The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making

Basile Garcia
basile.garcia@unige.ch
University of Geneva
Geneva, Switzerland

Crystal Qian
cjqian@google.com
Google DeepMind
New York City, NY, USA

Stefano Palminteri
stefano.palminteri@ens.fr
École Normale Supérieure
Paris, France

ABSTRACT

As large language models (LLMs) become increasingly integrated into society, their alignment with human morals is crucial. To better understand this alignment, we created a large corpus of human- and LLM-generated responses to various moral scenarios. We found a misalignment between human and LLM moral assessments; although both LLMs and humans tended to reject morally complex utilitarian dilemmas, LLMs were more sensitive to personal framing. We then conducted a quantitative user study involving 230 participants (N=230), who evaluated these responses by determining whether they were AI-generated and assessed their agreement with the responses. Human evaluators preferred LLMs' assessments in moral scenarios, though a systematic anti-AI bias was observed: participants were less likely to agree with judgments they believed to be machine-generated. Statistical and NLP-based analyses revealed subtle linguistic differences in responses, influencing detection and agreement. Overall, our findings highlight the complexities of human-AI perception in morally charged decision-making.

1 INTRODUCTION

Large language models (LLMs) are becoming widely used in applications ranging from conversational agents to decision-making systems capable of making consequential decisions, such as providing medical advice [20], legal advice [10], and mental well-being support [42]. As humans increasingly interact with LLMs, understanding our ability to detect and align with LLMs' judgments becomes crucial, particularly given the risk of misuse, such as the dissemination of disinformation by LLM-powered bots [15, 53].

While prior work has explored AI detection and alignment, the relationship between identification and agreement remains empirically under-investigated, especially in moral decision-making processes. Specifically, it remains unclear whether a participant's belief about the source of content affects their agreement. Our research directly addresses this gap, exploring humans' capacity to detect the source of moral judgments (either human or LLM), their agreement with these judgments, and critically, the relation between these two behavioral outcomes. Additionally, we explore the linguistic factors influencing identification and agreement [43, 51].

To investigate human-AI alignment in moral decision-making, we conducted a series of quantitative experiments involving 230 participants (N=230). First, we collected a corpus of moral judgments by presenting 60 diverse ethical scenarios to human participants and LLM models in the GPT-3.5 family. We then presented these judgments to a new group of participants, who were tasked with identifying the source (human or AI), expressing their agreement or disagreement with the judgment itself, as well as agreement with the accompanying justification. To control for detection bias, we created and evaluated additional corpora with "humanized" LLM responses. Here are the key highlights from our analysis:

- **LLMs exhibit a different moral code from humans, and from each other**: We found that LLMs are highly sensitive to personal vs. impersonal framing; GPT-3.5 `davinci-text-003`, in particular, was much more likely to agree with actions taken in impersonal moral scenarios (where they do not bear personal responsibility) than in personal moral reframings (where they bear personal responsibility). Furthermore, we found that the framing effect was greatly exacerbated between GPT-3.5 `davinci-text-002` and GPT-3.5 `davinci-text-003`, suggesting that moral judgments may be model-dependent.
- **Participants prefer AI justifications over human justifications in morally-complex scenarios**: Although participants preferred human justifications when the stakes were low (e.g., in non-moral scenarios), they significantly preferred LLM-generated justifications in personal moral scenarios (such as when explaining how they would handle the trolley problem), where LLMs exhibited much stronger utilitarian preferences than humans. Participants' preference for AI in these scenarios may stem from a preference for deliberative reasoning in high-stakes settings.
- **However, participants exhibit a strong anti-AI bias**: Even though participants favored the justifications produced by LLMs, they reported disagreement if they suspected that the output was LLM-generated. Across all types of scenarios, participants exhibited a notable anti-AI bias. This result is robust to our efforts to conceal the identity of the LLM through "humanizing" linguistic features, such as introducing typos.
- **Subtle contextual and linguistic cues can reveal AI authorship**: Participants were able to detect the source of generated justifications with moderate accuracy. The detection rate was higher in moral scenarios (such as the trolley problem) than in non-moral scenarios. Slight linguistic differences, such as an increased use of first-person pronouns in human explanations and more pedantic, analytical LLM-generated explanations, provided some signal.

2 BACKGROUND

2.1 Human moral psychology

Moral psychology investigates how people make ethical decisions and evaluate others' actions. Research indicates that moral judgments are often driven by immediate emotions rather than deliberate reasoning [37]. For example, in the trolley problem, individuals are asked if they would sacrifice one person to save five. Responses

vary depending on whether the scenario is framed *personally* (where one must actively push the person onto the tracks) or *impersonally*. This suggests that moral judgments are frequently inconsistent, influenced by context and cognitive biases [13, 18].

Dual process theories offer a popular explanation for these inconsistencies [29, 45]. These theories propose that moral judgment relies on two competing cognitive systems: one that is fast, intuitive, and emotion-driven ("hot"), and another that is slow, deliberative, and rational ("cold") [16, 34]. The deliberative system follows utilitarian principles, focusing solely on the outcomes of decisions, while the intuitive system is swayed by contextual factors unrelated to the final outcome, leading to automatic, emotional responses. Consequently, moral preferences can be inconsistent, as different framing of similar outcomes trigger varying responses [30, 52].

2.2 AI moral psychology

Moral scenarios can also be used to study ethics and alignment within AI systems [21, 32]. As LLMs increase their capacity for conversational decision-making, the practice of recycling tools from cognitive psychology to study LLMs' competencies in terms of decision-making and reasoning has emerged. Several recent studies took the challenge to recycle tools from cognitive psychology to the study of LLMs' competences in terms of decision-making and reasoning [4, 19, 55].

Scherrer et. al. finds that LLMs generally align with human moral values, but in ambiguous cases, their responses can vary based on question phrasing, with closed-source models demonstrating more consistent preferences [51]. This variation may arise from differences in pre-training data and fine-tuning processes [46, 58]. Thus, linguistic features play a significant role in assessing the moral quality of judgments from humans or AI.

2.3 Factors influencing AI detection

Determining whether a decision is made by a human or a machine is crucial: it enhances safety by revealing our susceptibility to manipulation, acts as an epistemological Turing test [27] for assessing AI conversational abilities, and guides the development of LLMs towards human-preferred outputs [11]. Concerningly, recent studies indicate that humans often struggle to reliably distinguish AI-generated texts from human texts, across diverse contexts such as poetry [12, 36] and media misinformation [35]. Additionally, strategies can be employed to "humanize" AI-generated content to increase the difficulty of detection. "Humanized" LLMs have sometimes been judged as more human-like than actual human-generated responses [25], and LLMs can be perceived as more empathetic than human responses when prompted appropriately [54].

2.4 Factors influencing AI alignment

Before LLMs, research into applied fields such as autonomous vehicles highlighted the need for alignment in human and machine moral decision making [1]. Prior research has shown a human tendency to favor human-generated decisions over machine-generated ones, a phenomenon known as algorithm aversion [6]. However, this phenomenon is context-dependent [9]. For instance, humans tend to prefer human judgement over AI judgement in the context of medical decision-making [8], but prefer AI judgement in numerical tasks [40].

Agreement with LLM-generated text hinges on factors like task nature, perceived authorship [14], prior AI interactions, and even cultural context [31]. While ChatGPT's responses in social scenarios have been rated as more balanced and empathetic than human advice [24], people still show a strong preference for human advice on moral issues [47]. AI authors are perceived as less competent, though humans still value their advice [7]. Moreover, in persuasive content creation (e.g. advertisements), AI efforts are often rated higher than human efforts. Revealing the source of content production lessens the quality gap between human- and AI-generated content, without affecting the assessed quality of AI-created content [56]. This suggests that human favoritism, rather than AI aversion, drives the perceived quality and perceived value [44].

To summarize, two competing hypotheses can help explain perceptions of LLMs in moral decision-making. A pro-AI bias may occur when machines are seen as authoritative sources of knowledge. In contrast, an anti-AI bias might emerge from societal or psychological prejudices against machines, often stemming from the belief that machines lack agency and the capacity for compassionate or morally sound decisions [2].

3 METHODS

Our experimental design is summarized follows (and shown in Figure 1).[1]

(1) **Corpus generation** First, we create two corpora of responses to scenarios of various types: non-moral, impersonal moral, or personal moral. For each scenario in corpus 1, 30 human participants and 30 API calls of GPT 3.5 `davinci-text-002` (**dv2**) provide a *response*, which includes a *judgement* (yes/no) and a *justification* (free text). Corpus 2 uses GPT 3.5 `davinci-text-003` (**dv3**) instead of dv2.
(2) **Corpus transformation** Because human- and LLM- generated justifications may have liguistic differences (such as typos, or response length), we create corpus 3, which "humanizes" the dv2 responses from corpus 1 by adding typos or shortening the text.
(3) **Corpus evaluation** Next, we have human raters evaluate each response from the 3 corpora. For each evaluation, they answer 1) whether they think the text was human- or LLM-generated, 2) whether they agree with the *judgement*, and 3) whether they agree with the *justification*.
(4) **Linguistic analysis** These remaining steps use statistical and computational analysis to figure out which specific signals are being communicated in the justification text to affect detection and alignment. In this step, we perform a linguistic analysis on potential linguistic differences between human- and LLM- generated text.

[1]The research was conducted in accordance with the principles and guidelines for experiments involving human participants as outlined in the Declaration of Helsinki (1964, revised in 2013). The study received approval from Paris School of Economics ethical committee (2024-007). Informed consent was obtained from all participants prior to their involvement in each experiment.

