Explanation quality.

5.2 Relevance Correlates with Human Judgments

We would like KALEIDO to predict whether a human would find a value, right, or duty relevant. However, its training data is synthetic, so the model's training objective is in fact closer to predicting whether a given value was likely to be generated for a particular situation by GPT-4. To test how well this proxy objective correlates with how humans judge relevance, we collect 18 relevance annotations each for 700 values/rights/duties and correlate the relevance score (token probability of "relevant" vs. "irrelevant") with the percentage of people who marked the value as relevant (See Table 6). We find correlations of 0.25-0.31 for the suite of model sizes[12] (all significant at $p < 10^{-10}$). Although we would like to explicitly train models to predict human relevance scores in future work, we take this as evidence that our synthetic relevance prediction task correlates positively with human judgments.

5.3 Zero-Shot Performance on ETHICS

While our model is explicitly trained to recognize values, rights, and duties, we want to understand how much the learned representations generalize to other frameworks as well. To do this, we test KALEIDO on the ETHICS benchmark (Hendrycks et al. 2023), which contains crowdsourced ethical judgments across several different frameworks. We design templates (prompts) in our values/rights/duties task setup that loosely correspond to the frameworks (see Appendix L) and test them in a zero-shot manner.

Subset	KALEIDO	ChatGPT	Random
Justice	**17.5 / 13.3**	**17.6 / 13.4**	6.3 / 6.3
Deont.	**19.8 / 15.1**	**20.6** / 13.8	6.3 / 6.3
Virtue	**33.1 / 22.2**	24.9 / **22.0**	8.2 / 8.2
Util.	**76.5 / 66.6**	59.4 / 55.1	50.0 / 50.0
Comm.	71.5 / 64.7	**80.3 / 68.8**	50.0 / 50.0
Average	**43.7 / 36.4**	40.6 / 34.6	24.2 / 24.2

Table 7: ETHICS few-shot performance. First/second number of each entry is performance on the test/hard test sets respectively. KALEIDO is zero-shot, ChatGPT is few-shot.

Results are in Table 7. On all five tasks, our model performs well over the random baseline. On all tasks but Commonsense, our model matches or exceeds (Justice, Deont., Virtue, Util.) ChatGPT's performance, while only having 3B parameters. Despite having only been trained to predict values, rights, and duties, our model meaningfully generalizes to other frameworks.

5.4 Interpretable Decision System and Zero-Shot On COMMONSENSENORMBANK

While the focus of the system is on modeling diverse values and not on making judgments, it can be easily extended to output the valence of an action $V(a)$:

$$V(a) = \sum_{v \in VRD} R(v|a) \times V(v|a)$$

where $v \in VRD$ are the generated values, rights, and duties from KALEIDO^{SYS}, $R(v|a)$ is the relevance of v given the

[12]Interestingly, we note that the correlation does not strictly improve with model size. While we are unsure of the reason for this, we note that 11B gives much more confident relevance scores, and hypothesize that this overconfidence may be miscalibrated to human judgments.

action, and $V(v|a)$ is the valence of v given the action. We will denote this decision system KALEIDO$^{\text{DEC}}$.

This system has the advantage of being interpretable, enabling direct inspection of how values linearly contribute to the outcome. It is also steerable, as users can easily assign a weight of zero to values they do not wish to take into consideration.

Zero-shot COMMONSENSENORMBANK performance We evaluate this system in a zero-shot manner on the four subportions of moral acceptability segment of COMMONSENSENORMBANK (Jiang et al. 2022) (results in Table 8). In all cases, the system performs at least as well as the majority class baseline, and much ($\geq$25%) better on ETHICS and Moral Stories.[13]

We observe that the model predictions are not well calibrated to the dataset statistics. To remedy this calibration issue, we fit a lightweight logistic regression on the model predictions. For SBIC and SocialChem it improves accuracy by about 5% and 15% respectively, suggesting that while the model is not initially well-calibrated to the datasets, relevant information can be linearly extracted. While KALEIDO$^{\text{DEC}}$ achieves non-trivial zero-shot performance, it unsurprisingly performs worse than supervised baselines such as Delphi.

