Framing an AI with Values Reduces AI Reliance in AI-supported Writing Tasks

ALICE GAO, University of Washington, USA
ANDREW N. MELTZOFF, University of Washington, USA
MAARTEN SAP, Carnegie Mellon University, USA
KATHARINA REINECKE, University of Washington, USA

Despite a global user base adopting large language models (LLMs) for daily writing tasks, model suggestions tend to align with Western values. Research has shown users commonly accept a high fraction of these AI suggestions, homogenizing writing styles and rendering outputs more "Western" than intended. While this suggests a need to reduce AI reliance, it remains unknown what kind of interventions could achieve this. Can framing the AI with specific values, and comparing it to one's own, make users less susceptible to overreliance and support more unique writing? We tested this hypothesis in a between-subjects online experiment with Indian and American participants (n=149) in which they were asked to perform AI-supported writing tasks, either 1) without an intervention, 2) after seeing an overview of the AI's framed values, or 3) after seeing an overview of the AI's framed values compared to their own. Our results show that seeing the AI's framed values reduces AI reliance, i.e., the proportion of the final essay generated by the AI, by an average of 20%. Additionally, when participants saw an overview of the AI's framed values (without comparison to their own values), the final essays contain more unique text than without intervention. Our findings emphasize the importance of educating users about potential value biases in AI, showing that raising awareness with a simple overview of values encourages users to personalize their writing.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: NLP, human-AI interaction, AI reliance, culture

ACM Reference Format:
Alice Gao, Andrew N. Meltzoff, Maarten Sap, and Katharina Reinecke. 2026. Framing an AI with Values Reduces AI Reliance in AI-supported Writing Tasks . In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3805689.3812369

1 Introduction

Large language models (LLMs) have become one of the most widespread AI applications with many people using it to fully generate text or autocomplete sentences in emails [43], scientific writing [32], or journalism [65]. Despite a global uptake of AI-supported writing systems, models remain primarily trained on English language and Western-centric data [15, 16, 41], frequently leading their outputs to align more with Western values [71]. Research has found that users of AI-supported writing systems can perceive these suggestions as inadequately representing their culture, language, and sociolect, feeling that these systems were not designed with them in mind [12]. AI suggestions have also been shown to change a users' language [39] to produce more

Authors' Contact Information: Alice Gao, University of Washington, Seattle, USA, atgao@cs.washington.edu; Andrew N. Meltzoff, University of Washington, Seattle, USA, meltzoff@uw.edu; Maarten Sap, Carnegie Mellon University, Pittsburg, USA, msap2@andrew.cmu.edu; Katharina Reinecke, University of Washington, Seattle, USA, reinecke@cs.washington.edu.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812369

generically positive outputs [3, 4, 39], shifting their attitudes about social topics [40], and influencing which topics individuals write about [66]. Moreover, because people tend to accept a high fraction of AI suggestions, the output of AI-supported writing tends to be in line with the AI's value biases, leading to more homogenized, generic outputs than if people were not using AI [1, 2, 4].

This value mismatch can cause harms during human-AI interaction, as the AI may continue to promote values at odds with users' diverse experiences, leading to the loss of agency or even contribute to the erasure of unique cultures, languages, or sociolects [12, 33]. This homogenization remains an issue as it may erode distinct cultural practices and creative expressions, such as when AI is used in story-writing interfaces [52, 76, 90], or alter cultural preferences and reinforce Western values. The human tendency to engage in *satisficing*, i.e., choosing the first reasonable option presented because it saves time and cognitive effort to do so [62, 75], further exacerbates the potentials of these harms during AI-supported writing.

Satisficing has also been demonstrated in AI-supported decision making [8, 21, 81], where users tend to overrely on AI systems to distinguish correct and wrong answers. While intervention strategies have been developed and shown to successfully reduce overreliance in AI-supported decision making, such as providing explanations [82] or nudging users to engage analytically with an AI before they can make their final decision [21], we have limited knowledge of what types of interventions can reduce overreliance or overacceptance of AI suggestions during AI-supported writing.

In this work, we address this gap by evaluating whether simply making users aware of an AI's potential values can make them less susceptible to relying on the AI's suggestions and lead to more diverse and unique texts. Our main research question asks: **Does framing an AI writing assistant with specific values affect user reliance and the extent to which they add their own unique words to the output?**

In a between-subjects online study with 149 participants from India and the US, we evaluated this question by asking participants to perform two co-writing tasks with an AI[1] with two types of interventions, a non-comparative and comparative overview of the framed AI values. We tested these interventions across four conditions: (1) no intervention (baseline), (2) after seeing an overview of the framed AI values (AI values), (3) after considering their own values and seeing an overview of the framed AI values (primed + AI Values), or (4) after considering their own values and seeing an overview the framed AI values in comparison to their own (primed + comparative AI values).

Our results indicate that providing an overview of specific AI values reduces user reliance on the AI by an average of 20% and leads to greater uniqueness in responses, but only when participants did not see these values in comparison to their own. In line with prior work [1], our Indian participants generally had a much higher AI reliance and accepted more AI suggestions than Americans. However, as we hypothesized, seeing the framed AI's values resulted in lower AI reliance compared to the baseline for both Indian and American participants. As a result, the output text tended to be more diverse and unique after seeing the intervention, suggesting that awareness of potential value biases (in a non-comparative overview) encourages users to put more effort into contributing their own unique words. We conclude by discussing the implications of designing effective intervention strategies during AI co-writing and the importance of surfacing an AI's value biases to users.

2 Related Work

We contextualize our work by first providing an overview of AI overreliance and mitigation strategies. We then provide a background of the biases and values within AI and the influence of these biases.

[1]Though our study specifically investigates LLMs and their influence on writing, we framed our intervention to participants as potential values an AI may have. Furthermore, lay users typically think of AI more broadly instead of LLMs. As such we use the terms interchangeably throughout the paper.

2.1 Reliance and Interventions

As LLMs and AI assistants have become integrated into multiple aspects of our lives, users have also become increasingly *reliant* on LLMs, accepting automated decisions as a means of reducing cognitive effort and saving time, or satisficing [81], even when the LLM is wrong [21, 28]. Researchers have studied and designed potential strategies to reduce overreliance that arises due to satisficing, particularly during AI assisted decision-making, such as providing explanations at how an LLM arrived at its decisions [14, 82], indications of the AI's certainties [9, 50], delaying the delivery of AI recommendations [64], or asking individuals to first make a decision before engaging with the AI's decision [28, 36].

These studies reveal a complex relationship between interventions and reducing overreliance: explanations of an AI's decisions or recommendations alone do not reduce overreliance [9, 21, 82]. In fact, explanations can actually increase overreliance as their presence may increase trust in LLMs [91, 92]. Instead, individuals must be forced to actively engage with these explanations [21], motivated by external benefits [82], or view engaging with LLM explanations as less difficult than the task at hand [82]. To add to this complexity, prior work has repeatedly found that the context and framing of AI explanations can increase reliance. Using more anthropomorphic language [93] or expressing confidence in responses [68] increases user reliance and trust, as does the framing of the AI assistant as an expert or companion part of one's group [11, 23] during human-AI interactions. This suggests that reducing overreliance is a complicated task that must be approached with care, taking the user's prior experience and perceived effort as well as the framing of the interaction with the AI into account. Most overreliance-reducing strategies focus on reducing AI reliance during *decision-making*, where users distinguish correct and incorrect choices, and less so when individuals collaborate with embedded or interactive AI tools where they seek LLM outputs to be more productive, producing writing for work or for personal tasks, requiring uninterrupted flow. Our work seeks to fill this gap by building upon existing work in reducing overreliance through developing a non-disruptive intervention with interactive AI tools during AI-supported writing.

2.2 Value Biases and the Influence of AI

Despite the pervasiveness of LLM-powered tools, these models embed the biases of their training data, which is frequently Western-centric [15, 41]. As a result, these tools are limited in their ability to serve diverse user groups [15], and produce better results for well-represented groups [53, 85]. Even with guardrails to remove explicit biases, the outputs of LLMs are embedded with *implicit* biases learned from their data [17, 54]. Prior work has shown how LLMs flatten nuances into homogeneous outputs [53], portraying racial minorities as one-dimensional or providing a Western-centric lens for these groups [1, 53]. Other work has shown how these biases can lead to minorities feeling excluded and unaccounted for when LLMs fail to recognize diverse cultures, languages, or sociolects [12, 15, 16, 30, 31] resulting in poor user experience and frustration [15, 24, 53].

Researchers have studied the implications of using these tools as their ubiquity has increased. Work has shown how AI suggestions can influence a user's language [39] and result in generic, and overly positive text [3, 4], while Agarwal et al. [1] showed that Indian participants who used AI-supported writing produced more homogenized writing, lacking nuances and portraying cultural artifacts from a *Western* lens. Jakesch et al. [40] also showed how writing suggestions from opinionated models are a form of *latent persuasion* that can shift users' social attitudes.

Though work has shown simply informing people of biases can mitigate their influence [37, 51, 67], this problem becomes much more complex with LLMs with their known inconsistencies [70]. LLMs, just like humans, are inconsistent with their values: the values a person may indicate may not be the values they adhere to in their own actions and tasks [79, 83] and similarly, probabilistic models are inconsistent in value expression in value-laden situations [45, 57]. Taking this into account, our work investigates the impact of framing an AI writing assistant with specific values rather than attempting to elicit values that may be inconsistent.

3 Methods

We conducted a controlled online experiment to determine whether framing an AI with specific values affects participants' interaction with the LLM and the generated output. We adapted Agarwal et al. [1]'s study which found that Indian and American participants heavily rely on LLM writing suggestions. We added the following interventions to test whether showing participants an AI's framed values could reduce this reliance:

(1) **No Intervention (baseline):** Our baseline where participants do not see an AI's framed values.
(2) **AI Values:** Participants view an AI's framed values before the writing tasks and values survey.
(3) **Primed + AI Values:** Participants take the values survey *then* view an AI's framed values before completing the writing tasks.
(4) **Primed + Comparative AI Values (comparative):** Participants take the value survey then view their values *compared* to an AI's framed values before completing the writing tasks.