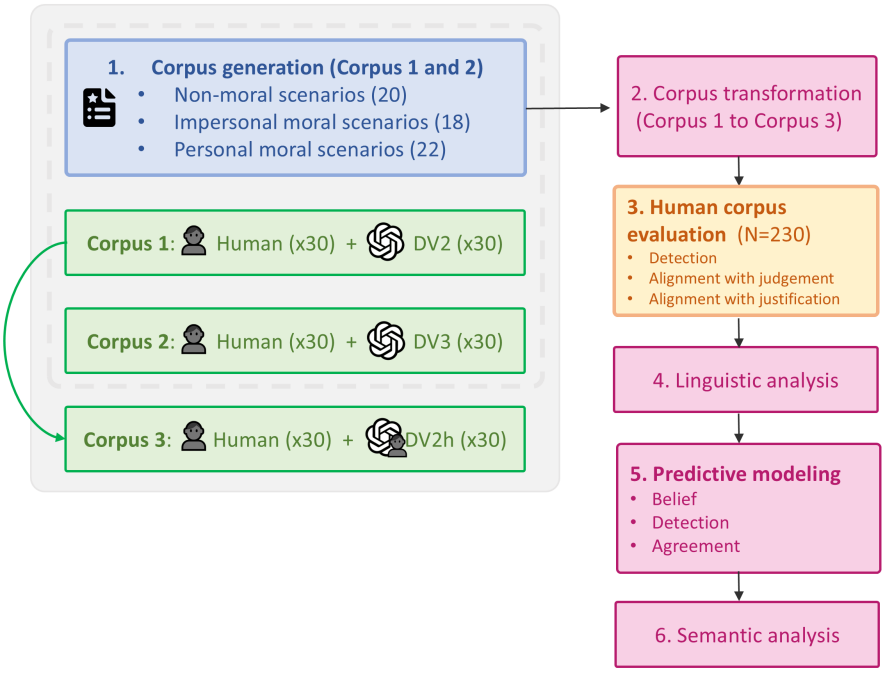

Figure 1: A diagram showing the experimental design. The yellow boxes (Steps 1 and 3) signal involvement and data aggregation from human participants, whereas the orange boxes (Steps 2, 4, 5, and 6) denote quantitative, statistical, and computational methods. The output of the experiment is an analysis on three corpora of human- and LLM- generated responses to various scenarios.

(5) **Predictive modeling** We use NLP techniques to evaluate whether the responses in the corpora have predictive signals on our outcomes of interest. We build models to predict 1) the true source of the text, 2) the human rater's belief of the source of the text, 3) whether the human rater's belief was correct, and 4) whether the human agrees with the justification.
(6) **Semantic analysis** Finally, we use interpretability methods to identify *which specific tokens* in the justification text have predictive power on the outcomes in described in 4.

3.1 Corpus generation

In the preliminary item-generating stage of this study (Figure 2A; corpus generation experiments), we presented 60 moral scenarios from Greene et al. [17] to participants. There were three categories of these scenarios:

- **Non-moral**: Scenarios that do not engage complex moral reasoning, e.g. "Is it appropriate to wait for a promotion before buying a new computer, versus buying it immediately?"
- **Impersonal moral**: Moral scenarios involving indirect intervention, e.g. "Is it appropriate to turn a runaway boxcar away from five people but towards one person?"
- **Personal moral**: Emotionally salient scenarios that are more engaging, e.g. "Is it appropriate to push one person off a bridge to stop a runaway boxcar from hitting five people?" These scenarios are similar to impersonal moral scenarios but involve a sense of personal agency.

Participants provided a yes/no decision and a free-text justification for each scenario. Each corpus contained 3600 responses, derived from 60 scenarios with 30 human responses and 30 LLM responses per scenario. The temperature of the LLM calls was set to 0.7, to maintain a form of inter-prompt variability similar to what is observed in humans. After removing ambiguous answers, we created two corpora:

- **Corpus 1**: 3542 responses from the `davinci-text-002` batch of experiments. Contains human and `dv2` responses.
- **Corpus 2**: 3420 responses from the `davinci-text-003` batch of experiments. Contains human and `dv3` responses.

3.2 Corpus transformation

To mitigate any potential linguistic differences, we used GPT-3.5 to "humanize" the LLMs' responses in corpus 1 by shortening them, introducing a few typos, and providing human examples to imitate; this process is shown in Figure 3. Our resulting `dv2_humanized` dataset showed no significant difference in length or number of typos from the human dataset, as evidenced by statistical t-tests. This resulting dataset of human and `dv2_humanized` responses is **corpus 3**.

3.3 Corpus evaluation

In the main stage of the study (Figure 2B), we presented the responses generated in the first stage to new human participants (N=230). 77 participants evaluated the 46 responses in corpus 1, 76 participants evaluated the 45 responses in corpus 2, and 77 participants evaluated the 46 items in corpus 3. For each response, participants had to answer three binary questions:

(1) **Detection**: Whether they think these answers were given by humans or AI (human/AI)
(2) **Agreement with the binary judgement**: Whether they agree with the binary moral judgment (yes/no)
(3) **Agreement with the justification**: Whether they agree with the provided justification (yes/no)

3.3.1 Participants. The corpus generation experiment involved 30 human participants (15 females, mean age = 34 ± 10.26) (Figure 2A). The detection and agreement experiments involved 230 participants (113 females; mean age = 35.57 ± 11.71) (Figure 2B). They were recruited through Prolific[2] with the screening requirement that they were fluent in English. Instructions were fully transparent, informing participants that they are expected to give meta-judgments on both human and AI-generated answers. In addition to a base rate, participants were incentivized in Turing test questions with a bonus of 5 cents for each correct identification (AI/human). The average final bonus was $1.46 ± 0.28, which was significantly higher than what they would have received on average for making random choices[3].

3.3.2 Statistical evaluation. The main outcome measures from the corpus evaluation steps are the binary responses to the three questions: detection (human- or LLM- generated), agreement with judgement (yes/no), and agreement with justification (yes/no). The responses to the two 'detection' questions were transformed into an accuracy measure. The two agreement answers were averaged, such that a full agreement was coded as 1, a partial agreement as 0.5, and a complete disagreement as 0.

All inferential tests were conducted using Python 3.9 and the Pingouin 0.5.4 library. Two-tailed t-tests were performed throughout. For multiple comparisons, we used the `pairwise_ttests` function with Bonferroni corrections systematically applied. Single t-tests were performed using the `ttest` function. We report the following statistics for each t-test: Student's t-value ($T(df)$), p-value significance (p), Cohen's d (d), and Bayesian factor (BF_{10}).

The anova function was used for comparing independent samples (one-way or two-way ANOVA), and the `rm_anova` function was used for repeated measures ANOVA (one-way ANOVA with repeated measures). For ANOVAs, we report the F-statistic ($F(df)$), p-value (p), and partial eta-squared (η_p^2).

3.4 Linguistic evaluation

We performed the corpus transformation step because we anticipated that linguistic differences between human- and LLM-generate text could be contributing to detection and alignment. Specifically, we hypothesize that 1. humans generate shorter justifications in length, 2. humans create more typing errors, and 3. humans tend to write more often in the first-person. We perform a statistical analysis to evaluate these differences.

3.5 Predictive modeling

Next, we wanted to understand whether state-of-the-art models could glean predictive signals for the outcomes of interest within the justification text.

Pre-processing. Lexical and semantic analyses were performed using `NLTK` tokenizers and stopwords[4], and a `TfidfVectorizer` to transform raw text data into a matrix of TF-IDF features [48]. The vectorizer was configured to remove common English stop words, exclude numbers, and include only alphabetic words. To limit the feature space, we set the `min_df` parameter to 3, excluding words that appeared in fewer than three documents, and capped the maximum number of features at 1000.

Transformer models. We fine-tuned a series of pre-trained transformer-based models locally (`DistilBERT`) [50], optimizing hyperparameters with `optuna`[5]. Given the pre-processed text data, these models were used to predict 1. the true source of the explanation text, 2. the participant's predicted source, and 3. the participant's agreement with the judgement. Label classes were encoded for these multi-class and binary classification models.

3.6 Semantic analysis

To understand which specific semantic features in the text could explain outcomes, we built random forest classifier with 100 estimators trained on the corpora's dense representation [5]. The decision to switch model architectures was made after confirming that performance scores were comparable across the transformer-based and tree-based implementations; the random forest implementation was less computationally intensive and easier to interpret through feature importance scores. To interpret the model's predictions, we applied SHAP (SHapley Additive exPlanations) values using `TreeExplainer` [41]. SHAP values decompose predictions into contributions from individual features, providing insights into how different features influenced the model's decisions. Positive SHAP values indicate that a feature contributes to a higher prediction, while negative values suggest a lower prediction.

4 RESULTS

For the purpose of readability, the statistical evidence for the results section are not embedded directly in the text. The results of two-tailed t-tests are shown in Table 1. ANOVA statistics are found in the footnotes. We claim that a result is statistically significant when the *p-value* of the accompanying t-test has a p-value $p < 0.001$.

4.1 Corpus evaluation (1 and 2): judgement

We generated corpus 1 (human- and dv2- generated responses) and corpus 2 (human and dv3- generated responses). Each response had a *judgement* (yes/no), and a *justification* free-text. Here, we evaluate the *judgement* values as a function of the type of the moral scenario: "Non moral", "Impersonal moral" and "Personal moral" (Figure 2A).