Model	SBIC	ETH.	MoSt	SoCh
KALEIDO$^{\text{DEC}}$	64.4	77.9	75.4	48.2
+label calibration	69.3	78.0	76.2	63.0
(improvement)	(+4.9)	(+0.1)	(+0.8)	(+14.8)
Majority class	63.1	51.6	50.0	46.7
Random	33.3	50.0	50.0	33.3
Delphi (SFT)	82.9	86.2	86.5	78.0

Table 8: Zero-shot Performance on COMMONSENSENORMBANK: Moral Acceptability.

5.5 Entropy as an Indicator of Decision Variability

When values support different decisions, it may be an indicator that the final judgment one may come to is highly dependent on which value is prioritized. Because of this, when KALEIDO$^{\text{DEC}}$ output has high entropy, we hypothesize that this may indicate higher variability in the distribution of decisions. To test this, we explore two datasets with variability indicators. MORALCHOICE (Scherrer et al. 2023) contains 687 low-ambiguity and 680 high-ambiguity moral scenarios. SOCIALCHEM (Forbes et al. 2021) is a corpus of social norms where, among other things, crowdworkers annotated for "What portion of people probably agree that [action] is [good / bad]?". We take those marked as $\geq$ 99% to have low controversialness, and those marked as $\leq$50% as having high controversialness. We run the corresponding scenarios through KALEIDO$^{\text{DEC}}$ and measure the entropy (Figure 3). We find that the entropy is predictive of these classes. In line with our hypothesis, the higher the entropy, the more likely a situation is to be ambiguous or controversial, even though the model was not explicitly trained to predict these features.

[13]For these two datasets, there is no "neutral" (i.e., lacks valence) class, so the "either" valence is zeroed out.

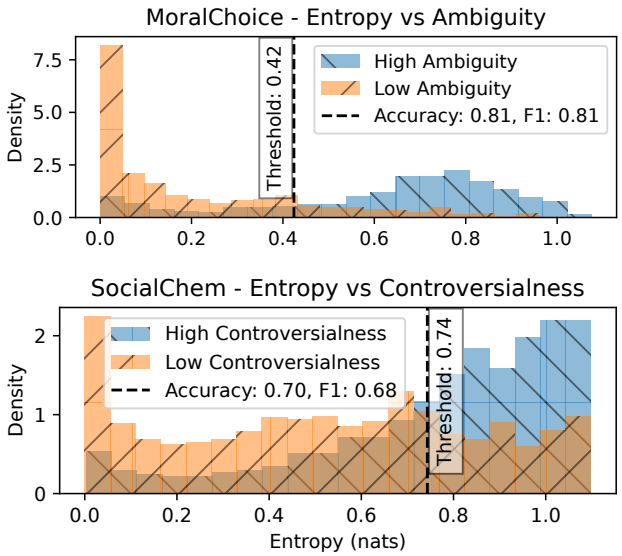

Figure 3: The output entropy of KALEIDO$^{\text{DEC}}$ is predictive of ambiguity in MoralChoice and controversialness in SocialChem. A threshold is chosen to maximize F1-score.

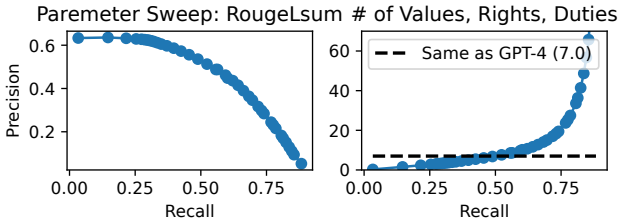

Figure 4: By sweeping KALEIDO$^{\text{SYS}}$'s parameters, we are able to trade precision for recall (w.r.t. to the GPT-4 generated test split of VALUEPRISM) and output many more (or fewer) values, rights, and duties.

6 Discussion

Strengths Over Teacher Model Although our model performs strongly against the teacher in value generation, it also has several other advantages. It is controllable, allowing users to generate either more or fewer values than GPT-4 by trading precision for recall (see Figure 4). Additionally, while GPT-4 provides only textual labels for valence, our model generates scalar valence and relevance scores (probabilities of the corresponding tokens). Lastly, our model, dataset, and code are openly accessible, enabling scientific review that is crucial for accountability and improvement.