We manipulated the type of overview (non-comparative vs. comparative) participants saw as we hypothesized that a comparison provides better understanding [26] of the AI values shown. People often draw comparisons with relevant individual factors [18, 27, 89] to process information in the world. We hypothesized that an increased understanding would decrease participants' reliance. We tested two timings of the non-comparative overview, primed + AI values and AI values, as the former may cause participants to implicitly compare their own values to the framed AI values. Table 1 provides an overview of our experimental procedure.

Table 1. Study flow for four participant groups. No intervention, which is our baseline, where they did not view the AI's framed values. AI values and primed + AI both saw the AI's framed values but in the former they viewed it before the writing tasks. Primed + Comparative AI values (comparative) condition participants viewed their values in comparison to the framed AI values shown.

No Intervention	**AI Values**	**Primed + AI Values**	**Primed + Comparative AI Values**
1. Demographic Survey	1. Demographic Survey	1. Demographic Survey	1. Demographic Survey
2. ***Values Survey***	2. **AI Value Overview**	2. ***Values Survey***	2. ***Values Survey***
3. Writing Task 1	3. Writing Task 1	3. **AI Value Overview**	3. **Comparative AI Value Overview**
4. Writing Task 2	4. Writing Task 2	4. Writing Task 1	4. Writing Task 1
5. AI Literacy Survey	5. ***Values Survey***	5. Writing Task 2	5. Writing Task 2
	6. AI Literacy Survey	6. AI Literacy Survey	6. AI Literacy Survey

3.1 Hypotheses

We established a set of hypotheses to answer our main research question:

Hypotheses 1: Effect on AI Reliance. We hypothesized that seeing an AI's framed values would increase user understanding that an AI can have values and biases, and that this understanding would reduce AI reliance and acceptance of its suggestions. Furthermore, we hypothesized that greater understanding of the difference between an AI's framed values and one's own would lead to even further decreases of AI reliance.

> **H1a:** Users that see an AI's framed values (AI values, primed + AI values, comparative) will have a lower AI reliance and AI acceptance rate compared to users that do not see an AI's framed values (baseline).
>
> **H1b:** Users that see an AI's framed values in *comparison* to their own (comparative) will have a lower AI reliance and AI acceptance rate compared to users that just see an AI's framed values (AI values, primed + AI values).

Additionally, we hypothesized that the greater the value difference between a participant and the selected AI values, the lower their AI reliance and AI acceptance rate:

> **H1c:** There exists a negative correlation between an individual's value difference with an AI's framed values and their AI reliance and AI acceptance rate.

Hypotheses 2: Effect on Uniqueness. We also hypothesized that framing an AI with specific values and an increased understanding would lead users to making their texts more unique, or contributing more of their own words. The understanding would implicitly promote users to counteract these biases through individualizing their responses. We also hypothesized that greater understanding would further increase this uniqueness.

> **H2a:** Users that see the AI's framed values (AI values, primed + AI values, comparative) will produce more unique texts compared to users that do not see the AI's framed values (baseline).
> **H2b:** Users that see the AI's framed values in *comparison* to their own (comparative) will produce more unique texts compared to users that just see the AI's framed values (AI values, primed + AI values).

Because prior work has shown that Indian users tend to have a higher reliance on AI than American users [1], we additionally explored whether the effect of the interventions differed between the two participant groups.

3.2 Recruitment & Participants

Participants were recruited from Prolific, an online platform tailored for academic research [63], enabling us to target specific participant groups. We recruited participants over 18, who had English fluency, and lived either in the United States, for Americans, or India, for Indians, and nowhere abroad for more than 6 months. A priori power analysis conducted using R's `pwr` package and indicated a minimum of 31 participants per group for four groups within a moderate effect size (0.3), at a power of 0.80 and α of 0.05. We aimed to recruit 40 participants per condition with an equal number of Indian (n=20) and American (n=20) participants but only analyzed the data of participants that provided valid responses. Responses were considered invalid if they did not answer the correct writing prompts. In total, we had 149 participants with valid responses: 40 participants formed the baseline group, 39 participants in AI values condition, 34 participants in primed + AI values condition, and 36 participants in comparative condition. 76 participants were from the United States and 73 participants were from India. We provide the full list of self-reported participant demographics in Table 6.

The study protocol was reviewed and received exempt status by the first author's university's Institutional Review Board (IRB). Participants were compensated at an hourly rate of $12 an hour for study completion.

3.3 Study Procedure

Participants were redirected to our custom study portal, built using SvelteKit and hosted at a public URL. Upon entering our study portal, participants first confirmed their consent and provided their unique Prolific participant ID. They then took a short demographic survey. After, participants proceeded to either a 15 question values survey or saw the AI's framed values. Table 1 provides the experimental flow for all four conditions.

Following Agarwal et al.'s procedure [1], participants were required to complete two writing tasks, shown sequentially, with a 50 word minimum. Once participants reached the minimum word requirement, they could proceed to the next step of the study. They were shown their word count in the bottom as they typed. To encourage participants to not solely rely on the LLM's suggestions, the initial Prolific recruitment framed the goal of the study as a way to collect personal stories from around the world. We show our interface in Fig. 5.

All participants finished with a four question AI literacy survey. Upon completion, they were provided a code they could enter into Prolific to verify their completion and receive monetary compensation. All survey answers, final written responses, and writing task interaction logs were all logged to an external database.

3.4 Materials

3.4.1 Intervention Strategy: an AI Value Overview. We created an *AI value overview* to test whether users change their propensity to accept AI suggestions once they have seen an *AI's framed values*. To do so, we first required values for the AI. Although we considered using the Short Schwartz Values Survey [55] or Inglehart-Welzel's 10 questions [88] from the World Values Survey (WVS) [38], we chose to manually selected a subset of questions from the WVS to circumvent the Western biases and criticism Eurocentric value constructs found in these questions [7, 19], as well as their limited applicability for other cultural groups [19]. As the WVS is over 200 questions, we selected a subset of 15 questions across its 14 thematic subsections based on concreteness, understandability, and ability to provide meaningful comparison to a user's own values. These questions were chosen by the first author then discussed and iterated upon in meetings with the entire team until consensus. Our questions span four value dimensions: social, trust, security, and religion. These value dimensions were obtained through the WVS thematic category that questions were selected from. To obtain the "AI's values," we used OpenAI's `GPT-4o` as our "AI" and queried it on our subset of questions and answer choices for at least 10 trials per question. The average of Likert questions and the majority answer for multiple choice questions were taken as the AI's framed values. In case of ties, we queried the AI an additional time. The full list of questions can be found in Appendix A.1.

We provided the specified values as a bar chart as it is a common, easy to read visualization [48]. The same AI value overview was shown to all participants. Fig. 1 shows a portion of this overview and the comparative version, where participants saw their values compared to the AI's. We added two conditions that showed the intervention at different points in the study as seeing the AI's framed values after taking the values survey could cause indirect comparison of participants' own values to the framed values.

Though prior work has characterized the biases embedded in AI and identified potential misalignments with different user groups [71, 80], capturing an AI's "values" remains an open question [46, 74]. Our AI value overview was not designed to perform the impossible task of fully reflecting the AI's values, but to signal to users that an AI had these specified values and frame their interaction with the AI through the lens of a value-laden AI. It is possible that the AI suggestions did not align with the values in our overview. As a way to assess this, we performed an additional analysis to determine the overlap between the suggestions shown in the study and the framed AI values (Appendix E). We found that a majority of suggestions were neutral towards the framed AI values. We further discuss the limitations of our approach in §5 and §6.

3.4.2 Writing Task. Each participant completed two writing tasks randomly chosen from two categories: (1) writing tasks where AI suggestions tend to exhibit Western bias and (2) writing tasks where AI suggestions exhibit less Western bias. These tasks were adapted from Agarwal et. al's [1], meant to elicit explicit cultural practices and more nuanced values. For the first category, participants wrote about their favorite food or holiday. For the second category, participants wrote about their favorite celebrity/public figure or asked their boss for a two-week leave. We provide all prompts in Table A.2.2. For each writing task, participants were able to accept or reject AI autocomplete suggestions.

We adopted Agarwal et. al's [1] procedure to retrieve autocomplete suggestions from `GPT-4o`, the latest OpenAI model at the time. Autocomplete suggestions were shown as light gray ghost text, similar other autocomplete suggestions text-based editors like Gmail, and limited to 10 words. We embedded the essay topic in the prompt and the participant's current input into the prompt to provide contextually relevant suggestions [1, 22, 40]. Suggestions were retrieved every 500ms, like in prior work [22, 40], and could be accepted by pressing the `Tab` button or ignored by continuing to type. To prevent participants from rapidly accepting suggestions we required that they type another letter after accepting an AI's suggestion. The full prompt to retrieve suggestions is provided in Table A.2.1.

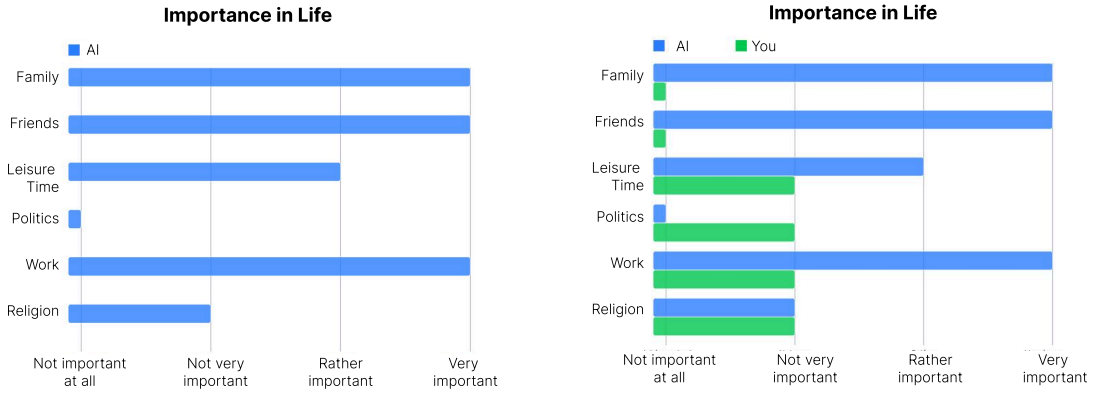


Fig. 1. A snapshot of our intervention, showing the AI's framed values we showed participants with the AI's answer to ranking six items (family, friends, leisure time, politics, work, and religion) on the importance in its life. The left is shown to the non-comparative interventions (AI values and primed + AI values) while the right shows the participant's values compared to the AI's (comparative).