[2] https://prolific.co/
[3] $T(229) = 17.28, p < 0.001^{***}, d = 1.14, \mathrm{BF}_{10} = 9.768 \times 10^{39}$
[4] https://www.nltk.org/
[5] https://optuna.org/

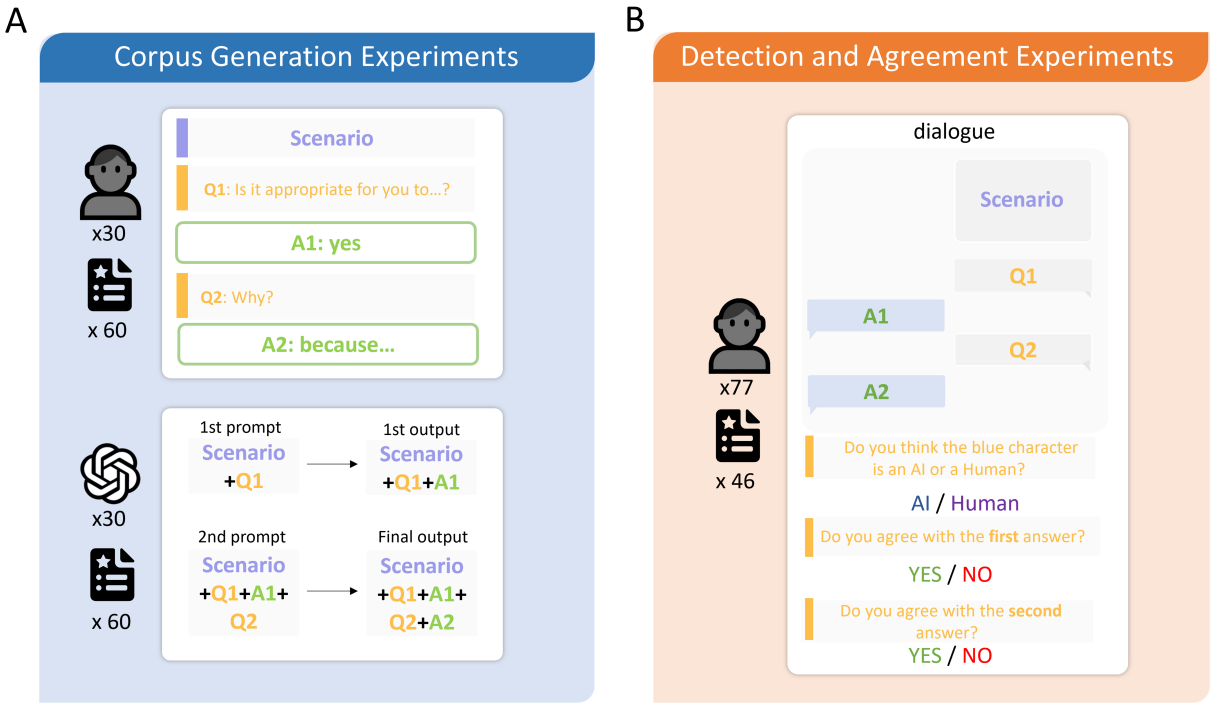

Figure 2: (A) Schematic interface and method used in the experiment we used to generate corpus 1 (human and dv2 responses) and corpus 2 (human and dv3 responses). (B) Schematic interface used in the *3. corpus evaluation* step.

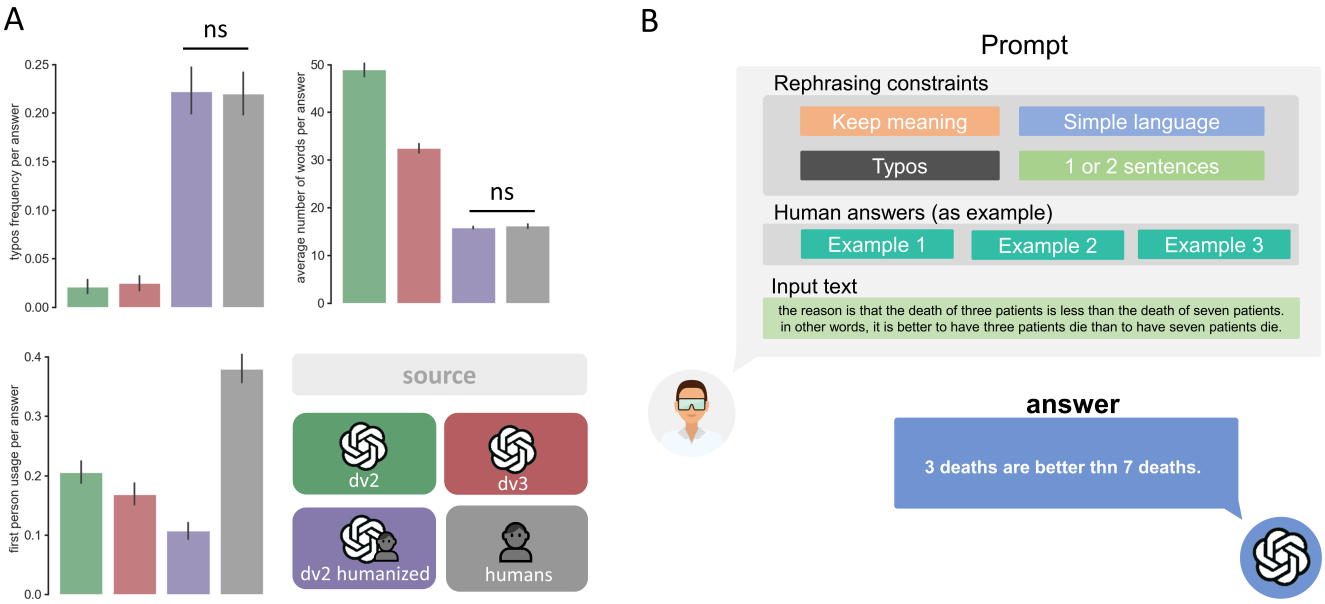

Figure 3: (A) Identified linguistic features which have been found to be different between human- and LLM-generated responses. (dv2: text-davinci-002; dv3: text-davinci-003, dv2h: humanized dv2). (B) Schematized prompting strategy to generate the humanized LLM response, by reducing size and including typos.

Table 1: This table contains statistical values from two-tailed t-tests; each section corresponds to a section in the results. We report the following statistics for each t-test: Student's t-value ($T(df)$), the significance of the p-value (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, n.s.: not significant), Cohen's d (d), and Bayesian factor (BF_{10}). The source value, when numeric, refers to the corresponding corpora.

Source	Feature	T(df)	p	d	BF_{10}
4.1. Judgement					
Human	Impersonal vs. personal scenarios	T(30) = 2.8	*	0.69	5.02
dv2	Impersonal vs. personal scenarios	T(29) = 10.05	***	2.23	1.65e8
dv3	Impersonal vs. personal scenarios	T(29) = 49.28	***	12.11	1.09e26
4.2 Detection					
1	DV2 detection	T(76) = 10.49	***	1.19	3.21e13
2	DV3 detection	T(75) = 13.48	***	1.59	2.19e19
1	Impersonal moral vs. personal moral	T(75) = 3.93	***	0.57	116.47
2	Personal moral vs. non-moral	T(75) = 4.52	***	0.56	819.58
1, 2	Impersonal moral vs. personal moral	T(152) = 3.50	**	0.32	30.06
1, 2	Personal moral vs. non-moral	T(152) = 30.6	**	0.28	8
4.3. Alignment					
1	Personal moral vs. non-moral	T(76) = 4.18	***	0.71	263.91
2	Personal moral vs. non-moral	T(75) = 5.51	***	0.80	3.06e4
1	Impersonal moral vs. personal moral	T(75) = 5.76	***	0.77	7.97e4
1, 2	Personal moral vs. non-moral	T(152) = 6.86	***	0.75	5.32e7
1, 2	Impersonal moral vs. personal moral	T(152) = 4.68	***	0.48	2180.9
1, 2	Non-moral	T(152) = 5.12	***	0.58	1.35e10
1, 2	Moral	T(152) = 4.35	***	0.19	0.39
1.2	Impersonal	T(152) = 1.73	n.s.	0.19	0.39
4.3. Alignment; conditioning on belief					
1, 2	Non-moral	T(143) = 5.76	***	0.67	2.19e5
1, 2	Impersonal moral	T(143) = 2.7	*	0.31	3.08
1, 2	Personal moral	T(143) = 4.31	***	0.5	498.6
4.4 Corpus 3					
3	dv2_humanized detection	T(76) = 8.24	***	0.93	2.3e09
3	Conditioned on belief (aggregate)	T(75) = 3.57	***	0.4	37.82
4.5 Linguistic analysis					
1	Length	T(76) = 8.34	***	1.25	3.44e9
2	Length	T(75) = 11.39	***	1.8	1.139e15
1	Typos	T(75) = 4.02	***	0.58	149.147
2	Typos	T(75) = 10.33	***	1.41	1.391e13
3	Typos	T(76) = 1.55	n.s.	0.19	0.39
1	First-person usage	T(76) = 7.31	***	1.04	4.423e07
2	First-person usage	T(75) = 13.28	***	1.68	2.143e18
3	First-person usage	T(76) = 8.72	***	1.09	1.786e10
4.6 Predictive modeling					
1	Length	T(76) = 0.36	n.s.	0.05	0.133
2	Length	T(75) = 0.49	n.s.	0.08	0.142
1	Typos	T(75) = 0.76	n.s.	0.12	0.167
2	Typos	T(75) = 0.1	n.s.	0.02	0.127
1	First-person usage	T(76) = 0.81	n.s.	0.12	0.173
2	First-person usage	T(75) = 2.07	n.s.	0.31	0.939
3	First-person usage	T(76) = 2.67	*	0.37	3.377

In previous studies, decisions with utilitarian outcomes (benefiting the group) are more readily endorsed when framed in an "impersonal" manner [16, 34]. That is, participants are less keen to sacrifice one person in the *Trolley Problem* scenario when they themselves are the one who must push the person onto the tracks.