KALEIDO is Sensitive to Contextual Variations One of the strengths of our approach is that the signal can be conditioned on variations in a situation, leading to changes in values' relevance and valence. For example, consider three vari-

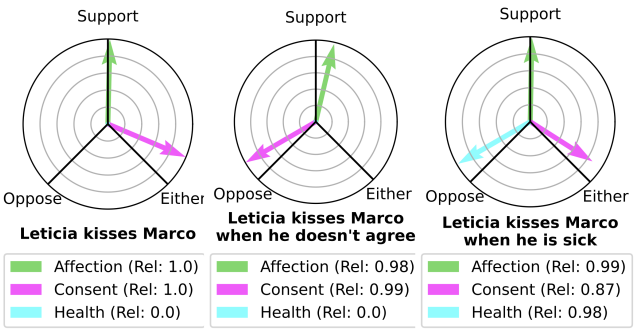

Figure 5: KALEIDO is sensitive to subtle changes in inputs, changing relevance and valence scores accordingly.

ations of a situation: "Leticia kisses Marco," "Leticia kisses Marco when he doesn't agree," and "Leticia kisses Marco when he is sick" (see Figure 5). In all three situations, affection and consent are relevant values, as reflected by their relevance scores. However, the valence changes: consent can either support or oppose the action in the two underspecified situations, but opposes "when Marco doesn't agree." Additionally, the value of health is not usually relevant in the typical context of kissing; however, "when Marco is sick," health becomes relevant and opposes the action. This demonstrates the ability of KALEIDO to adjust to subtle input changes.[14]

False Balance and Extreme Inputs One potential danger when generating diverse values is coming up with a contrived reason why something might be good or bad, even if no reasonable person may hold such a value in such a situation (This is similar to false balance, or "bothesidesism", in news reporting (Imundo and Rapp 2021; Boykoff and Boykoff 2004)). To probe at this, we hand-write 20 situations (10 bad/10 good, in App. J) for which we cannot come up with reasonable values, rights, or duties that would support or oppose them respectively. We run them through KALEIDO^{SYS} after development and find no generated supporting values/rights/duties for the extreme bad actions nor any opposing for the good actions. We take this as limited evidence that our system can avoid false balance.

Universal Declaration of Human Rights Inspired by (Prabhakaran et al. 2022), we think that an ideal dataset containing human rights would contain all rights listed in the United Nation's Universal Declaration of Human Rights[15] (UDHR). We manually extract all 41 human rights we could find from the UDHR and find the 20 closest rights in the dataset as measured by entailment score with WANLI (Liu et al. 2022). We then go through all 41 sets manually and label each for whether or the right is included. We are able to find matches in VALUEPRISM for 97.5% of the UDHR's human rights, demonstrating that the dataset has broad coverage of the UDHR.[16]

[14]While this is a qualitative and not a quantitative experiment, this is not a cherry-picked example —this behavior occurs for other tested situational variations.

[15]https://www.un.org/en/about-us/universal-declaration-of-human-rights

[16]See App. K for all human rights and corresponding matches.

7 Related Work

Value Representations of Language Models Scholars from diverse disciplines have engaged in extensive discussions regarding the incorporation of human ethics and values into LLMs (Wallach and Allen 2008; Jiang et al. 2022; Hendrycks et al. 2023), understanding cultural influences (Santy et al. 2023), examining opinion alignment (Santurkar et al. 2023), and using LLMs as proxies for studying specific human sub-populations in social science research (Argyle et al. 2023b). Jiang et al. (2022) introduced Delphi, a framework trained to reason about ethical perspectives, and showed the ethical limitations of out-of-the-box LLMs. Another noteworthy dimension is the multicultural nature of LLMs. Santy et al. (2023) explored the cultural disparities manifest in LMs and their implication for diverse societies. Tasioulas (2022) criticized the prevailing preference-based utilitarian approach (i.e., which act is likely to yield the optimal fulfillment of human preferences) in AI ethics, pointing out its limitations and proposing as a guide an alternative "humanistic" ethical framework that accounts for additional factors such as pluralism and procedural/participatory considerations. Moreover, Santurkar et al. (2023) and Durmus et al. (2023) introduced novel opinion datasets, quantitatively analyzed the opinions conveyed by LMs, and unveiled substantial misalignments between the stated "viewpoints" of current LLMs and specific demographic groups within the United States.