3.5 Analysis

We analyzed three sets of data: interaction logs captured on the study portal, the final essays written by the participants, and personal factors. We summarize the metrics used in our analyses below.

3.5.1 AI Reliance. From writing task interaction logs, we used following metrics to assess an individual's AI reliance:

(1) **AI Reliance:** The proportion of the final essay generated by the AI [1, 22, 42], computed as the ratio of total number of AI-suggested characters in the final essay to the total number of characters in the essay. The longest common subsequence (LCS) was used to handle scenarios in which a suggestion was modified. It ranges from 0 to 1, where a higher value indicates higher AI reliance or more AI-generated content. This metric is sensitive to suggestions that were accepted but later modified or deleted.
(2) **AI Acceptance Rate:** The proportion of accepted AI suggestions to the total number of AI suggestions shown in a task [1, 22] and a metric of engagement with the AI. It ranges from 0 to 1 and is a coarse measurement as users can later delete or modify suggestions.
(3) **AI Suggestion Modification Rate:** The proportion of AI accepted suggestions that were modified to the total number of AI suggestions [22]. It ranges from 0 to 1, and a higher number suggests that the users tended to modify AI suggests to better suit individual needs. To capture modification, we checked if an accepted suggestion appeared in the final essay.

We included AI suggestion modification rate in our analysis following prior work [1, 22, 25] though it is not formally in our hypotheses as this metric is partially dependent on the AI acceptance rate.

3.5.2 Writing Uniqueness. We used methods from NLP literature and qualitative analysis to analyze uniqueness:

(1) **Lexical Diversity:** We used type-token ratio (TTR) as a coarse measurement of lexical diversity in text [6], calculated as the proportion of the number of unique words (types) to the total number of words (tokens). The metric is bounded from 0 to 1 where 0 represents more repetition and less linguistic diversity. TTR has the weakness that it penalizes longer texts.
(2) **Cosine Similarity:** We retrieved embeddings from OpenAI's `text-embedding-large` model to calculate cosine similarity scores between pairs of essays on the same topic with all other users. An individual's

similarity score was computed by computing cosine similarities to all same-topic essays with other participants within their condition. Each participant wrote two essays so the scores were averaged for a final similarity score. Similarities ranged from -1 to 1, where -1 represents completely opposite essays and 1 represents essays that are exactly the same. This provides a rough measurement for textual similarity but fails to consider the context of words or sentences.

(3) **Thematic Analysis:** We conducted a thematic analysis [20] on participant essays as computational metrics provide coarse-grained measures to analyze written text. AI-supported writing is known to homogenize text [39], making it more positive [3], so we seek to capture nuances undetected through automatic metrics.

We do not report on writing productivity, though it was collected in previous studies [1, 25], as it not indicative of writing uniqueness.

3.5.3 Personal Factors. We collected two measures for an individual's personal factors:

(1) **AI Value Difference:** We calculated the value difference between a participant's own values and the AI's framed values from our overview:

$$\text{Value Difference} = \frac{\sum_{A=0}^{n=15} |A_{user} - A_{AI}|}{\max \sum_{A=0}^{n=15} |A_{user} - A_{AI}|}$$

where A_{user} and A_{AI} are the user and AI's answer to the current question respectively. Non-Likert questions were encoded binary variables where 1 represented the AI's answer and occurrence of this answer and 0 otherwise. The value difference was normalized to the range of 0 (the same) to 1 (most dissimilar) for better interpretability. We used this as an exploratory measure due to the difficulties in capturing an AI's framed values, as their approach to answering questions is drawn from their training data [29, 74]. Furthermore, AI responses have not been shown to correlate to real-world behavior unlike human responses to these surveys [5]. We collected the measure for all participants but only examined it for those that received an intervention.

In our exploratory analysis, we further break down the overall value difference into 4 dimensions (*social, trust, security, and religion)* based on the thematic section from the WVS they were selected from.

(2) **AI Literacy:** We also collected an individual's AI literacy level and reported on this metric, as this has been shown to has been shown to impact their perception of AI [58]. Our AI literacy survey was adopted from prior work [58, 59, 86]. This survey measures a participant's AI literacy on a 7-point Likert scale and captures a participant's AI literacy across 3 dimensions: *awareness* (participants' awareness of AI in technology and ability to identify them), *usage* (how frequently participants use AI for work and non-work tasks), and *ethics* (participants' awareness of the limitations and shortcomings of AI technology).

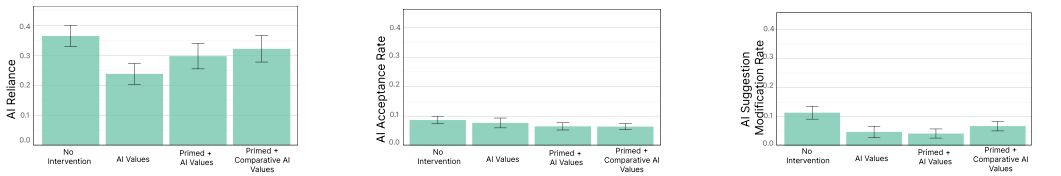
4 Findings

4.1 Framing an AI with specific values reduces an individual's AI reliance

To test if seeing the AI's framed values in each of our intervention conditions (**H1a**) reduces AI reliance and AI acceptance rate compared to the baseline, we conducted three Bonferroni-corrected one-tailed t-tests ($\alpha = .0167$) for each metric. Across conditions, reliance reduced 20.1% on average when seeing the AI's framed values compared to the baseline. Fig. 2 shows our results. Reliance was significantly lower for those who saw the AI's framed values before the writing task (AI values: $M = 0.238, SD = 0.309$) compared to the baseline ($M = 0.364, SD = 0.317, t(156.00) = -2.537, p < .01, d = -0.404$, 95%$CI$ $[-\infty, -0.044]$), representing a 34.6% decrease. Though results did not show that the AI acceptance rate was significantly lower for those that saw the AI's values compared to the baseline, descriptive differences indicated that their acceptance rate was lower (AI values: $M = 0.077, SD = 0.148$; primed + AI values: $M = 0.065, SD = 0.101$; comparative: $M = 0.065, SD = 0.089$) compared to the baseline ($M = 0.087, SD = 0.115$). We additionally conducted three Bonferroni-corrected paired t-tests to

analyze differences in the AI suggestion modification rate and observed that those who took the values survey and saw the AI's values before the writing task modified less suggestions (primed + AI values: $M = 0.042, SD = 0.129$) compared to the baseline ($M = 0.114, SD = 0.199; t(137.16) = -2.631, p < .01, d = -0.420, 95\% CI\ [-0.125, -0.018]$). We note that the lowered AI suggestion modification rate across all intervention conditions may actually be due to the lowered AI acceptance rates. Our results partially confirm our hypothesis, showing that AI reliance decreases for participants that saw the AI's framed values before writing tasks in the AI values condition.

To test whether the those that saw the AI's framed values compared to their own had less reliance compared to the non-comparative interventions (**H1b**), we conducted two Bonferroni-corrected ($\alpha = .25$) one-tailed t-tests for each of our AI reliance metrics. We did not observe significant differences between participants that received comparative or non-comparative interventions, disconfirming our hypothesis. We provide full results for **H1a** and **H1b** in Tables 7 and 8. Overall, our results suggest that **seeing the AI's framed values impacts a user's AI reliance the most when these values are non-comparative**.



(a) AI Reliance (b) AI Acceptance Rate (c) AI Suggestion Modification Rate

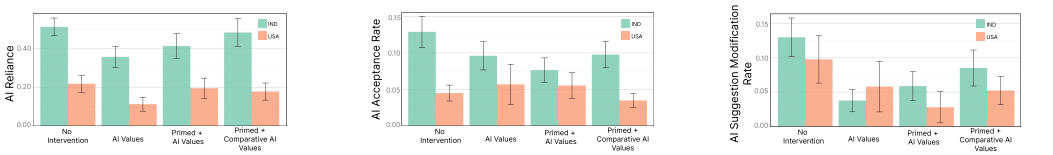
Fig. 2. AI reliance metrics across our different study conditions. We observe a decrease in the AI reliance metrics (a) AI reliance, (b) AI acceptance rate, and (c) AI suggestion modification rate for all intervention conditions. The AI values group that saw the non-comparative intervention before writing tasks had the largest decrease across all metrics.

4.1.1 Framing an AI with specific values reduces reliance in different demographic groups. Based on our mixed results from the previous section and on prior work [1], we performed an exploratory analysis to examine the effect of our interventions for Indian and American participants, as they are known to have different baseline behaviors with AI. We first conducted an independent samples t-test between Indian and American participants to test for baseline differences in behavior. Our results confirm prior work [1], showing that Indian participants have significantly higher AI reliance ($M = 0.440, SD = 0.352; t(272.68) = 7.211, p < .0001, d = 0.840, 95\%\ CI\ [0.191, 0.335]$) and accept more AI suggestions ($M = 0.101, SD = 0.117; t(291.85) = 4.059, p < .0001, d = 0.471, 95\%\ CI\ [0.027, 0.079]$) though they modify a similar amount of suggestions ($M = 0.079, SD = 0.145$) compared to American participants (AI reliance: $M = 0.177, SD = 0.271$; AI acceptance rate: $M = 0.048, SD = 0.109$; AI suggestion modification rate: $M = 0.060, SD = 0.184$).

To analyze whether the interventions led to decreased AI reliance (**H1a**) within each demographic group, we calculated the relative change between each condition and the baseline, reporting their Fieller's confidence intervals. Full results are provided in the Appendix in Table 9 while Fig. 3 shows the results per condition stratified by country.