However, in our study, humans are not the ones to exhibit this behavior. Figure 4B shows the likelihood of endorsing across the three scenarios (with accompanying statistical values in Table 1). We find that there is no statistical difference, in our human sample, between the amount of impersonal moral scenario and personal moral scenario endorsements; however, humans significantly endorse non-moral scenarios over personal moral scenarios. Interestingly, the `dv2` and `dv3` responses appear to follow a different moral code: `dv2` significantly endorses impersonal moral scenarios more than personal moral scenarios. DV3 agents (`text-davinci-003`) displayed an even greater sensitivity to this framing effect. These results suggest that that human moral preferences are context-dependent, that different versions of models (`dv2` vs. `dv3`) can express different moral preferences, and that there can be misalignment between human and LLM moral judgements.

4.2 Corpus evaluation (1 and 2): detection

In Figure 5, row A, we show the probability of correctly detecting the source of the judgement as a function of the scenario type. We find that participants significantly but imperfectly identify the source of the item under evaluation, significantly higher than random guessing. In Corpus 1 (human vs. `dv2` justifications) correctly identified the source 64% of the time. Identification was even higher for Corpus 2 (human vs. `dv3`), reaching 71% accuracy.

We also find that the context of the scenario matters: justifications related to to personal and impersonal moral scenarios were more easily detected than non-moral scenarios. That is, participants were more likely to correctly detect that content was AI- or human-generated in moral scenarios (both personal and impersonal), as compared to non-moral. The LLMs, as compared to humans, were much more likely to endorse utilitarian scenarios; participants appear to pick up on this difference. These differences persisted when corpus 1 and 2 were grouped.

4.3 Corpus evaluation (1 and 2): alignment

We investigated whether agreement rates varied based on the source of the justification, and how this varied with scenario type. Experiments across both GPT models revealed a significant main effect of scenario type.[6].

In Row B of Figure 5, we see that participants' alignment with human- or LLM- responses depends on the scenario type. Participants significantly preferred human judgements over AI judgements for non-moral scenarios. Conversely, they significantly preferred AI judgements for personal moral scenarios. In impersonal moral scenarios however, we did not observe any significant bias[7]. This result suggests that when faced with morally complex scenarios, humans tend to agree more with justifications emitted by machines rather than humans.

In Row C of Figure 5, we condition this alignment on whether the participant *believes* the justification was human- or LLM- generated. This exposes a significant anti-AI bias across all scenarios. Participants showed significantly more anti-AI bias in non-moral scenarios compared to personal moral ones, and in impersonal compared to personal moral scenarios for DV2. These results were consistent when experiments were grouped. These results suggest that participants' agreement with LLM judgments differs based on their belief about the source, particularly for complex (personal) moral scenarios.

4.4 Corpus 3

Comparing the results of `dv2`, `dv3` and `dv2_humanized`, humanization was successful in significantly reducing the detection rate,[8] which dropped below 60%, but still remained above 50% (Figure 6A). The difference in agreement between human and machine items remained unchanged, regardless of considering the actual source of the item (Figure 6B) or the belief (Figure 6C).[9]. Although reduced, the accuracy in the detection task was still significant in Corpus 3, suggesting that participants were not dependent on features such as text length and the number of typos for detection.

4.5 Linguistic analysis

Thus far, we've learned that participants are weakly able to detect LLM-generated moral justifications, and that this detection varies across scenario type. Is it the case that linguistic features are driving this detection ability? In Figure 7, we quantify each hypothesis (columns A, B, and C, respectively) and plot the probability of detection (row A) and alignment (row B). In corpus 1 and corpus 2, we see that length, typos, and first-person usage all significantly affect detection rates; participants are, in fact, using these linguistic features to influence their detection.

As expected, in corpus 3 (where we targeted reducing these differences), these effects are greatly diminished. While the use of first-person markers still significantly predicted human identification, the presence of typos as well the length of the text is no longer significant. This suggests that humans are capable of picking up on subtle linguistic cues, such as first-person pronoun usage, in the absence of more explicit cues such as text length and typos.

We also test whether typos, length and first-person pronoun usage predicted alignment. The only significant effect was observed for first-person usage in the humanized condition, although this effect was not replicated in other conditions. As linguistic features did not significantly predict agreement, this suggests that deeper concepts, semantic features, and subjective preferences affect alignment.

4.6 Predictive modeling

Here, we show the model performance of the multi-class and binary classifiers, built using hyperparameter-optimized, transformer-based models. Our models, like the participants, were able to detect provenance with moderate accuracy. Accuracy scores are still low, but higher than random selection. Models are also able to predict

[6] $F(2, 152) = 8.37, p < 0.001^{***}, \eta_p^2 = 0.09$; $F(2, 150) = 20.79, p < 0.001^{***}, \eta_p^2 = 0.21$

[7] We define bias as a significant deviation from 0.

[8] $F(2, 227) = 27.57, p < 0.001^{***}, \eta_p^2 = 0.19$

[9] $F(2, 227) = 2.55, p = 0.07, \eta_p^2 = 0.02$; $F(2, 223) = 2.52, p = 0.08, \eta_p^2 = 0.02$

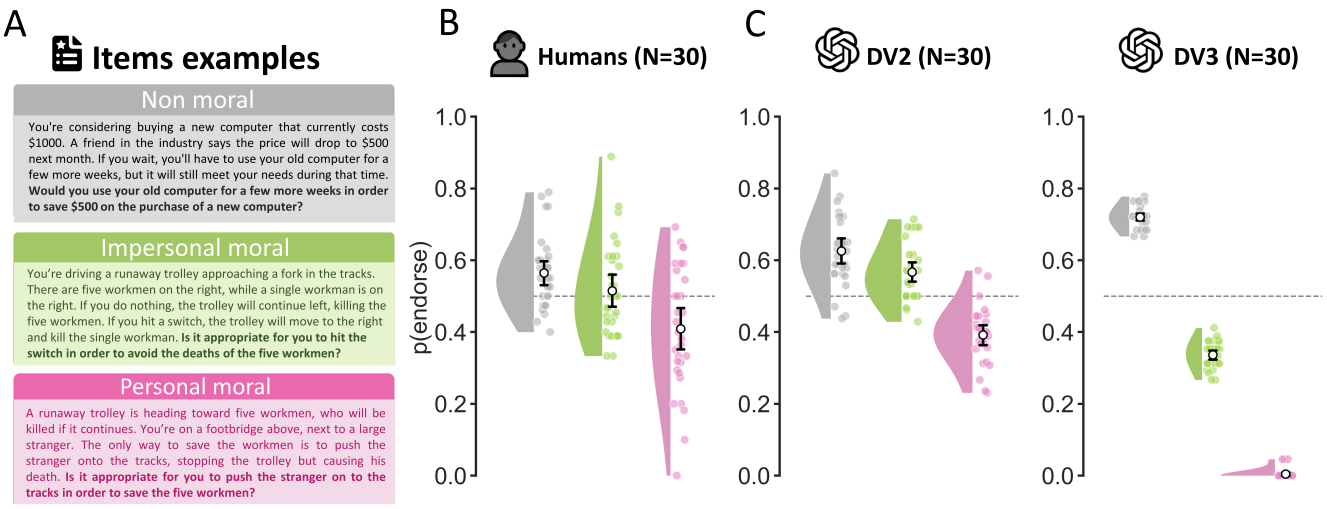

Figure 4: (A) Example of scenarios across three categories (taken from Greene et al. 2004). (B) Endorsement of the different moral actions as a function of category of scenario; 'non moral' refers to scenarios with no moral stakes; 'impersonal moral' refers to scenarios with moral scenario whose resolution does not involve a direct, personal involvement of the participant (emotionally non-engaging); 'personal moral' refers to moral scenario whose resolution involve a direct involvement of the participant (emotionally engaging). Note, what is asked in moral scenario is judging the appropriateness of the utilitarian response. (C) Same as (B), but for the two considered LLMs; DV2= text-davinci-002, DV3: text-davinci-003.

	F1 score	Accuracy
1. Multiclass provenance classifier		0.61
dv2	0.66	
dv2_humanized	0.58	
dv3	0.74	
human	0.49	
2. Binary provenance classifier		0.63
llm	0.63	
human	0.64	
3. Agreement predictor		0.63
disagree	0.64	
agree	0.62	
4. Identification predictor		0.62
incorrect identification	0.63	
correct identification	0.59	

Table 2: Performance table of 4 transformer models to predict outcomes based on the justification free text. 1. and 2. predicts the provenance of the justification text. 3. predicts whether the human rater agreed yes/no with the decision. 4. predicts whether the human correctly identified the explanation as human- or AI-generated.

alignment and detection, indicating that there are potentially quantifiable features within the text that can inform these higher-level choices.

While training and tuning these models, the model performance was sensitive to the distributions within the training set, particularly for performance tradeoffs between dv2_humanized- and human-generated text. That is, the models would easily confuse the two (Table 2, 1.), but could easily differentiate dv3- generated text.

4.7 Semantic analysis

To evaluate the semantic features influencing human responses on detection and agreement, we applied SHAP analysis to a random forest classifier trained on data from Experiments 2 and 3, comparing the human vs. dv2, dv3 and dv2_humanized results.