Alignment of Large Language Models Several computational approaches have been proposed to address the challenge of aligning LLMs with desired values and objectives. Reinforcement learning (RL) has historically been used in multiple NLP tasks to ensure that the generated text is optimized for an arbitrary non-differentiable reward (Johnson et al. 2017; Nguyen, Daumé III, and Boyd-Graber 2017; Ramamurthy et al. 2022; Pyatkin et al. 2023). Lu et al. (2022) optimized a reward function that quantifies an undesired property, while not straying too far from the original model via a KL-divergence penalty. (Bai et al. 2022) explored RL techniques for training LLMs to adhere to legal and ethical guidelines encoded in a constitution, naming it "Constitutional AI." Wu et al. (2023) used fine-grained human feedback as an explicit training signal to train and learn from reward functions in a RLHF fashion. Additionally, Lu et al. (2023) proposed an inference-time algorithm to efficiently tailor LLMs without no fine-tuning, addressing tasks like ensuring safety and fidelity in dialogue models.

Automatic Dataset Curation Previous research in automatic data generation has focused on creating datasets for various tasks, such as commonsense reasoning (West et al. 2022a; Bhagavatula et al. 2023; Wang et al. 2023a; Kim et al. 2023), dialogues (Kim et al. 2023; Xu et al. 2023; Chiang et al. 2023), summarization (Sclar et al. 2022; Jung et al. 2023), and contextual reasoning about offensive statements (Zhou et al. 2023) . West et al. (2022a) introduce the symbolic knowledge distillation framework, which has been extended in subsequent studies through iterative distillation (Sclar et al. 2022; Jung et al. 2023; Bhagavatula et al. 2023; West et al. 2023). In addition, Liu et al. (2022) propose a human-AI collaboration approach to generate high-quality

datasets with challenging examples.

Human Disagreement and Machine Learning Previous work has argued for the importance of modeling annotator disagreement in machine learning (Gordon et al. 2022; Davani, Díaz, and Prabhakaran 2022). Aroyo et al. (2023) measured disagreements in safety judgments across demographic groups and Lu (2023) proposed a framework to explore ambiguity, while Argyle et al. (2023a) explored LLMs' ability to facilitate productive conversations between people who disagree. Baan et al. (2022) argued that common metrics can be misleading when dealing with ambiguous data.

8 Conclusion

In this work, we contribute VALUEPRISM and KALEIDO in the hopes of leading to better value-pluralistic modeling. We validate VALUEPRISM's quality with two human studies, and find that KALEIDO outperforms the teacher's strong performance at generating relevant values, rights, and duties for a given situation. We also show that KALEIDO can help explain variability in human decisions and generalizes to data and frameworks outside of its training scope.

Ethical Impact

Machine-Generated Data. We use GPT-4's open-text generative capabilities to collect VALUEPRISM, leveraging the wide variety of knowledge about human values, rights, and duties latent in LLM's pretraining data. However, we also recognize that in doing so we run the potential for introducing the majority's bias: the generated data may be limited to the values of certain majority groups. In an effort to assess the extent of value plurality and represention, we make a deliberate effort to conduct the validation of the VALUEPRISM by collecting annotations from annotators of various social and demographic backgrounds as described in §4.2. The human annotators find the majority of our data as high-quality at a high agreement rate. Additionally, less than 1% of the validated situations were found to be lacking. Nevertheless, a more extensive study that focuses on the type and nature of values covered by VALUEPRISM remains a compelling direction for future research.