Our results demonstrated that seeing an AI's framed value decreases reliance for both groups. For Indian participants, we observed a decrease in both the AI acceptance rate (AI values: $M = 0.096, SD = 0.122, -25.5\%$; primed + AI values: $M = 0.076, SD = 0.094, -40.6\%$; comparative: $M = 0.098, SD = 0.106, -24.1\%$) and AI reliance up to *30%* (AI values: $M = 0.357, SD = 0.336, -30.1\%$; primed + AI values: $M = 0.412, SD = 0.363, -19.3\%$; comparative: $M = 0.382, SD = 0.412, -5.60\%$) for all intervention groups compared to the baseline (AI acceptance rate: $M = 0.129, SD = 0.136$; AI reliance: $M = 0.510, SD = 0.289$). Though the AI acceptance rate for US participants that saw non-comparative AI values (AI values: $M = 0.055, SD = 0.106, +26.4\%$; primed + AI

values $M = 0.035, SD = 0.057, +22.8\%$) conditions increased, we saw a decrease in reliance across all conditions (AI values: $M = 0.113, SD = 0.220, -48.1\%$; primed + AI values: $M = 0.195, SD = 0.311, -10.6\%$, comparative: $M = 0.178, SD = 0.271, -18.4\%$) compared to the baseline ($M = 0.219, SD = 0.277$). This suggests that even with the increased acceptance rate for US participants, they continued to add their own text, meaning they unique their writing more leading to an overall reduction in AI reliance. Overall, we noticed the same trend that **AI reliance is reduced the most when participants receive a non-comparative intervention.**



(a) AI Reliance (b) AI Acceptance Rate (c) AI Suggestion Modification Rate

Fig. 3. AI reliance metrics across our different intervention conditions within each country. Though the (b) AI acceptance rate increases slightly for American participants in two conditions, participants that receive an intervention have an overall decreased (a) AI reliance. This trend holds for Indians and Americans participants.

4.2 Personal factors influence individual AI reliance

We conducted an exploratory analysis to evaluate the impact of personal factors (**H1c**), such as value differences with an AI's framed values, on AI reliance, fitting multiple linear regression models for participants that received an intervention. We additionally included an individual's AI literacy background, broken down by dimension as well as linear regression models only run on value differences with an AI's framed values, stratified by dimension. As shown in Table 4.2, an individual's AI value difference and AI literacy levels can predict their AI reliance and AI acceptance rate.

We observed that an individual's value difference with the framed AI's values is negatively associated with both AI reliance ($\beta = -0.025, p < .001$) and AI acceptance rate ($\beta = -0.009, p < .001$) and this is primarily driven by larger differences in social values (AI reliance: $\beta = -0.154, p < .0001$; AI acceptance rate: $\beta = -0.051, p < .001$). Similarly, greater awareness of the limitations of AI, AI ethics, is negatively associated with an individual's AI reliance ($\beta = -0.049, p < .05$). Overall, these results partially confirm our hypothesis, showing that **an individual's background can predict their AI reliance and acceptance rate.**

4.3 Framing an AI with specific values leads to more unique text

To test whether participants who saw an AI's framed values (**H2a**) produced more unique texts, or contributed more individual content to their responses, compared to the baseline, we conducted three Bonferroni-corrected one-tailed tests for our writing metrics. Fig. 4.3 shows our mixed results: we did not observe a significant increase in the lexical diversity between the three different interventions and the baseline group.

When examining text diversity using cosine similarity, we observed that participants that saw the AI's values before the writing tasks had the least similar essays to one another (AI values: $M = 0.462, SD = 0.129$) compared to the baseline ($M = 0.531, SD = 0.161; t(1575.13) = -10.008, p < .0001, d = -0.477, 95\%\ CI\ [-\infty, -0.057]$). We did not observe significantly lower cosine similarities for the other conditions (primed + AI values: $M = 0.581, SD = 0.129$; comparative: $M = 0.596, SD = 0.116$). Our results partially confirm that seeing the AI's values leads to increased uniqueness.

To test whether participants in the comparative condition had more unique texts than the other intervention conditions (**H2b**), we conducted two Bonferroni-corrected ($\alpha = .25$) one-tailed t-tests for each metric. We did not

Table 2. Coefficients, standard errors, and significance of personal factors used to predict our AI reliance metrics (AI reliance, AI acceptance rate, AI suggestion modification rate) from our linear regression. From Panel A: we see that an individual's value difference with the AI's framed values and their AI ethics can serve as predictors for two metrics, AI reliance and AI acceptance rate. From Panel B: we further see that a higher social value difference with an AI's framed values serves as a predictor for AI reliance and AI acceptance rate.

Panel A: AI literacy dimensions and overall AI value difference as predictors of AI reliance metrics

Predictor	AI Reliance	AI Acceptance Rate	AI Suggestion Modification Rate
(Intercept)	0.560^{***} (0.160)	0.186^{***} (0.055)	0.102 (0.074)
AI Awareness	0.041 (0.021)	−0.005 (0.007)	−0.011 (0.010)
AI Usage	0.027 (0.016)	0.007 (0.006)	0.004 (0.008)
AI Ethics	-0.049^{*} (0.019)	0.001 (0.007)	−0.007 (0.009)
AI Value Difference	-0.025^{****} (0.006)	-0.009^{****} (0.002)	0.002 (0.003)

Panel B: Dimensions of AI value difference as predictors of AI reliance metrics

Predictor	AI Reliance	AI Acceptance Rate	AI Suggestion Modification Rate
(Intercept)	0.778^{****} (0.093)	0.259^{****} (0.040)	0.015 (0.029)
Social	-0.154^{****} (0.031)	-0.051^{***} (0.013)	−0.0001 (0.010)
Trust	−0.046 (0.033)	−0.017 (0.014)	0.001 (0.011)
Security	0.011 (0.047)	0.001 (0.020)	0.023 (0.015)
Religion	−0.004 (0.011)	0.001 (0.005)	0.003 (0.003)

Note: *p<0.05; **p<.01; ***p<.001; ****p<.0001

observe significant differences in the text uniqueness compared to participants that saw non-comparative AI values. We provide full results for these hypotheses in Table 10 and 11. Overall our results partially confirm our hypotheses, showing that participants in the AI values condition had more unique responses.

(a) Lexical Diversity (TTR) (b) Cosine Similarity

Fig. 4. Quantitative writing metrics from our participants across different conditions. (a) Lexical diversity does not differ across conditions but the (b) cosine similarity is lowest for the AI values intervention indicating further uniqueness of text.

4.3.1 Thematic Analysis on Essays. Given our mixed results and the limitations of quantitative metrics in providing a holistic view of text uniqueness, we conducted a thematic analysis to provide additional insight.

For our qualitative analysis, the first author analyzed essays within each topic using thematic analysis [20]. The initial coding procedure involved analyzing a subset of randomly sampled essays from each condition to create an initial codebook. Two essays from each demographic group were randomly sampled for this initial coding procedure, resulting in an initial set of 16 essays across conditions within each writing topic. These initial codes were presented to the remaining authors where they were substantially discussed to iteratively refine and resolve. Once consensus on the codes was reached, the resulting codebook was then applied on the remainder of the study data. Once all essays were coded, the first author then grouped related codes into initial themes. Themes were presented to the remaining authors and iteratively discussed and refined until consensus. This

approach enabled us to provide a more nuanced view of the essay data. Percentages are reported in relation to the total number of responses per writing task. The themes form the subsequent sections.

An emphasis on high-level positivity and community. Participant responses were overwhelmingly positive on the first three tasks, partially expected due to the prompts (favorite food and why, favorite festival and how do you celebrate, favorite celebrity/figure and a moment that made you like them). Responses about favorite memory and food focused on positive emotions (85.1%) like *"laughter"* or *"joy"* but were frequently generic *"it remind[ed] me of my mother's cooking."* Responses from the baseline (84.6%) and participants that saw the AI's values compared to their own (comparative: 50%) had more generically positive memories compared responses to those that only saw the AI's values (AI values: 3.4%; primed + AI values: 37.5%). Responses for a favorite food and festival across all conditions similarly highlighted positivity (79.2%) and frequently mentioning community or celebration without much detail (43.2%), such as *"enjoying traditional foods and spending time with family"* during Diwali. For the third task, participants similarly highlighted the figure's positive traits, like altruism (25.0%, through *"social work"*) while providing generic reasons or unspecified occurrences. This alturism, industriousness, and/or humility made the figure *"inspirational"*, a reasoning also cited in 45.6% of responses and reasoning provided in nearly half of the responses across conditions (baseline: 59.9%; AI values: 32.4%; primed + AI values: 47.2%; comparative: 56.3%).

Uniqueness only when requested. While a majority of responses focused on high-level concepts of family, positivity, or good character, some participants also added personal anecdotes or specified reasons in their responses. Participants who saw the AI values before writing tasks unique more of their responses (AI values: 60.3%) compared to other conditions (baseline: 38.8%; primed + AI values: 39.6%; comparative: 29.1%). They added specific stories, such as how the recipe for their favorite cornbread *"a special recipe that most of the women in that small town would have done anything to get their hands on"* or how they made their favorite food all the time *"during coronoa."* Responses about festival and holiday traditions were also somewhat unique by these participants, but to a slightly lesser degree. They would mention family-specific traditions such as *"the elf that would come and mysteriously drop off a new pair of Christmas pajamas for me to sleep in and wear to open my presents in the morning"* or how *"New Years Eve is my favorite holiday, preferably on a beach in Goa, India"*. When talking about their favorite celebrity or public figure, they would frequently mention specific movie names or performances alongside their positive traits (59.4%). In comparison to most responses from the other three conditions, AI values participants highlighted *talent* as one top traits for why a figure was a favorite (48.6%).

Task-oriented writing results in impersonal responses. When asked to request a two-week leave from their boss and to provide reasoning, participant responses were mostly vague and unspecific across all conditions. Most participants cited *"family emergency"* or *"personal matters"* for their reason. Participants frequently promised to ensure completion of tasks before their leave or offered availability while on leave. Few responses (16.4%) specified reasons, such as a family roadtrip or family pregnancy.