In dv2 and dv3, SHAP analysis showed that the word "appropriate" was a strong indicator of AI-generated content (Figure 8; first row). LLM responses often reworded the question. For example if the question was worded like, "Is it appropriate to do X to save Y?" the LLM would begin its response with, "Yes, it is appropriate because…" Participants picked up on this, often identifying justifications containing the term "appropriate" as AI-generated. In particular, dv3 appeared more pedantic and repetitive, which may attribute to its high detection rate. The following are examples of dv3-generated justifications that the predictive models were able to correctly identify with high probability:

> *"it is not **appropriate** to leave this man by the side of the road in order to preserve your leather upholstery because human life is more important than material possessions. taking the time to help this man in need is the right thing to do and it is your moral responsibility to prioritize his health and safety over your car's upholstery. it may be inconvenient to have to clean up the mess and deal with the ruined upholstery, but it is a small price to pay for saving someone's life."* —dv3

> *"policy A has a 90% chance of causing no deaths at all and has a 10% chance of causing 1000 deaths. policy B has a 92% chance of causing no deaths and an 8% chance of causing 10, 000 deaths. overall, policy A is a*

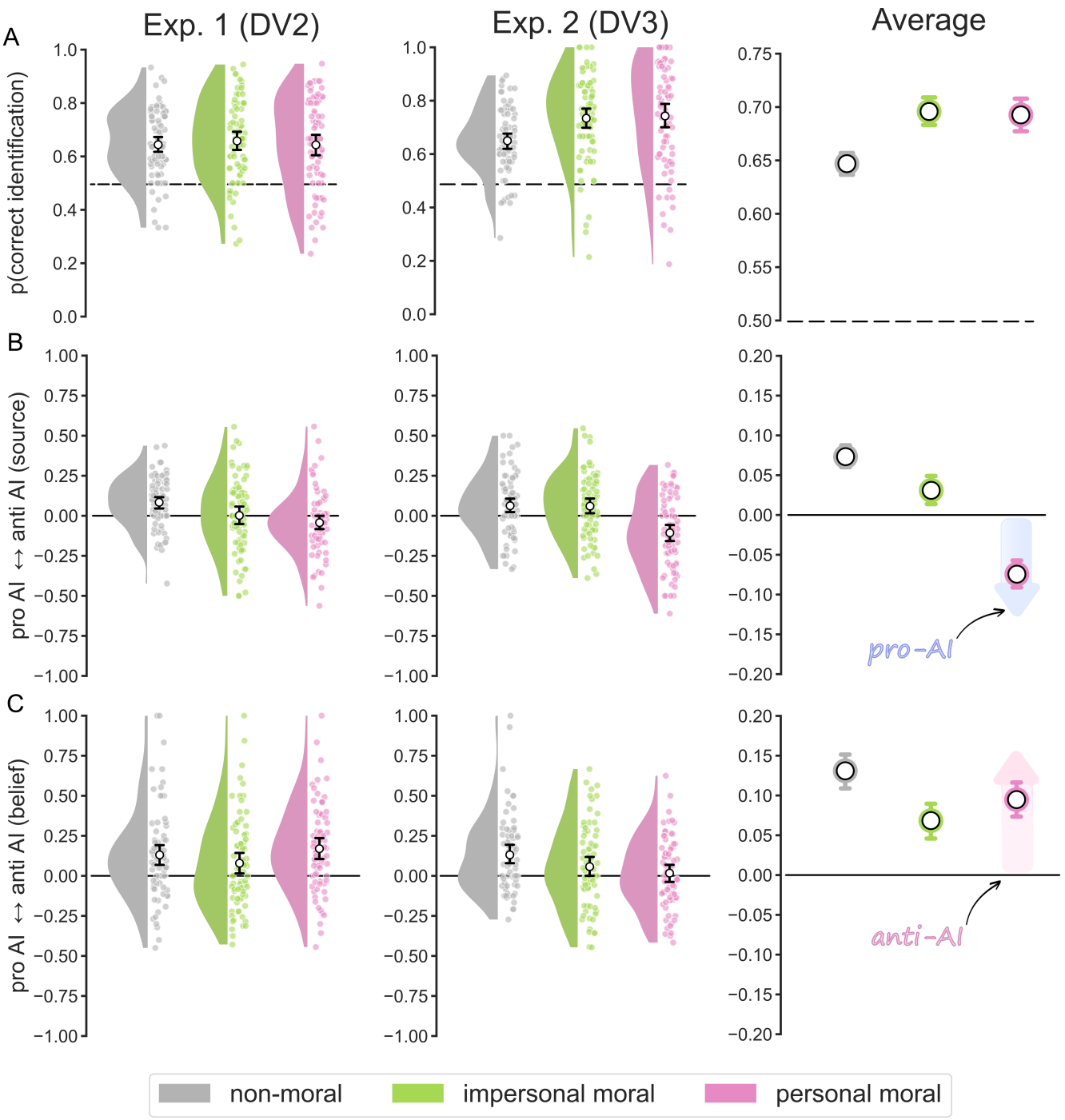

Figure 5: (A) Probability of correctly detecting the source of the judgement (p(correct identification)), as a function of the scenario type in Corpus 1 (leftmost column; Exp. 1 DV2), Corpus 2 (central column; Exp. 2 DV3), and on average (rightmost column; Average). (B) Difference in agreement between the trials featuring human-generated items and those featuring LLM-generated items as a function of the scenario type. (C) Difference in agreement between the trials the participant declared as being human-generated and those declared to be LLM-generated (belief).

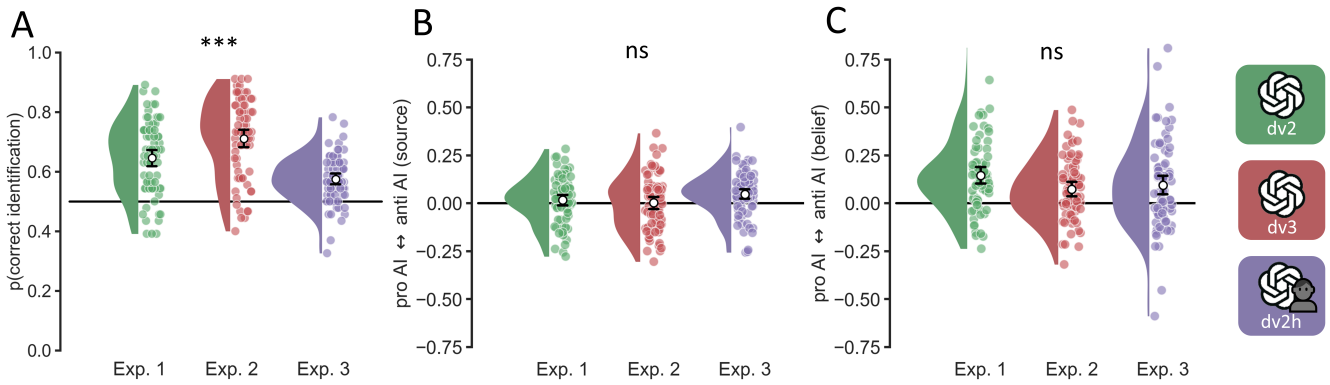

Figure 6: (A) Correct response in the detection task rate across corpus 1: dv2; corpus 2: dv3; corpus 3: dv2 humanized. (B) Actual source-oriented agreement differential. (C) Declared source-oriented (belief) agreement differential.

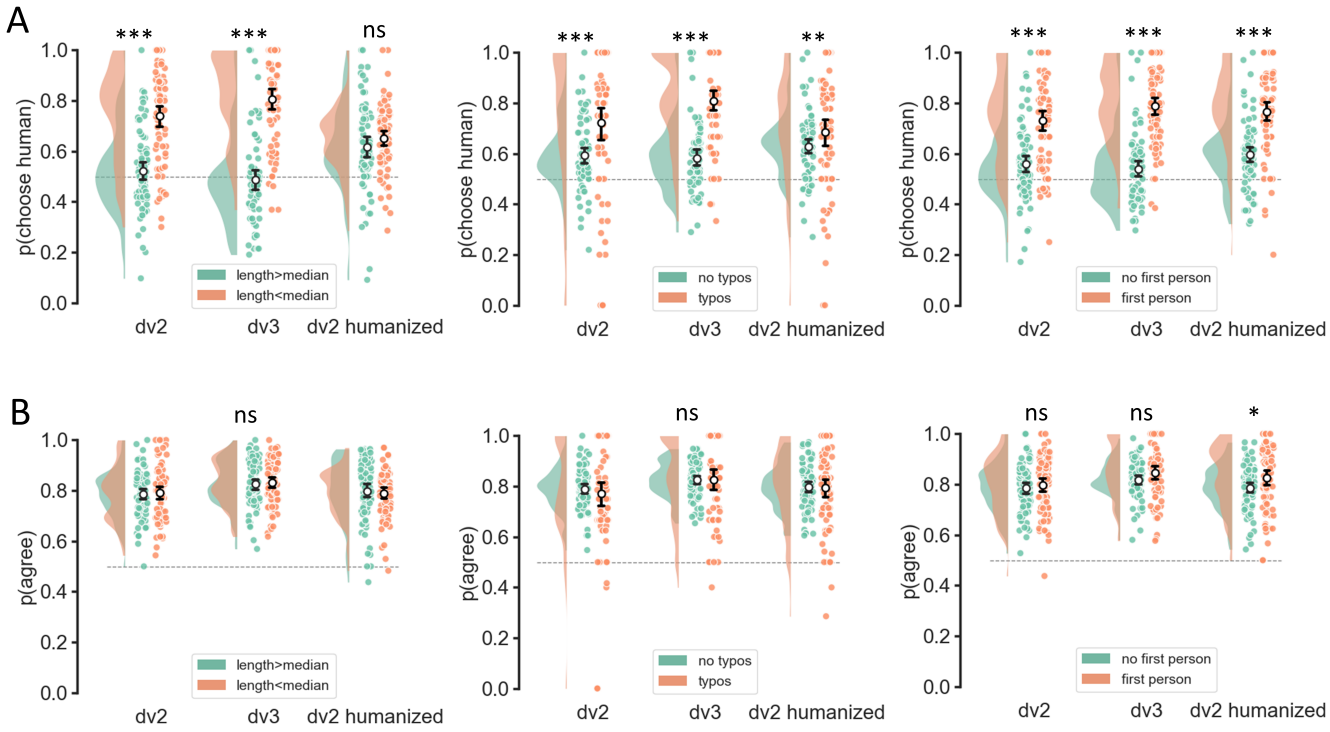

Figure 7: (A) Probability of choosing 'human' in the detection task as a function of different linguistic features. Items are split as a function of their length (leftmost column), the presence or not of typos (central column), and the utilization of first-person marker (rightmost column). (B) Probability of agreeing with a justification as a function of linguistic features.