Intended Use. We make VALUEPRISM openly available by individual request with the hope and intention that it furthers research in value pluralism in NLP and AI. However, it is possible that our data can be used in malicious and unintended application (e.g., speech policing or promotion of certain values). We do not endorse its use in such capacity and emphasize that the use of our dataset and model should be limited to research purposes only. Additionally, we limit the data and model available only by individual request to try to prohibit non-research use cases and ensure fair use.

Acknowledgments

The authors thank Ronan LeBras, Jared Moore, Hyunwoo Kim, Jenny Liang, and Sebastin Santy for helpful discussions; Alane Suhr for the example situation in Figure 2; Jared Moore, Dhruba Ghosh, and David Atkinson for draft feedback; and Michael Wilson, Michael Guerquin, and John Borchardt from the AI2 ReViz team for help with the demo. This research was supported in part by DARPA under the ITM program (FA8650-23-C-7316) and the Allen Institute for AI.

References

Alexander, L.; and Moore, M. 2021. Deontological Ethics. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

Archive, G. D. 2011. European Values Study Longitudinal Data File 1981-2008 (EVS 1981-2008). *EVS*.

Argyle, L. P.; Bail, C. A.; Busby, E. C.; Gubler, J. R.; Howe, T.; Rytting, C.; Sorensen, T.; and Wingate, D. 2023a. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41): e2311627120.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023b. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 1–15.

Aroyo, L.; Taylor, A. S.; Diaz, M.; Homan, C. M.; Parrish, A.; Serapio-Garcia, G.; Prabhakaran, V.; and Wang, D. 2023. DICES Dataset: Diversity in Conversational AI Evaluation for Safety. arXiv:2306.11247.

Baan, J.; Aziz, W.; Plank, B.; and Fernández, R. 2022. Stop Measuring Calibration When Humans Disagree. arXiv:2210.16133.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Bhagavatula, C.; Hwang, J. D.; Downey, D.; Bras, R. L.; Lu, X.; Qin, L.; Sakaguchi, K.; Swayamdipta, S.; West, P.; and Choi, Y. 2023. I2D2: Inductive Knowledge Distillation with NeuroLogic and Self-Imitation. arXiv:2212.09246.

Boykoff, M. T.; and Boykoff, J. M. 2004. Balance as bias: global warming and the US prestige press. *Global Environmental Change*, 14(2): 125–136.

Britannica Editors. 2002. Pluralism — Ideology, Diversity & Tolerance — britannica.com. https://www.britannica.com/topic/pluralism-politics. [Accessed 07-08-2023].

Brosch, T.; and Sander, D. 2013. Neurocognitive mechanisms underlying value-based decision-making: from core values to economic value. *Frontiers in Human Neuroscience*, 7.

Casella, G.; and George, E. I. 1992. Explaining the Gibbs Sampler. *The American Statistician*, 46(3): 167–174.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krasheninnikov, D.; Chen, X.; Langosco, L.; Hase, P.; Bıyık, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv:2307.15217.

Chang, R. 1997. *Incommensurability, Incomparability, and Practical Reason*. Cambridge, MA, USA: Harvard.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416.

Dancy, J. 2004. *Ethics Without Principles*. New York: Oxford University Press.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Driver, J. 2022. The History of Utilitarianism. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Durmus, E.; Nyugen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; et al. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv preprint arXiv:2306.16388*.

Feng, S.; Park, C. Y.; Liu, Y.; and Tsvetkov, Y. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. arXiv:2305.08283.

Festinger, L. 1962. Cognitive dissonance. *Sci. Am.*, 207(4): 93–102.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2021. Social Chemistry 101: Learning to Reason about Social and Moral Norms. arXiv:2011.00620.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H. M.; Daumé, H.; and Crawford, K. 2018. Datasheets for datasets. *Communications of the ACM*, 64: 86 – 92.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Gill, M. B.; and Nichols, S. 2008. Sentimentalist Pluralism: Moral Psychology and Philosophical Ethics. *Philosophical Issues*, 18(1): 143–163.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. ACM.

Gowans, C. 2021. Moral Relativism. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition.

Griffiths, N. 2021. Personal Values & Decision-Making Biases. In *Personal Values & Decision-Making Biases*.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2023. Aligning AI With Shared Human Values. arXiv:2008.02275.