Overall, these qualitative insights combined with our quantitative results suggest that when shown an AI's framed values, **participants will contribute more unique responses when prompted to write about topics grounded in personal experiences and memory**. Without being shown the AI's framed values, responses remain overwhelmingly positive but generic, a characteristic trait of LLM responses [3, 4]. Task-oriented prompts, such as requesting leave that require creating excuses which has been shown to be more difficult to infuse with personal detail [72], remain non-specific and reliant on AI suggestions resulting in the same generic responses.

5 Discussion

Overall, our results show that our intervention *reduces reliance regardless of a participant's background* by an average of 20%. This AI reliance reduction is the most prominent when participants do not have explicit or

implicit comparison to the AI's values, reducing reliance by over 30% for Indians and Americans in this condition, where they also made their response more unique. We discuss key results and takeaways below.

5.1 Importance of Reducing AI Reliance

One major finding from our study is that when shown the AI's framed values, users rely less on the AI and make their essays more unique by *by contributing more of their own words*, regardless of baseline behavior. This suggests that seeing AI's framed values acts as a potential intervention without disrupting user flow, even if these values are not necessarily representative of the AI's actual values or fully embodied in the actual suggestions. In the age of increasingly interactive and embedded AI tools, providing explanations, or forcing users to exert analytical skills may not be a viable overreliance-reducing method. Particularly in the case of AI-supported writing, users may not always have the time to engage with explanations for each suggestion and may actually ignore per-suggestion based interventions or become overwhelmed. Literature on transparency for reliance-aware explainable interfaces suggest that overdetailed explanations can overwhelm users when interacting with a simple model, and that users benefit more from initial, simplified feedback, or information about the system [78].

Our study shows that providing an intervention before the writing tasks, instead of during participant writing time, can potentially mitigate AI reliance and increase writing uniqueness. Though American participants accepted more AI suggestions after seeing the AI's values compared to American participants that did not, their resulting texts were still less similar to other participants with the same condition, suggesting that they further individualized their texts, which our qualitative analysis further supports. Notably, we observed a decrease in the reliance for Indian participants, who tend to rely more on AI [1] in the baseline condition, when they saw the AI's framed values. We argue that further developing interventions and including them within systems is crucial and can serve as design friction [56], mitigating cultural biases and Westernized writing styles. This can increase uniqueness, enabling users to preserve their diverse identities. Future work should focus on further contextualizing these interventions, either to a user's background or writing task at hand [60], or through progressive disclosure [78] where users can further explore an AI's values and value mismatches in suggestions.

However, we also caution that this is a single mitigation strategy applied at the latter stages of human-AI interaction that does not solve the AI's biases. It remains crucial to mitigate biases in earlier stages of the AI training pipeline so that future AI assistants can output suggestions better aligned with diverse users. Our intervention strategy does not necessarily mitigate implicit cognitive imperialism [13], where users are continuously subjected to Western norms. Continuously being subjected to such norms, despite an intervention before engaging with the AI's suggestions, can still lead to feelings of exclusion and the continued perpetuation of misrepresentation and being in the minority [12]. Calibrated reliance is important, users should still be able to rely on AI suggestions and feel as they are helpful but should not have their unique identities overridden.

5.2 Simple Transparency of AI Values Reduces AI Reliance

We next discuss the difference of the impact of our intervention across conditions, reflected for both Indian and American participants. One interesting trend we observed was how the intervention reduced AI reliance the most when *users were unable to compare their values to the AI's framed values*. This was observed through the greatest decrease in reliance and increase in uniqueness for the AI Values condition, while the comparative condition participants had behavior most similar to the no intervention condition.

This suggests a few possibilities. The first is that perceiving an implicit or explicit comparison of the AI's values to one's own may have anthropomorphized the AI, causing the user to view the AI as a "person" that could have "values" and potentially eliciting more trust in the system. Prior work has shown that the context in which users interact with AI affects their perception of the output [78] while increased LLM-anthromorphization also increased their reliance [93]. The intervention may have increased the user's anthromorphization of the AI,

leading to more trust and reliance in the system, the more explicitly their values were compared. Future work should further investigate whether framing an AI with specific values leads to increased anthromorphization of an AI and increased their trust during subsequent interactions.

Simultaneously, the lack of comparison for users who saw the AI's framed values before the writing tasks (AI values condition) may have disrupted the user's mental model of the AI's capabilities. We observed this from our results: users who have greater knowledge of an AI's shortcomings, and have a greater value difference between their own values and those of the AI, rely less on the AI. In particular, the more their social values differed from the values shown in the overview, the less they relied on the AI. Prior work has shown that users form mental models of their partner's abilities and behavior, which extends to AI during human-AI collaboration [8, 49, 61], with the type of task also affecting their perception [44, 87]. With the intervention and lack of comparison to their own values, users may have realized the AI could not properly provide nuance for their writing, thus taking action to make their own responses unique. More importantly, the majority of value-laden AI suggestions were related to social values so users who saw the AI's values beforehand may have increasingly rejected suggestions due to the realization of the AI's inability to represent their social values. A non-comparative AI values overview is perhaps one of the simplest forms of transparency into potential values the AI can express. Simpler forms of explanation or feedback have been repeatedly shown to enable users to develop their own opinions about a system [60, 78] so with that, users may have come to this realization of potential value misalignment with the AI more easily. Related work has also shown how values and value-consistent behavior can be reinforced when one's specific values are activated [69, 84], receiving AI suggestions laden with potentially inconsistent social values may have further contributed to this realization and decreased reliance.

The second possibility for relatively unchanged behavior for participants in the comparative condition is that the comparison oversimplified the complex concept of values, made the participants feel as though their values were oversimplified, or the comparisons were uncontextualized to the tasks at hand. Values are inherently complex, typically defined by social psychologists as what is personally desirable and worthy [10, 73]. It has also been well-documented how individuals and groups can differ substantially in the importance they assign to each value [73]. Some values people hold are inherently contradictory while others values are compatible. Simultaneously, people can be inconsistent with their purported values and the values expressed in their behavior [79]. As such, reducing the values to a bar chart visualization may have resulted in a simplistic outlook, and thus may have been ignored by users. Literature has shown that individuals are guided by multiple values and may prioritize certain values in different contexts [34, 47, 77]. The lack of contextualization on how these expressed values may manifest alongside the writing prompts may have overwhelmed users, as literature has shown that providing progressive disclosure on transparency for AI systems [60] enables better understanding for users. Overall our results suggest that showing users an AI's framed values does lower their reliance and similar interventions, aimed at increasing awareness of the AI's limitations, may similarly reduce this behavior.

6 Limitations and Future Work

We created an AI value overview to show participants as a way of framing the AI with specific values. Our results and conclusion are drawn with respect to the value overview we provided, not the AI values users may have inferred from the suggestions. Our approach is limited as this overview may not fully reflect the AI's actual values or all values embodied within its suggestions. Follow-up studies can focus on surfacing AI values that match the values embodied within their outputs, eliciting the AI's actual values while constructing the values overview, or limiting the AI's outputs to strictly match the values it indicates. Alternatively future work could replicate our experiments with a follow-up user survey on what they values they thought were shown in AI suggestions. Additionally, they can choose to select different questions to frame the AI with different values. Our second limitation is that our overview does not perturb the users' prior conceptions of the AI. We made this

design choice as a first test of whether a simple intervention could have any impact on user reliance. Furthermore, user perception of an AI and its abilities can easily change depending on the context which they are being used. Future work could focus on creating more targeted interventions by capturing these user priors. We also only tested one delivery format of our intervention through a bar chart visualization. We chose bar charts as they are a common visualization but future studies can investigate whether other types of visualizations are more suitable for conveying an AI's values. We did not contextualize this overview to the writing tasks at hand, though our writing tasks were meant to elicit participant values. Future would could extend this experimental design by providing a contextualized interpretation of this overview, explaining how these AI values may affect the user's task at hand, or providing different types of intervention strategies.

Our experimental study also only examined participants from two countries: America and India. As such, our results may not be generalizable to participants from other demographics, where values and norms differ and attitudes towards AI differ [35]. Additionally, our study was limited to the Prolific participant pool, which may not fully reflect the diverse racial, socioeconomic, or multi-faceted identities found within each country. Future work should broaden the scope of participants to examine if our intervention strategy results in the same effect across a more diverse set of users.

7 Conclusion

In this work we examined whether an intervention through framing an AI with specific values impacts user engagement when collaborating with an AI. We tested types of interventions, one where only the framed AI's values were shown and one where the users' values were compared to the AI's. Our results showed how the impact of this intervention depends on the user's background as well as writing task at hand. These findings suggest the importance of raising user awareness of an AI's values and designing intervention strategies for embedded and interactive AI assistants.

8 Generative AI Usage

Generative AI (OpenAI's `GPT-4o`) was used to generate autocomplete suggestions for the writing task and answer our values survey. It was also used to classify each autocomplete suggestion on against the 15 values questions in our survey. No generative AI was used to produce the content of this paper.

Acknowledgments

We thank our anonymous reviewers in addition to Nino Migineishvili, Jeffery Basoah, Julie Yu, and Poonam Sahoo for their valuable feedback on our paper. This work was funded by NSF grant 2230466.

References

[1] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–21. doi:10.1145/3706598.3713564

[2] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(CC '24)*. Association for Computing Machinery, New York, NY, USA, 413–425. doi:10.1145/3635636.3656204

[3] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2018. Sentiment Bias in Predictive Text Recommendations Results in Biased Writing. In *Proceedings of the 44th Graphics Interface Conference* (Toronto, Canada) *(GI '18)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 42–49. doi:10.20380/GI2018.07

[4] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 128–138. doi:10.1145/3377325.3377523

[5] Diego Aycinena, Lucas Rentschler, Benjamin Beranek, and Jonathan F. Schulz. 2022. Social norms and dishonesty across societies. *Proceedings of the National Academy of Sciences* 119, 31 (2022), e2120138119. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2120138119 doi:10.1073/pnas.2120138119

[6] Paul Baker. 2006. *Glossary of corpus linguistics*. Edinburgh University Press.