> *better choice because it has a lower chance of causing deaths."* —dv3

Words like "additionally" and "reason", which imply logical reasoning, were correctly recognized by both humans and our predictive models as markers of AI-generated judgments (Figure 8; second row). This suggests that both the classifier and participants detected formal reasoning language as a distinctive feature of AI-produced judgments. This following example shows a correctly-predicted dv2 instance demonstrating logical reasoning:

> *"killing someone, even your own child, is never an **appropriate** solution. human life is inherently valuable, and no one should be asked to make the impossible decision of sacrificing a life for the sake of others. **therefore**, it is not appropriate for you to kill your oldest son in order to save your husband and [other] children."* —dv2

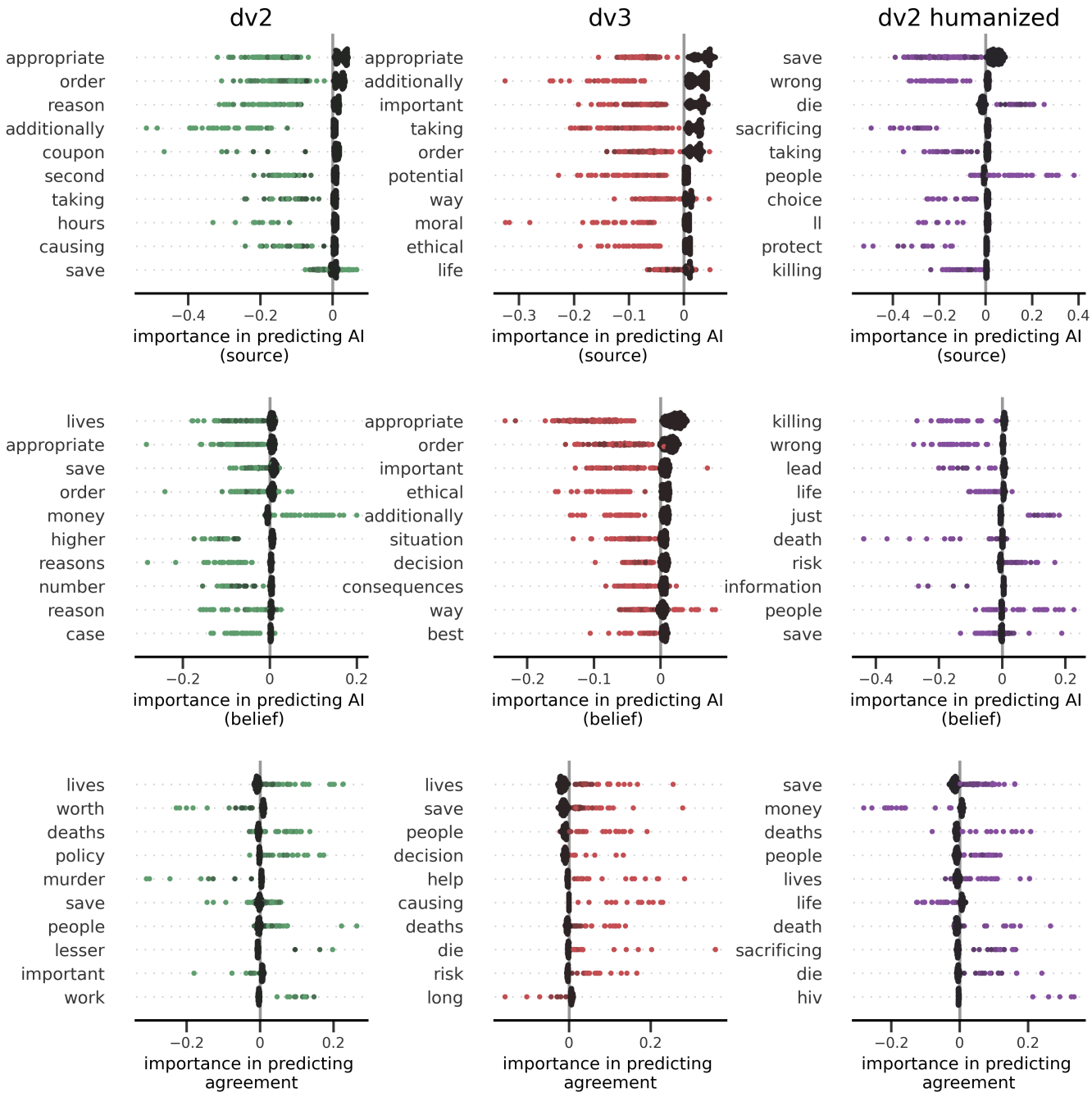

Figure 8: SHAP beeswarm plot summarizing the impact of key features on the random forest classifier predictions. The model predicted three variables (rows: source, belief, agreement) for each experiment (columns; dv2, dv3, dv2 humanized). Each point represents a data instance, with the x-axis showing the SHAP value, which reflects the importance of a word in influencing the prediction. The color represents the feature value, which in this case is the word's frequency. Black dots indicate a low frequency of the word for a given sample of moral judgements, while colored dots signify a higher frequency. Thus, the feature value (color) goes from black (low word occurrence) to brighter colors for words that appear more frequently. Features are ordered by average absolute SHAP value, highlighting their relative importance. Features with higher SHAP values have a larger influence on the model's output, with the most important features appearing at the top. We only display the 10 most important features.

Words tied to utilitarian reasoning (e.g., "lives," "save," "deaths") were predictive of disagreement with justifications (Figure 8; third row). These words often appear in justifications to personal moral scenarios, which was the category of scenarios receiving the least endorsement by participants. For example, one such scenario asks whether it is appropriate for a doctor to sacrifice one patient to save five others. Participants often disagreed with the utilitarian outcomes in this case, highlighting the moral conflict. In contrast, the presence of terms like "murder" predicted agreement, often used in moral statements like "the murder of innocents is always wrong" aligning with deontological principles [23].

The predictive terms in the `dv2_humanized` corpus, where GPT-3.5 was prompted to imitate human responses, were different, both in predicting the source of the judgements (Figure 8; first row) and participants' beliefs (Figure 8; second row). Words like "save," "sacrificing," "choice," and "killing", associated with utilitarian personal moral scenarios, frequently appeared in AI-generated content. Participants recognized these cues even without formal features (e.g., text length and typos). This reflects large language models' tendency toward utilitarian reasoning, which participants likely used as a heuristic to detect AI-generated responses.

Agreement with the humanized dv2 model (Figure 8; third row) followed similar patterns as dv2 and dv3: words evoking personal moral scenarios, such as "deaths," 'lives," and "sacrificing," often led to participant rejection. These terms reflect emotionally charged conflicts between utilitarian outcomes and necessary actions.

5 DISCUSSION

5.1 Summary

In the first part of our study, we set the stage to examine differences in human- and LLM- reasoning, administering a well-established psychology task designed to elicit contrasting moral preference across a diversity of scenarios [16, 34].

We found that human preferences are scenario-dependent; they agree more with judgements that are not morally complex (i.e. non-moral scenarios). We administered the same task to LLMs [4, 19, 55] and found there was some misalignment between human- and LLM-generated judgements, especially across scenario type [21, 33, 46].

Then, we asked raters to evaluate the responses from our corpora. We found that participants were only somewhat able to distinguish between the moral judgements generated by humans and LLMs, and that the context of the scenario was important. For judgments on relatively trivial matters, participants generally showed greater agreement with human justifications. However, participants preferred AI-generated responses to complex moral scenarios. This pro-AI bias for complex moral scenarios was not consciously recognized by participants; rather, it coexisted with a rather pervasive belief-based anti-AI bias, according to which higher agreement was given to justifications that our participants believed coming from humans, even if it was not the case.

5.2 Relationship between detection and alignment

Newer LLMs may exhibit more "correct answer bias". There was a significant decrease in alignment within responses in `dv3`, as compared to `dv2`, potentially indicative of a "correct answer" bias [46]. This bias suggests that newer LLMs may be trained and fine-tuned to generate socially accepted responses, leading to reduced diversity in their outputs.