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-k.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 8003–8017. Toronto, Canada: Association for Computational Linguistics.

Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. arXiv:2010.05953.

Imundo, M.; and Rapp, D. 2021. When Fairness is Flawed: Effects of False Balance Reporting and Weight-of-Evidence Statements on Beliefs and Perceptions of Climate Change. *Journal of Applied Research in Memory and Cognition*, 11.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Liang, J.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Borchardt, J.; Gabriel, S.; Tsvetkov, Y.; Etzioni, O.; Sap, M.; Rini, R.; and Choi, Y. 2022. Can Machines Learn Morality? The Delphi Experiment. arXiv:2110.07574.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5: 339–351.

Jung, J.; West, P.; Jiang, L.; Brahman, F.; Lu, X.; Fisher, J.; Sorensen, T.; and Choi, Y. 2023. Impossible Distillation: from Low-Quality Model to High-Quality Dataset & Model for Summarization and Paraphrasing. arXiv:2305.16635.

Kant, I. 1785/2002. *Groundwork for the Metaphysics of Morals*. Yale University Press.

Keeney, R. L. 1992. *Value-Focused Thinking: A Path to Creative Decisionmaking*. Harvard University Press. ISBN 9780674931978.

Kekes, J. 1993. *The Morality of Pluralism*. Princeton University Press.

Kim, H.; Hessel, J.; Jiang, L.; West, P.; Lu, X.; Yu, Y.; Zhou, P.; Bras, R. L.; Alikhani, M.; Kim, G.; Sap, M.; and Choi, Y. 2023. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. arXiv:2212.10465.

Komppula, R.; Honkanen, A.; Rossi, S.; Kolesnikova, N.; et al. 2018. The impact of values on sustainable behaviour-A study among Russian and Finnish university students. *European Journal of Tourism Research*, 19: 116–131.

Landemore, H. 2013. Deliberation, cognitive diversity, and democratic inclusiveness: an epistemic argument for the random selection of representatives. *Synthese*, 190: 1209–1231.

Larmore, C. E. 1987. *Patterns of Moral Complexity*. New York: Cambridge University Press.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liscio, E.; Araque, O.; Gatti, L.; Constantinescu, I.; Jonker, C.; Kalimeri, K.; and Murukannaiah, P. K. 2023. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14113–14132. Toronto, Canada: Association for Computational Linguistics.

Litman, L.; Robinson, J.; and Abberbock, T. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, 49(2): 433–442.

Liu, A.; Swayamdipta, S.; Smith, N. A.; and Choi, Y. 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. arXiv:2201.05955.

Lu, X. 2023. Learning Ambiguity from Crowd Sequential Annotations. arXiv:2301.01579.

Lu, X.; Brahman, F.; West, P.; Jang, J.; Chandu, K.; Ravichander, A.; Qin, L.; Ammanabrolu, P.; Jiang, L.; Ramnath, S.; et al. 2023. Inference-Time Policy Adapters (IPA): Tailoring Extreme-Scale LMs without Fine-tuning. *arXiv preprint arXiv:2305.15065*.

Lu, X.; Welleck, S.; Hessel, J.; Jiang, L.; Qin, L.; West, P.; Ammanabrolu, P.; and Choi, Y. 2022. Quark: Controllable text generation with reinforced unlearning. *Advances in neural information processing systems*, 35: 27591–27609.

Martí, J. L. 2017. Pluralism and consensus in deliberative democracy. *Critical Review of International Social and Political Philosophy*, 20(5): 556–579.

Mason, E. 2006. Value pluralism.

Nguyen, K.; Daumé III, H.; and Boyd-Graber, J. 2017. Reinforcement Learning for Bandit Neural Machine Translation with Simulated Human Feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1464–1474.

Plank, B. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10671–10682. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Prabhakaran, V.; Mitchell, M.; Gebru, T.; and Gabriel, I. 2022. A Human Rights-Based Approach to Responsible AI. arXiv:2210.02667.