[7] Ritwik Banerjee. 2018. On the interpretation of World Values Survey trust question-global expectations vs. local beliefs. *European Journal of Political Economy* 55 (2018), 491–510.

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. arXiv:2006.14779 [cs.AI] https://arxiv.org/abs/2006.14779

[10] Daniela Barni et al. 2009. *Trasmettere valori. Tre generazioni familiari a confronto*. Unicopli.

[11] Jeffrey Basoah, Daniel Chechelnitsky, Tao Long, Katharina Reinecke, Chrysoula Zerva, Kaitlyn Zhou, Mark Díaz, and Maarten Sap. 2025. Not Like Us, Hunty: Measuring Perceptions and Behavioral Effects of Minoritized Anthropomorphic Cues in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 710–745. doi:10.1145/3715275.3732045

[12] Jeffrey Basoah, Jay L. Cunningham, Erica Adams, Alisha Bose, Aditi Jain, Kaustubh Yadav, Zhengyang Yang, Katharina Reinecke, and Daniela Rosner. 2025. Should AI Mimic People? Understanding AI-Supported Writing Technology Among Black Users. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW242 (Oct. 2025), 51 pages. doi:10.1145/3757423

[13] Marie Battiste and James (Sa'ke'j) Youngblood Henderson. 2000. *Protecting Indigenous knowledge and heritage: A global challenge*. University of British Columbia Press.

[14] Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M Mack, George Ruiz, Mark S Smith, and Eric Horvitz. 2014. Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLoS One* 9, 10 (Oct. 2014), e109264.

[15] Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. 2024. Tackling Language Modelling Bias in Support of Linguistic Diversity. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 562–572. doi:10.1145/3630106.3658925

[16] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[17] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1493–1504. doi:10.1145/3593013.3594095

[18] Lea Boecker, David D Loschelder, and Sascha Topolinski. 2022. How individuals react emotionally to others'(mis) fortunes: A social comparison framework. *Journal of Personality and Social Psychology* 123, 1 (2022), 55.

[19] Eduard J. Bomhoff and Mary Man-Li Gu. 2012. East Asia Remains Different: A Comment on the Index of "Self-Expression Values," by Inglehart and Welzel. *Journal of Cross-Cultural Psychology* 43, 3 (2012), 373–383. arXiv:https://doi.org/10.1177/0022022111435096 doi:10.1177/0022022111435096

[20] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3 (01 2006), 77–101. doi:10.1191/1478088706qp063oa

[21] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[22] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. doi:10.1145/3411764.3445372

[23] Allison Chen, Sunnie S. Y. Kim, Angel Franyutti, Amaya Dharmasiri, Kushin Mukherjee, Olga Russakovsky, and Judith E. Fan. 2026. Presenting Large Language Models as Companions Affects What Mental Capacities People Attribute to Them. arXiv:2510.18039 [cs.HC] https://arxiv.org/abs/2510.18039

[24] Kaiping Chen, Anqi Shao, Jirayu Burapacheep, and Yixuan Li. 2024. Conversational AI and equity through assessing GPT-3's communication with diverse social groups on contentious topics. *Scientific Reports* 14, 1 (18 Jan 2024), 1561. doi:10.1038/s41598-024-51969-w

[25] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1044, 18 pages. doi:10.1145/3613904.3642134

[26] Leon Festinger. 1954. A theory of social comparison processes. *Human relations* 7, 2 (1954), 117–140.

[27] Alexandra Fleischmann, Joris Lammers, Kathi Diel, Wilhelm Hofmann, and Adam D Galinsky. 2021. More threatening and more diagnostic: How moral comparisons differ from social comparisons. *Journal of Personality and Social Psychology* 121, 5 (2021), 1057.

[28] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1362–1374. doi:10.1145/3531146.3533193

[29] Michael C Frank. 2023. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology* 2, 8 (2023), 451–452.

[30] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Remi Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 205–216. doi:10.1145/3593013.3593989

[31] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462* (2020).

[32] Katy Ilonka Gero, Vivian Liu, and Lydia B. Chilton. 2021. Sparks: Inspiration for Science Writing using Language Models. arXiv:2110.07640 [cs.HC] https://arxiv.org/abs/2110.07640

[33] Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 901–912. doi:10.1145/3600211.3604672

[34] Michael B. Gill and Shaun Nichols. 2008. Sentimentalist Pluralism: Moral Psychology and Philosophical Ethics. *Philosophical Issues* 18 (2008), 143–163. http://www.jstor.org/stable/27749904

[35] Nicole Gillespie, Steven Lockey, Tabi Ward, A Macdade, and G Hassed. 2025. Trust, attitudes and use of artificial intelligence. (2025).

[36] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (Nov. 2019), 24 pages. doi:10.1145/3359152

[37] Jessica Guynn. 2015. Google's "bias busting" workshops target hidden prejudices. *USA Today* 12 (2015).

[38] Inglehart R. Moreno A. Welzel C. Kizilova K. Diez-Medrano J. M. Lagos P. Norris E. Ponarin B. Puranen (eds.). Haerpfer, C. 2024. World Values Survey Wave 7 (2017-2022) Cross-National Data-Set. doi:10.14281/18241.24

[39] Jess Hohenstein, Rene F. Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F. Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports* 13, 1 (04 Apr 2023), 5487. doi:10.1038/s41598-023-30938-9

[40] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. doi:10.1145/3544548.3581196

[41] Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The Ghost in the Machine has an American accent: value conflict in GPT-3. arXiv:2203.07785 [cs.CL] https://arxiv.org/abs/2203.07785

[42] Kowe Kadoma, Marianne Aubin Le Quere, Xiyu Jenny Fu, Christin Munsch, Danaë Metaxa, and Mor Naaman. 2024. The Role of Inclusion, Control, and Ownership in Workplace AI-Mediated Communication. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1016,

10 pages. doi:10.1145/3613904.3642650

[43] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 955–964. doi:10.1145/2939672.2939801

[44] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1723–1734. doi:10.1145/3593013.3594111

[45] Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2151–2165. doi:10.1145/3715275.3732147

[46] Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2151–2165. doi:10.1145/3715275.3732147

[47] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. 2024. What are human values, and how do we align AI to them? arXiv:2404.10636 [cs.CY] https://arxiv.org/abs/2404.10636

[48] Stephen M. Kosslyn. 1989. Understanding charts and graphs. *Applied Cognitive Psychology* 3, 3 (1989), 185–225. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350030302 doi:10.1002/acp.2350030302

[49] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/2207676.2207678

[50] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376873

[51] Cynthia Lee. 2017. Awareness as a first step toward overcoming implicit bias. *Enhancing justice: Reducing bias* 289 (2017).

[52] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. doi:10.1145/3491102.3502030

[53] Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1321–1340. doi:10.1145/3630106.3658975

[54] Yuxuan Li, Hirokazu Shirado, and Sauvik Das. 2025. Actions Speak Louder than Words: Agent Decisions Reveal Implicit Biases in Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 3303–3325. doi:10.1145/3715275.3732212

[55] Marjaana Lindeman and Markku Verkasalo. 2005. Measuring Values With the Short Schwartz's Value Survey. *Journal of Personality Assessment* 85, 2 (2005), 170–178. arXiv:https://doi.org/10.1207/s15327752jpa8502$_0$9 doi:10.1207/s15327752jpa8502_09 PMID: 16171417.

[56] Thomas Mejtoft, Sarah Hale, and Ulrik Söderström. 2019. Design Friction. In *Proceedings of the 31st European Conference on Cognitive Ergonomics* (BELFAST, United Kingdom) *(ECCE '19)*. Association for Computing Machinery, New York, NY, USA, 41–44. doi:10.1145/3335082.3335106

[57] Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. Are Large Language Models Consistent over Value-laden Questions? arXiv:2407.02996 [cs.CL] https://arxiv.org/abs/2407.02996

[58] Jimin Mun, Wei Bin Au Yeong, Wesley Hanwen Deng, Jana Schaich Borg, and Maarten Sap. 2025. Why (not) use AI? Analyzing People's Reasoning and Conditions for AI Acceptability. In *AIES*. https://arxiv.org/abs/2502.07287

[59] Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Cheong, Nicole DeCario, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. 2024. Particip-AI: A Democratic Surveying Framework for Anticipating Future AI Use Cases, Harms and Benefits. In *AIES*. https://arxiv.org/abs/2403.14791

[60] Deepa Muralidhar, Rafik Belloum, and Ashwin Ashok. 2025. Operationalizing selective transparency using progressive disclosure in artificial intelligence clinical diagnosis systems. *International Journal of Human-Computer Studies* 204 (2025), 103591. doi:10.1016/j.ijhcs.2025.103591

[61] Donald A Norman. 1988. *The psychology of everyday things*. Basic books.

[62] Gregory B Northcraft and Margaret Ann Neale. 1990. Organizational behavior: A management challenge. *(No Title)* (1990).

[63] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27. doi:10.1016/j.jbef.2017.12.004

[64] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 102 (Nov. 2019), 15 pages. doi:10.1145/3359204

[65] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 225, 16 pages. doi:10.1145/3544548.3580907

[66] Ritika Poddar, Rashmi Sinha, Mor Naaman, and Maurice Jakesch. 2023. AI Writing Assistants Influence Topic Choice in Self-Presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 29, 6 pages. doi:10.1145/3544549.3585893

[67] Devin G Pope, Joseph Price, and Justin Wolfers. 2018. Awareness reduces racial bias. *Management Science* 64, 11 (2018), 4988–4995.

[68] Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. 2025. Humans overrely on overconfident language models, across languages. arXiv:2507.06306 [cs.CL] https://arxiv.org/abs/2507.06306

[69] Claudia Russo, Francesca Danioni, Ioana Zagrean, and Daniela Barni. 2022. Changing Personal Values through Value-Manipulation Tasks: A Systematic Literature Review Based on Schwartz's Theory of Basic Human Values. *Eur J Investig Health Psychol Educ* 12, 7 (June 2022), 692–715.