LLMs' more deliberate reasoning may be preferred in complex scenarios. Participants exhibited a strong preference for AI-generated judgments in personal moral scenarios, which typically involve more deliberation and evoke stronger emotional responses (e.g., pushing one person off a bridge to save five). In contrast, for less emotionally engaging, impersonal scenarios (e.g., diverting a runaway boxcar), participants slightly favored human judgments, although this preference was not statistically significant. According to Dual Process Theory, moral judgments rely on two cognitive systems: fast, emotion-driven intuitions, and slower, deliberate reasoning [17]. In personal moral scenarios, where emotions run high, the theory predicts more engagement with deliberate reasoning. This may explain the preference for AI judgments, which might be perceived as more reasoned compared to human ones. However, this framework is debated, with some arguing that the distinction between intuitive and rational processes is not always clear-cut [28].

Participants' anti-AI bias may stem from a preference for "human-like" responses. The second major finding, concerning the influence of participants' beliefs on their agreement with moral judgments, reveals a complex interaction between belief-based and source-based biases. Participants often rejected judgments they perceived as AI-generated, reflecting an anti-AI bias. Paradoxically, the same participants showed greater agreement with LLM-generated judgments in morally challenging scenarios, indicating a pro-AI bias when it came to content (pro-AI source). These findings align with previous research [44, 57], suggesting that humans do not exhibit a simple aversion to AI [2, 6, 49]. Instead, they may favor judgments perceived as more "human-like," regardless of actual authorship.

Participants may exhibit ingroup favoritism. One explanation is that participants might initially agree with a judgment, but to avoid the cognitive dissonance of preferring an AI judgment, they unconsciously attribute it to a human source. This phenomenon may reflect an ingroup/outgroup bias, where individuals prefer to assign positive qualities to their own group (humans) over an outgroup (machines) [22]. Interestingly, LLMs themselves have been shown to exhibit similar ingroup favoritism [38]. Research in moral contexts suggests that ingroup favoritism intensifies as participants perceive their morally aligned group as embodying core values, leading to stronger biases [3]. For instance, individuals may rationalize harmful actions committed by their ingroup as justified to protect core values, while condemning the same actions by the outgroup as immoral [39]. This effect may be amplified in the context of AI-generated judgments, as participants may reject the notion that AIs can possess moral agency [2]. In our experiments, textual and linguistic features played a key role in determining ingroup identification, as these were the only cues available to participants.

5.3 Linguistic determinants of detection and alignment

Textual features predict detection but not agreement. We conducted several analyses to understand how the different sources (human, dv2, dv3, and dv2_humanized) varied from each other and how specific features influenced detection and agreement. First, we examined linguistic features such as justification length and the presence of typos. As expected, these features significantly aided human detection, based on the assumption that LLMs tend to produce longer, typo-free responses [24]. Reducing typos and shortening responses substantially lowered detection rates when evaluating corpora 3 (human vs. dv2_humanized). However, detection accuracy remained above chance, even after these modifications. Consistent with previous findings, humans demonstrated a strong tendency to use first-person discourse more frequently than LLMs [25]. Notably, neither the textual features (length, typos) nor syntactic features (first-person discourse) were linked to agreement judgments, suggesting that human moral reasoning is unaffected by changes in these basic linguistic aspects.

Semantic cues drive AI detection and moral judgment divergence. To further explore these findings, we applied SHAP interpretations from a classifier model to predict the source of the text. The analysis revealed that terms indicating structured reasoning (e.g., "additionally," "reason") were strong predictors of AI-generated content, recognized both by the model and by participants.

Interestingly, while detection and source prediction overlapped in some cases, they diverged in corpora 3 (where GPT-3.5 mimicked human responses). In these cases, humanized responses removed many typical cues. However, semantic patterns still revealed that utilitarian cues in the predictive tokens were still identifiable, even when formal textual features were diminished. Participants likely relied on these patterns to achieve modest but significantly above-chance detection rates.

Semantic patterns also revealed that utilitarian terms (e.g., "lives," "save") were associated with disagreement, particularly in personal moral scenarios like sacrificing one person for many. This suggests that participants' agreement was influenced by the perceived alignment of moral judgments with either utilitarian or deontological reasoning [16, 23, 34]. While basic textual and syntactic features affected detection, semantic elements tied to moral reasoning played a more nuanced role in agreement with judgments.

5.4 Limitations

A key limitation of our study is that the findings may be specific to GPT-3.5 and might not generalize to other models. The behavior of LLMs can vary based on their architecture and the specific version used, as different models encode distinct moral values and exhibit varying behaviors [51]. However, the alignment results from the 'corpus' generating experiments were not central to our main claim regarding how LLM judgments are detected and evaluated.

Additionally, participants' imperfect detection of AI-generated judgments may stem from linguistic factors, such as subtle differences in phrasing or style. Despite efforts to humanize LLM responses in corpora 3, participants still detected LLM-generated judgments above chance. As shown in Figure 7, they relied on (notably) first-person cues as a decision heuristic [26, 43]. Moreover, with careful prompting, AI-generated judgments can become even harder to detect, and in some cases, LLMs have been rated as more human-like or empathetic than actual human responses [25, 54].

5.5 Conclusions and perspectives

Our findings reveal a potential dissociation between participants' attribution of moral agency to AI systems and their evaluation of AI-generated moral judgments. While participants might reject the notion that AIs can act as true moral agents, as supported by previous research [2], they nonetheless find AI-generated judgments persuasive, especially in complex scenarios that challenge their own moral intuitions. This tension suggests a form of cognitive dissonance or compartmentalization, where participants maintain an anti-AI bias concerning moral agency but exhibit a pro-AI bias when practically evaluating the quality of moral judgments.

Our study further demonstrates that large language models (LLMs) exhibit human-like reasoning that can deviate from utilitarian standards depending on how the moral scenario is framed. These deviations mirror those observed in humans, and, in the case of GPT-3.5, may even be more pronounced. Notably, participants often struggled to differentiate between human and AI-generated moral justifications, raising concerns about the potential of LLMs to mislead human users. In addition to generating responses that are difficult to detect, LLMs can also be leveraged to make their outputs even harder to distinguish from human-generated content.

Moreover, human agreement with LLM justifications was influenced by the nature of the moral scenario, with participants showing a stronger preference for AI judgments in more complex scenarios. This finding suggests a possible role for LLMs as advisors or mediators in human moral decision-making. However, this pro-AI bias often occurred without participants' awareness, as higher agreement was consistently given to justifications believed to come from humans, regardless of their actual source. This discrepancy between the perceived and actual competence of human and machine judgments highlights how anti-AI biases or human chauvinism may hinder the integration of LLMs into human moral decision-making processes.

ACKNOWLEDGMENTS

The authors thank Nicolas Yax for help concerning the LLM experiment. The authors thank Michael Xieyang Liu, Lucas Dixon, and James Wexler for feedback on an early draft of the manuscript. SP is funded by the European Research Council consolidator grant (RaReMem: 101043804) and three Agence Nationale de la Recherche grants (CogFinAgent: ANR-21-CE23-0002-02; RELATIVE: ANR-21-CE37- 750 0008-01; RANGE: ANR-21-CE28-0024-01), the Alexander Von Humbolt foundation and a Google unrestricted gift.

REFERENCES

[1] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (Nov. 2018), 59–64. https://doi.org/10.1038/s41586-018-0637-6 Number: 7729 Publisher: Nature Publishing Group.

[2] Yochanan E. Bigman and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181 (2018), 21–34. Publisher: Elsevier.

[3] Ennio Bilancini, Leonardo Boncinelli, Valerio Capraro, Tatiana Celadin, and Roberto Di Paolo. 2020. "Do the right thing" for whom? An experiment on ingroup favouritism, group assorting and moral suasion. *Judgment and Decision Making* 15, 2 (March 2020), 182–192. https://doi.org/10.1017/S1930297500007336

[4] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (Feb. 2023), e2218523120. https://doi.org/10.1073/pnas.2218523120 Publisher: Proceedings of the National Academy of Sciences.

[5] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[6] Jason W. Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2020), 220–239. https://doi.org/10.1002/bdm.2155 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.2155.

[7] Robert Böhm, Moritz Jörling, Leonhard Reiter, and Christoph Fuchs. 2023. People devalue generative AI's competence but not its advice in addressing societal and personal challenges. *Communications Psychology* 1, 1 (Nov. 2023), 1–10. https://doi.org/10.1038/s44271-023-00032-x Publisher: Nature Publishing Group.

[8] Romain Cadario, Chiara Longoni, and Carey K. Morewedge. 2021. Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour* 5, 12 (Dec. 2021), 1636–1642. https://doi.org/10.1038/s41562-021-01146-0 Publisher: Nature Publishing Group.

[9] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56, 5 (Oct. 2019), 809–825. https://doi.org/10.1177/0022243719851788 Publisher: SAGE Publications Inc.

[10] Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2454–2469. https://doi.org/10.1145/3630106.3659048

[11] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. https://doi.org/10.48550/arXiv.2403.04132 arXiv:2403.04132 [cs].

[12] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. https://doi.org/10.48550/arXiv.2107.00061 arXiv:2107.00061 [cs].

[13] Fiery Cushman, Liane Young, and Marc Hauser. 2006. The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm. *Psychological Science* 17, 12 (Dec. 2006), 1082–1089. https://doi.org/10.1111/j.1467-9280.2006.01834.x Publisher: SAGE Publications Inc.

[14] Julian De Freitas, Stuti Agarwal, Bernd Schmitt, and Nick Haslam. 2023. Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour* (Nov. 2023), 1–10. https://doi.org/10.1038/s41562-023-01734-2 Publisher: Nature Publishing Group.

[15] Jaiv Doshi, Ines Novacic, Curtis Fletcher, Mats Borges, Elea Zhong, Mark C. Marino, Jason Gan, Sophia Mager, Dane Sprague, and Melinda Xia. 2024. Sleeper Social Bots: a new generation of AI disinformation bots are already a political threat. https://doi.org/10.48550/arXiv.2408.12603 arXiv:2408.12603 [cs].