Pyatkin, V.; Hwang, J. D.; Srikumar, V.; Lu, X.; Jiang, L.; Choi, Y.; and Bhagavatula, C. 2023. ClarifyDelphi: Reinforced Clarification Questions with Defeasibility Rewards for Social and Moral Situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11253–11271.

Páez, J.; De-Juanas, A.; García-Castilla, F.; and Muelas, A. 2020. Relationship Between Basic Human Values and Decision-Making Styles in Adolescents. *International Journal of Environmental Research and Public Health*, 17(22): 8315. [Accessed 31-07-2023].

Ramamurthy, R.; Ammanabrolu, P.; Brantley, K.; Hessel, J.; Sifa, R.; Bauckhage, C.; Hajishirzi, H.; and Choi, Y. 2022. Is Reinforcement Learning (Not) for Natural Language Processing: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization. In *The Eleventh International Conference on Learning Representations*.

Rytting, C. M.; Sorensen, T.; Argyle, L.; Busby, E.; Fulda, N.; Gubler, J.; and Wingate, D. 2023. Towards Coding Social Science Datasets with Language Models. arXiv:2306.02177.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? arXiv:2303.17548.

Santy, S.; Liang, J. T.; Bras, R. L.; Reinecke, K.; and Sap, M. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. arXiv:2306.01943.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: Association for Computational Linguistics.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. M. 2023. Evaluating the Moral Beliefs Encoded in LLMs. arXiv:2307.14324.

Sclar, M.; West, P.; Kumar, S.; Tsvetkov, Y.; and Choi, Y. 2022. Referee: Reference-Free Sentence Summarization with Sharper Controllability through Symbolic Knowledge Distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9649–9668. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Stocker, M. 1990. *Plural and Conflicting Values*. New York: Oxford University Press.

Talat, Z.; Blix, H.; Valvoda, J.; Ganesh, M. I.; Cotterell, R.; and Williams, A. 2022. On the Machine Learning of Ethical Judgments from Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 769–779. Seattle, United States: Association for Computational Linguistics.

Tasioulas, J. 2022. Artificial Intelligence, Humanistic Ethics. *Daedalus*, 151(2): 232–243.

Wallach, W.; and Allen, C. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wang, P.; Wang, Z.; Li, Z.; Gao, Y.; Yin, B.; and Ren, X. 2023a. SCOTT: Self-Consistent Chain-of-Thought Distillation. arXiv:2305.01879.

Wang, Y.; Ivison, H.; Dasigi, P.; Hessel, J.; Khot, T.; Chandu, K. R.; Wadden, D.; MacMillan, K.; Smith, N. A.; Beltagy, I.; and Hajishirzi, H. 2023b. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. arXiv:2306.04751.

Wenar, L. 2023. Rights. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2023 edition.

West, P.; Bhagavatula, C.; Hessel, J.; Hwang, J.; Jiang, L.; Le Bras, R.; Lu, X.; Welleck, S.; and Choi, Y. 2022a. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4602–4625. Seattle, United States: Association for Computational Linguistics.

West, P.; Bhagavatula, C.; Hessel, J.; Hwang, J. D.; Jiang, L.; Bras, R. L.; Lu, X.; Welleck, S.; and Choi, Y. 2022b. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. arXiv:2110.07178.

West, P.; Bras, R. L.; Sorensen, T.; Lin, B. Y.; Jiang, L.; Lu, X.; Chandu, K.; Hessel, J.; Baheti, A.; Bhagavatula, C.; and Choi, Y. 2023. NovaCOMET: Open Commonsense Foundation Models with Symbolic Knowledge Distillation. arXiv:2312.05979.

Williams, B. 1985. Moral Luck. *Critica*, 17(51): 101–105.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. *arXiv preprint arXiv:2306.01693*.

Xu, C.; Guo, D.; Duan, N.; and McAuley, J. 2023. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. arXiv:2304.01196.

Zhou, X.; Zhu, H.; Yerukola, A.; Davidson, T.; Hwang, J. D.; Swayamdipta, S.; and Sap, M. 2023. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. In *Findings of ACL*.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2023. Can Large Language Models Transform Computational Social Science? arXiv:2305.03514.