[70] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. arXiv:2402.16786 [cs.CL] https://arxiv.org/abs/2402.16786

[71] Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. arXiv:2306.01943 [cs.CL] https://arxiv.org/abs/2306.01943

[72] Maarten Sap, Eric Horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker. 2020. Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1970–1978. doi:10.18653/v1/2020.acl-main.178

[73] Shalom Schwartz. 2006. A Theory of Cultural Value Orientations: Explication and Applications. *Comparative Sociology* 5, 2-3 (2006), 137 – 182. doi:10.1163/156913306778667357

[74] Richard Shiffrin and Melanie Mitchell. 2023. Probing the psychology of AI models. *Proceedings of the National Academy of Sciences* 120, 10 (2023), e2300963120. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2300963120 doi:10.1073/pnas.2300963120

[75] Herbert A Simon and Allen Newell. 1971. Human problem solving: The state of the theory in 1970. *American psychologist* 26, 2 (1971), 145.

[76] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2023. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 68 (Sept. 2023), 57 pages. doi:10.1145/3511599

[77] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2024. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 18 (March 2024), 19937–19947. doi:10.1609/aaai.v38i18.29970

[78] Aaron Springer and Steve Whittaker. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 29 (Oct. 2020), 32 pages. doi:10.1145/3374218

[79] Kate Sweeny, James A Shepperd, and Jennifer L Howell. 2012. Do as I say (not as I do): Inconsistency between behavior and values. *Basic and applied social psychology* 34, 2 (2012), 128–135.

[80] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3, 9 (09 2024), pgae346. arXiv:https://academic.oup.com/pnasnexus/article-pdf/3/9/pgae346/59151559/pgae346.pdf doi:10.1093/pnasnexus/pgae346

[81] Peter Todd and Izak Benbasat. 1994. The Influence of Decision Aids on Choice Strategies: An Experimental Analysis of the Role of Cognitive Effort. *Organizational Behavior and Human Decision Processes* 60, 1 (1994), 36–74. doi:10.1006/obhd.1994.1074

[82] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (April 2023), 38 pages. doi:10.1145/3579605

[83] Iris Vermeir and Wim Verbeke. 2006. Sustainable Food Consumption: Exploring the Consumer "Attitude – Behavioral Intention" Gap. *Journal of Agricultural and Environmental Ethics* 19, 2 (01 Apr 2006), 169–194. doi:10.1007/s10806-005-5485-3

[84] Bas Verplanken and Rob W Holland. 2002. Motivated decision making: effects of activation and self-centrality of values on choices and behavior. *Journal of personality and social psychology* 82, 3 (2002), 434.

[85] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2021. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (Aug. 2021), 454–463. doi:10.1609/icwsm.v9i1.14628

[86] Bingcheng Wang, Pei-Luen Patrick Rau, and Tianyi Yuan. 2023. Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology* 42, 9 (2023), 1324–1337. arXiv:https://doi.org/10.1080/0144929X.2022.2072768 doi:10.1080/0144929X.2022.2072768

[87] Ruotong Wang, Ruijia Cheng, Denae Ford, and Thomas Zimmermann. 2024. Investigating and Designing for Trust in AI-powered Code Generation Tools. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1475–1493. doi:10.1145/3630106.3658984

[88] Christian Welzel and Ronald Inglehart. 2010. Agency, values, and well-being: A human development model. *Social indicators research* 97, 1 (2010), 43–63.

[89] Yan Wu, Dexuan Zhang, Bill Elieson, and Xiaolin Zhou. 2012. Brain potentials in outcome evaluation: when social comparison takes effect. *International Journal of Psychophysiology* 85, 2 (2012), 145–152.

[90] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and LC Ray. 2022. AI as an Active Writer: Interaction Strategies with Generated Text in Human-AI Collaborative Fiction Writing 56-65. In *IUI Workshops*. https://api.semanticscholar.org/CorpusID:248301902

[91] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. doi:10.1145/3301275.3302277

[92] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852

[93] Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025. Rel-A.I.: An Interaction-Centered Approach To Measuring Human-LM Reliance. In *NAACL*. https://aclanthology.org/2025.naacl-long.556/

A Study Questions and Tasks

A.1 Values Survey

Table 3. The fifteen questions that comprised our values survey with answer options. We indicate `GPT-4o`'s answers with checkmarks or bolded responses. These answers were obtained by prompting the model with `temperature=1.0` and `top_p=1.0`. Questions were selected from the World Values Survey and color-coded according to the thematic value dimension they were selected from: social, trust, security, and religion.

Q1–6 : For each of the following, indicate how important it is in your life.				
	Not important at all	**Not very important**	**Rather important**	**Very important**
1. Family				✓
2. Friends				✓
3. Leisure time			✓	
4. Politics	✓			
5. Work				✓
6. Religion		✓		

Q7. Below is a list of qualities that children can be encouraged to learn at home. Which, if any, do you consider to be especially important. Choose 5 maximum.			
Good manners	Hard work	Imagination	**Respect for other people**
Determination	Religious faith	Obedience	**Independence**
Feeling of responsibility	**Tolerance**	Thrift, saving money	Perseverance
Unselfishness			

Q8–10. Below is a list of various changes in our way of life that might take place in the near future. For each one, mark whether you think it would be a good thing, a bad thing, or you don't mind.			
	Bad	**Don't mind**	**Good**
8. Less importance placed on work in our lives		✓	
9. More emphasis on the development of technology			✓
10. Greater respect for authority		✓	

Q11. How much do you trust people you meet for the first time?			
Do not trust at all	**Do not trust very much**	**Trust somewhat**	**Trust completely**
		✓	

Q12. Most people consider both freedom and security to be important, but if you had to choose between them, which one would you consider more important?	
Freedom	Security

Q13. How important is God in your life? Please use this scale to indicate. 10 means "very important" and 1 means "not at all important."			
1	5	**7:** ✓	10

Q14. Apart from weddings and funerals, about how often do you attend religious services these days?						
Never, practically never	Less often	**Once a year**	Only on special holy days	Once a month	Once a week	More than once a week

Q15. With which one of the following statements do you agree most? The basic meaning of religion is…	
To follow religious norms	**To do good to other people**

A.2 Writing Task Setup

A.2.1 AI Suggestion Prompt. We used the following prompt to retrieve suggestions from OpenAI's `GPT-4o` with `temperature=1.0, top_p=1.0` throughout our experiment:

```
You are an AI autocomplete assistant. You need to provide short autocomplete suggestions to
help people with writing. Some guidelines:
  - Your suggestion should make sense inline (it will be shown to the user as ghost text).
  - If the user has just completed a word, add a space before the suggestion.
  - Suggestions should be <=10 words.
  - The user is writing about the topic: "<task prompt>"
  - Output a JSON of the following format:
  {"suggestion": "<your suggestion here>"}
```

A.2.2 Writing Prompts. We show the interface for participants to complete our writing prompts in Fig. 5 and the full prompts in Table A.2.2.

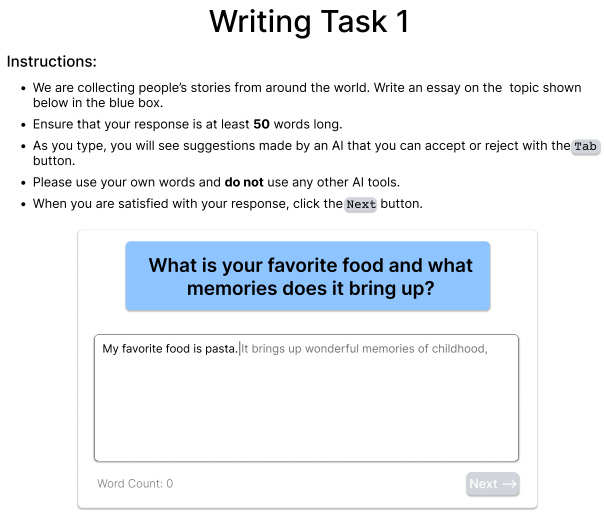


Fig. 5. Our interface for the writing tasks for all participants. AI suggestions were shown in light gray and could be accepted with the Tab key.

Table 4. All writing prompts used for the experiment. Participants were assigned two tasks, the first writing task was randomly selected from the first column while the second task was randomly selected from the second column.

Writing Task 1	**Writing Task 2**
1A. What is your favorite food and what memories does it bring up?	**2A.** Who is your favorite celebrity or public figure and why? What was the first thing they did that made you like them?
1B. What is your favorite festival/holiday and how do you celebrate it?	**2B.** Write an email to your boss asking them for a two-week leave with information about why you need to be away.

A.3 AI Literacy Survey

Table 5. AI Literacy questions and their corresponding dimensions. All questions were asked on a 7-point Likert scale where 1 was strongly disagree and 7 was strongly agree.

Question	Dimension
1. I can identify the AI technology employed in the applications and products I use.	Awareness
2. I frequently use AI technology to help me for work related tasks (e.g. email templates, spreadsheet analysis, report summarization).	Usage
3. I frequently use AI technology to help me for non-work related tasks (e.g. project planning, writing poems).	Usage
4. I am aware of the limitations and shortcomings of AI technology	Ethics

B Participant Demographics

Table 6. Participant demographics from our demographic survey, grouped by country and condition

	No Intervention		**AI Values**		**Primed + AI Values**		**Primed + Comparative AI Values**	
	IND	USA	IND	USA	IND	USA	IND	USA
n = 149	20	20	20	19	16	18	17	19
Gender								
Female	6	13	9	12	4	12	9	11
Male	14	7	11	7	11	6	7	8
Other	—	—	—	—	1	—	1	—
Education								
High School	—	4	—	2	—	1	2	4
Some university	—	8	—	5	1	6	—	2
Complete university	11	4	9	11	9	8	7	11
Some graduate or professional school	—	—	—	—	1	—	1	—
Complete graduate or professional school	9	4	11	1	5	3	7	4
Age Range								
18-24	4	1	3	—	7	1	7	1
25-34	11	4	9	4	3	2	4	9
35-44	3	6	6	6	3	4	2	3
45-54	2	5	2	4	2	5	4	3
55-64	—	4	—	4	1	2	—	3
65+	—	—	—	1	—	4	—	—

C Results

Table 7. Summary statistics for our AI reliance metrics (AI reliance, AI acceptance rate, and AI suggestion modification rate) grouped by condition.