[16] greene. 2004. The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron* 44, 2 (Oct. 2004), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027 Publisher: Cell Press.

[17] Joshua D. Greene, Leigh E. Nystrom, Andrew D. Engell, John M. Darley, and Jonathan D. Cohen. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44, 2 (Oct. 2004), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

[18] Joshua D. Greene, R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science* 293, 5537 (Sept. 2001), 2105–2108. https://doi.org/10.1126/science.1062872 Publisher: American Association for the Advancement of Science.

[19] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3, 10 (2023), 833–838. https://www.nature.com/articles/s43588-023-00527-x Publisher: Nature Publishing Group US New York.

[20] Claudia E. Haupt and Mason Marks. 2023. AI-generated medical advice—GPT and beyond. *Jama* 329, 16 (2023), 1349–1350. https://jamanetwork.com/journals/jama/article-abstract/2803077 Publisher: American Medical Association.

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. https://doi.org/10.48550/arXiv.2008.02275 arXiv:2008.02275 [cs].

[22] Miles Hewstone, Mark Rubin, and Hazel Willis. 2002. Intergroup Bias. *Annual Review of Psychology* 53, Volume 53, 2002 (Feb. 2002), 575–604. https://doi.org/10.1146/annurev.psych.53.100901.135109 Publisher: Annual Reviews.

[23] Keith J. Holyoak and Derek Powell. 2016. Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin* 142, 11 (2016), 1179. https://psycnet.apa.org/fulltext/2016-47729-001.html?casa_token=-4xn_jCipYcAAAAA:fCKRBrcTmMcOyKrsRa3_CGNq7KBeNuZuZ1NtCTk5UPeVUSqliap-nvcW91PQbqbdB547ZSCP9jISnEwdmfBAzRce Publisher: American Psychological Association.

[24] Piers Douglas Lionel Howe, Nicolas Fay, Morgan Saletta, and Eduard Hovy. 2023. ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology* 14 (2023), 1281255. https://doi.org/10.3389/fpsyg.2023.1281255

[25] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (March 2023), e2208839120. https://doi.org/10.1073/pnas.2208839120 Publisher: Proceedings of the National Academy of Sciences.

[26] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (March 2023), e2208839120. https://doi.org/10.1073/pnas.2208839120 Publisher: Proceedings of the National Academy of Sciences.

[27] Cameron R. Jones and Benjamin K. Bergen. 2024. Does GPT-4 pass the Turing test? https://doi.org/10.48550/arXiv.2310.20216 arXiv:2310.20216 [cs].

[28] Guy Kahane. 2012. On the Wrong Track: Process and Content in Moral Psychology. *Mind & Language* 27, 5 (2012), 519–545. https://doi.org/10.1111/mila.12001 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mila.12001.

[29] Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux* (2011). https://www.pdcnet.org/pdc/bvdb.nsf/showopenaccess?open&repid=852577BA0050DC0D&docid=2BE9C7FA6C507E4DC1257ADA0072A186&solarid=inquiryct_2012_0027_0002_0054_0057

[30] Daniel Kahneman and Amos Tversky. 1984. Choices, values, and frames. *American Psychologist* 39, 4 (1984), 341–350. https://doi.org/10.1037/0003-066X.39.4.341 Place: US Publisher: American Psychological Association.

[31] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3491102.3517533

[32] Emre Kazim and Adriano Soares Koshiyama. 2021. A high-level overview of AI ethics. *Patterns* 2, 9 (Sept. 2021). https://doi.org/10.1016/j.patter.2021.100314 Publisher: Elsevier.

[33] Mehdi Khamassi, Marceau Nahon, and Raja Chatila. 2024. Strong and weak alignment of large language models with human values. *Scientific Reports* 14, 1 (Aug. 2024), 19399. https://doi.org/10.1038/s41598-024-70031-3 Publisher: Nature Publishing Group.

[34] Michael Koenigs, Liane Young, Ralph Adolphs, Daniel Tranel, Fiery Cushman, Marc Hauser, and Antonio Damasio. 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446, 7138 (April 2007), 908–911. https://doi.org/10.1038/nature05631 Number: 7138 Publisher: Nature Publishing Group.

[35] Sarah E. Kreps, Miles McCain, and Miles Brundage. 2020. All the News that's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. https://doi.org/10.2139/ssrn.3525002

[36] Nils Köbis and Luca D. Mossink. 2021. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior* 114 (Jan. 2021), 106553. https://doi.org/10.1016/j.chb.2020.106553

[37] Daniel K. Lapsley. 2018. *Moral psychology*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9780429498824/moral-psychology-daniel-lapsley

[38] Walter Laurito, Benjamin Davis, Peli Grietzer, Tomáš Gavenčiak, Ada Böhm, and Jan Kulveit. 2024. AI AI Bias: Large Language Models Favor Their Own Generated Content. https://doi.org/10.48550/arXiv.2407.12856 arXiv:2407.12856 [cs].

[39] Bernhard Leidner, Emanuele Castano, Erica Zaiser, and Roger Giner-Sorolla. 2010. Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality & Social Psychology Bulletin* 36, 8 (Aug. 2010), 1115–1129. https://doi.org/10.1177/0146167210376391

[40] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (March 2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[41] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[42] Alexander Marrapese, Basem Suleiman, Imdad Ullah, and Juno Kim. 2024. A Novel Nuanced Conversation Evaluation Framework for Large Language Models in Mental Health. http://arxiv.org/abs/2403.09705 arXiv:2403.09705 [cs].

[43] Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model

for Detecting Short ChatGPT-generated Text. https://doi.org/10.48550/arXiv.2301.13852 arXiv:2301.13852 [cs].

[44] Carey K. Morewedge. 2022. Preference for human, not algorithm aversion. *Trends in Cognitive Sciences* 26, 10 (Oct. 2022), 824–826. https://doi.org/10.1016/j.tics.2022.07.007 Publisher: Elsevier.

[45] Wim De Neys. 2006. Dual Processing in Reasoning: Two Systems but One Reasoner. *Psychological Science* 17, 5 (May 2006), 428–433. https://doi.org/10.1111/j.1467-9280.2006.01723.x Publisher: SAGE Publications Inc.

[46] Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. 2023. "Correct answers" from the psychology of artificial intelligence. https://doi.org/10.48550/arXiv.2302.07267 arXiv:2302.07267 [cs].

[47] Sebastian Proksch, Julia Schühle, Elisabeth Streeb, Finn Weymann, Teresa Luther, and Joachim Kimmerle. 2024. The impact of text topic and assumed human vs. AI authorship on competence and quality assessment. *Frontiers in Artificial Intelligence* 7 (May 2024). https://doi.org/10.3389/frai.2024.1412710 Publisher: Frontiers.

[48] Shahzad Qaiser and Ramsha Ali. 2018. Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications* 181, 1 (2018), 25–29. https://www.researchgate.net/profile/Shahzad-Qaiser/publication/326425709_Text_Mining_Use_of_TF-IDF_to_Examine_the_Relevance_of_Words_to_Documents/links/5b4cd57fa6fdcc8dae245aa3/Text-Mining-Use-of-TF-IDF-to-Examine-the-Relevance-of-Words-to-Documents.pdf

[49] Md Jabir Rahman, Huigang Liang, and Yajiong Xue. 2023. AI Aversion: A Task Dependent Multigroup Analysis. *PACIS 2023 Proceedings* (July 2023). https://aisel.aisnet.org/pacis2023/86

[50] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108 arXiv:1910.01108 [cs].

[51] Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the Moral Beliefs Encoded in LLMs. http://arxiv.org/abs/2307.14324 arXiv:2307.14324 [cs].

[52] Walter Sinnott-Armstrong. 2008. Framing moral intuitions. (2008). https://psycnet.apa.org/record/2007-14533-005 Publisher: Boston Review.

[53] Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges. https://doi.org/10.48550/arXiv.2403.18249 arXiv:2403.18249 [cs].

[54] Anuradha Welivita and Pearl Pu. 2024. Are Large Language Models More Empathetic than Humans? https://doi.org/10.48550/arXiv.2406.05063 arXiv:2406.05063 [cs].

[55] Nicolas Yax, Hernán Anlló, and Stefano Palminteri. 2024. Studying and improving reasoning in humans and machines. *Communications Psychology* 2, 1 (June 2024), 1–16. https://doi.org/10.1038/s44271-024-00091-8 Publisher: Nature Publishing Group.

[56] Yunhao Zhang and Renée Gosline. 2023. Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making* 18 (2023), e41. https://www.cambridge.org/core/journals/judgment-and-decision-making/article/human-favoritism-not-ai-aversion-peoples-perceptions-and-bias-toward-generative-ai-human-experts-and-humangai-collaboration-in-persuasive-content-generation/419C4BD9CE82673EAF1D8F6C350C4FA8 Publisher: Cambridge University Press.

[57] Yuyan Zhang, Jiahua Wu, Feng Yu, and Liying Xu. 2023. Moral Judgments of Human vs. AI Agents in Moral Dilemmas. *Behavioral Sciences (Basel, Switzerland)* 13, 2 (Feb. 2023), 181. https://doi.org/10.3390/bs13020181

[58] Jian-Qiao Zhu, Haijiang Yan, and Thomas L. Griffiths. 2024. Language Models Trained to do Arithmetic Predict Human Risky and Intertemporal Choice. https://doi.org/10.48550/arXiv.2405.19313 arXiv:2405.19313 [cs, econ, q-fin].

Received 12 September 2024