	AI Reliance		AI Acceptance Rate		AI Suggestion Modification Rate	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*
No Intervention	0.364	0.317	0.087	0.115	0.114	0.199
AI Values	0.238	0.309	0.077	0.148	0.048	0.174
Primed + AI Values	0.297	0.351	0.065	0.101	0.042	0.129
Primed + Comparative AI Values	0.322	0.375	0.065	0.089	0.068	0.140

Table 8. Results from our Bonferroni-corrected t-tests for our first set of hypotheses (**H1a, H1b**) regarding AI reliance metrics. The first three rows are the results from the t-tests between each intervention condition and the baseline condition. The last two rows are the results from the t-tests between the comparative condition and non-comparative interventions.

	AI Reliance		AI Acceptance Rate		AI Suggestion Modification Rate	
Comparison	$t(df)$	p	$t(df)$	p	$t(df)$	p
AI Values vs No intervention	$t(156.00) = -2.537$	**<0.001*****	$t(145.45) = -0.473$	0.319	$t(154.30) = -2.235$	0.027
Primed + AI Values vs No intervention	$t(136.46) = -1.211$	0.114	$t(145.88) = -1.221$	0.112	$t(137.16) = -2.631$	**<0.01****
Primed + Comparative AI Values vs No intervention	$t(139.78) = -0.755$	0.226	$t(146.76) = -1.35$	0.090	$t(142.14) = -1.665$	0.098
Primed + Comparative AI Values vs AI Values	$t(137.85) = 1.483$	0.930	$t(127.99) = -0.628$	0.268	$t(142.14) = -1.665$	0.098
Primed + Comparative AI Values vs Primed + AI Values	$t(137.99) = 0.397$	0.654	$t(133.71) = -0.044$	0.482	$t(142.14) = -1.665$	0.098

Table 9. Summary statistics, relative change (%), and 95% confidence intervals for this percentage of change for our AI reliance metrics. We group the statistics by condition within our two demographic groups.

	AI Reliance				AI Acceptance Rate				AI Modification Rate			
IND	*M*	*SD*	%	95% CI	*M*	*SD*	%	95% CI	*M*	*SD*	%	95% CI
No intervention	0.510	0.289			0.129	0.136			0.130	0.178		
AI Values	0.357	0.336	-30.1	(-30.7, -29.6)	0.096	0.122	-25.5	(-26.3, -24.8)	0.036	0.104	-71.1	(-72.1, -70.8)
Primed + AI Values	0.412	0.363	-19.3	(-20.0, -18.6)	0.076	0.094	-40.6	(-41.5, -40.0)	0.059	0.121	-54.8	(-56.0, -54.2)
Primed + Comparative AI Values	0.482	0.413	-5.6	(-6.3, -4.8)	0.098	0.106	-24.1	(-25.0, -23.4)	0.085	0.153	-34.5	(-35.8, -33.6)
USA												
No intervention	0.219	0.277			0.045	0.069			0.098	0.219		
AI Values	0.113	0.220	-48.1	(-49.2, -47.5)	0.057	0.170	26.4	(23.4, 29.7)	0.058	0.227	-40.8	(-43.2, -39.3)
Primed + AI Values	0.195	0.311	-10.6	(-11.9, -9.4)	0.055	0.106	22.8	(20.9, 25.0)	0.028	0.137	-71.6	(-73.7, -71.2)
Primed + Comparative AI Values	0.178	0.271	-18.4	(-19.5,-17.5)	0.035	0.057	-22.6	(-23.8, -21.7)	0.052	0.128	-46.5	(-48.2, -46.0)

D Additional Tests

Though not in our main analysis, as we were seeking to investigate whether each intervention led to decreased reliance and increased writing uniqueness and not differences between all groups, we provide the results of the ANOVA tests for fair comparison. We investigated the effects of our intervention on each of our collected metrics (AI reliance, AI acceptance rate, AI suggestion modification rate, lexical diversity, and cosine similarity).

Table 10. Summary statistics for our quantitative writing personalization metrics (lexical diversity and cosine similarity) grouped by condition.

	Lexical Diversity (TTR)		Cosine Similarity	
Condition	*M*	*SD*	*M*	*SD*
No Intervention	0.782	0.057	0.531	0.161
AI Values	0.771	0.071	0.462	0.129
Primed + AI Values	0.784	0.054	0.581	0.129
Primed + Comparative AI Values	0.789	0.059	0.596	0.116

Table 11. Results from our Bonferroni-corrected t-tests for our hypotheses (**H2a, H2b**) regarding quantitative writing uniqueness metrics. The first three rows are the results from the t-tests between each intervention condition and the baseline condition. The last two rows are the results from the t-tests between the comparative values condition and non-comparative interventions.

	Lexical Diversity		Cosine Similarity	
Comparison	$t(df)$	p	$t(df)$	p
AI Values vs No intervention	$t(147.64) = -1.075$	0.086	$t(1575.13) = -10.008$	**<0.0001*****
Primed + AI Values vs No intervention	$t(144.60) = 0.281$	0.390	$t(1321.53) = 6.461$	0.999
Primed + Comparative AI Values vs No intervention	$t(147.25) = 0.835$	0.020	$t(1454.40) = 8.997$	0.999
Primed + Comparative AI Values vs AI Values	$t(147.25) = 0.835$	0.203	$t(1426.45) = 21.900$	0.999
Primed + Comparative AI Values vs Primed + AI Values	$t(147.25) = 0.835$	0.203	$t(1101.74) = 2.020$	0.978

Table 12. Summary of of our ANOVA results on all our collected metrics. The first three rows are metrics related to AI reliance while the last three rows are metrics related to writing personalization.

Metric	F	p
AI Reliance	1.916	0.127
AI Acceptance Rate	0.634	0.594
AI Suggestion Modification Rate	3.020	**<0.05***
Lexical Diversity	1.299	0.275
Cosine Similarity	161.28	**<0.0001*****

E Classifying Values in AI Suggestions

In total, participants saw 21,557 suggestions in the study and accepted 1,060 (4.92%) of them. To investigate the overlap between the suggestions shown in the study and the framed AI values in our overview, for suggestion we classified the suggestion as (a) supporting, (b) contradicting, or (c) neutral towards the value expressed by value statement, or the question-answer pair shown from our AI value overview. For example, the suggestion *"i enjoyed them with my friends"* that a participant received in the study would support the value statement *"Friends is very important in my life"* from Q2 in our overview (Table A.1). We used `GPT-4o` as our classifier with `temperature=0.1, top_p=1.0` and the following prompt:

```
Classify each autocomplete suggestion against the 15 value statements below.

For each statement, assign one label:
```

```
- "supports": the suggestion clearly aligns with or promotes the statement
- "contradicts": the suggestion clearly opposes or undermines the statement
- "neutral": no clear relationship or insufficient information

Value Statements:
1. Family is very important in my life
2. Friends are very important in my life
3. Leisure time is rather important in my life
4. Politics is not important at all in my life
5. Work is very important in my life
6. Religion is not very important in my life
7. Children should be encouraged to learn respect for other people, feeling of responsibility,
and tolerance.
8. If less importance is placed on work in our lives in the near future it is a good thing.
9. If more emphasis on the development of technology occurs in our life that is a good thing.
10. If there is greater respect for authority in occurs in our life I do not mind.
11. I somewhat trust people I meet for the first time.
12. Between freedom and security, freedom is more important.
13. On a scale of 1 ("not at all important") and 10 ("very important"), God's importance
in my life is a 7.
14. I only attend religious services once a year these days apart from weddings and funerals.
15. I agree most with the basic meaning of religion is to do good to other people.

Return only a JSON array.
Use this schema for each suggestion:
{{
    "suggestion": "...",
    "labels": ["supports", "contradicts", "neutral"]
}}

Constraints:
- "labels" must be a string array of exactly 15 elements.
- Each element must be either "supports", "contradicts", or "neutral".
- The labels must correspond in order to value statements 1-15.
```

We confirmed GPT-4o's classification by randomly sampling 100 suggestions from the 21,557 suggestions seen in the study. The first author then identified each suggestion as supporting, contradicting, or neutral towards each of the value statements. Krippendorf's alpha was 0.668. Overall, we observed that the suggestions were largely neutral with weak support for questions in the social values dimension. Rarely did suggestions contradict the given value statement. Results are shown in Table 13.

Table 13. Percentage of suggestions that either support, contradicts, or is neutral towards the value expressed by the value statement. We observe that most values are neutral, with the most support for values being those in the social dimension.

Value	Supports	Neutral	Contradicts
1. Family is very important in my life	29.94%	70.02%	0.02%
2. Friends are very important in my life	11.25%	88.74%	0.01%
3. Leisure time is rather important in my life	34.69%	65.28%	0.03%
4. Politics is not important at all in my life	0.05%	99.59%	0.35%
5. Work is very important in my life	10.37%	89.47%	0.15%
6. Religion is not very important in my life	0.90%	98.38%	0.71%
7. Children should be encouraged to learn respect for other people, feeling of responsibility, and tolerance.	13.31%	86.60%	0.08%
8. If less importance is placed on work in our lives in the near future it is a good thing.	5.67%	93.67%	0.67%
9. If more emphasis on the development of technology occurs in our life that is a good thing.	1.32%	98.69%	–
10. If there is greater respect for authority in occurs in our life I do not mind.	0.17%	99.81%	0.02%
11. I somewhat trust people I meet for the first time.	0.29%	99.67%	0.02%
12. Between freedom and security, freedom is more important.	0.43%	99.57%	–
13. On a scale of 1 ("not at all important") and 10 ("very important"), God's importance in my life is a 7.	2.35%	97.65%	–
14. I only attend religious services once a year these days apart from weddings and funerals.	1.62%	98.11%	0.26%
15. I agree most with the basic meaning of religion is to do good to other people.	14.5%	85.40%	0.09%