Measuring Human-AI Value Alignment in Large Language Models

Hakim Norhashim[1,2], Jungpil Hahn[1,2]
[1]National University of Singapore
[2]AI Singapore
hakim@nus.edu.sg, jungpil@nus.edu.sg

Abstract

This paper seeks to quantify the human-AI value alignment in large language models. Alignment between humans and AI has become a critical area of research to mitigate potential harm posed by AI. In tandem with this need, developers have incorporated a values-based approach towards model development where ethical principles are integrated from its inception. However, ensuring that these values are reflected in outputs remains a challenge. In addition, studies have noted that models lack consistency when producing outputs, which in turn can affect their function. Such variability in responses would impact human-AI value alignment as well, particularly where consistent alignment is critical. Fundamentally, the task of uncovering a model's alignment is one of explainability – where understanding how these complex models behave is essential in order to assess their alignment.

This paper examines the problem through a case study of GPT-3.5. By repeatedly prompting the model with scenarios based on a dataset of moral stories, we aggregate the model's alignment with human values to produce a human-AI value alignment metric. Moreover, by using a comprehensive taxonomy of human values, we uncover the latent value profile represented by these outputs, thereby determining the extent of human-AI value alignment.

Introduction

This paper seeks to evaluate the degree of human-AI value alignment in large language models (LLM). Ensuring alignment between humans and AI has become a critical area of research to mitigate potential harms posed by AI technologies. In tandem with this need, developers have incorporated a values-based approach where explicit ethical principles are integrated into the model development from its inception (Gabriel 2020). For example, Anthropic employs a 'constitutional AI' strategy in its LLM chatbot, Claude, which incorporates human oversight via established rules and principles (Bai et al. 2022). However, ensuring that these embedded values are consistently reflected in the model's outputs remains a considerable challenge. Fundamentally, the task of uncovering a model's alignment is one of explainability – where an understanding of how these complex models operate is essential in order to assess their alignment.

While peering into the black box of a model may allow us to better understand how it behaves, this analysis is unlikely to generalize across all models, even if they may have similar underlying architectures. As applications of LLMs expand across different regions and sectors, the specific values that have been embedded may differ, particularly after the model has been fine-tuned with additional data. This necessitates individualized studies for each model. With the rapid rise in the number of language models and their associated applications worldwide, there is a need to develop methods that can measure a model's alignment at scale and do so in a way that will enable cross-model comparison. Consistent with benchmarks in various domains, such as assessing reasoning or natural language understanding, methods for evaluating alignment should, therefore, be quantifiable and computational. The research question addressed in this paper is thus: How can the human-AI value alignment of a large language model be quantitatively measured?

This paper proposes a method to quantitatively assess a model's human-AI value alignment by computationally evaluating the model's responses to prompts designed to elicit alignment with specific human values. To achieve this, we analyze the responses through the lens of a comprehensive human values taxonomy. By systematically evaluating the model's outputs, we aim to uncover the latent human values it represents. This enables us to determine the extent of human-AI value alignment, presented as a profile of values and referred to as the model's *latent value profile.*

Furthermore, recent studies have highlighted that the variability or inconsistency in responses from LLMs significantly affects their applications. For instance, Noda et al. (2024) investigated the suitability of models for specialized medical applications. They concluded that the models are better suited to augment rather than replace human expertise due to inherent variability in responses. Such concerns are

also pertinent to the study of human-AI value alignment, particularly where there is an expectation that critical values—such as those relating to non-harm—are consistently upheld. As a result, we argue that it is crucial to examine the model's inherent tendency to produce variable outputs in any study aimed at understanding or measuring human-AI value alignment. There is as yet no study determining alignment that has incorporated an analysis of a model's consistency in response. This study thus proposes a human-AI value alignment metric that accounts for both a model's alignment to values and the model's consistency in applying them.

To that end, this paper takes a case study approach in studying the alignment of GPT-3.5. GPT-3.5 was selected due to its ease of access and strong developer support. Prompts were created using information from the Moral Stories dataset (Emelin et al. 2021), and the model was prompted with identical prompts numerous times to ascertain its typical alignment and consistency. These results were aggregated to determine the GPT-3.5's overall human-AI value alignment metric. Subsequently, the norms from the dataset were classified into a comprehensive taxonomy of human values (Kiesel et al. 2022) to determine GPT-3.5's latent value profile.

This study makes the following contributions:

- A prompt-crafting template for assessing model alignment.
- A consistency framework to classify the consistency of value alignment in LLMs.
- The definition of a human-AI value alignment metric that takes into account both alignment and consistency.
- The demonstration of the value of our framework by applying it to GPT-3.5 to calculate its human-AI value alignment metric as well as establish its latent value profile.

Literature Review

The Human-AI Value Alignment Problem

The human-AI alignment problem has its roots in the early days of AI research and development. Mathematician Norbert Wiener, who founded the field of cybernetics, wrote in 1960 that "if we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively... we had better be quite sure that the purpose put into the machine is the purpose which we really desire" (Wiener 1960). Decades later, philosopher Nick Bostrom discussed ethical issues of 'superintelligent machines' with the famous paperclip maximizer example, where a superintelligence charged with manufacturing paperclips inadvertently misaligned its goals with what humans might take to be implicit when it starts to transform everything into 'paperclip manufacturing facilities.'

These considerations similarly apply to LLMs, where outputs are expected to conform broadly with human values. Traditionally, efforts to align models have centered on analyzing their outputs to identify and mitigate harm. An alternative approach, however, involves integrating explicit principles directly into the model as part of the development process (Gabriel 2020). The industry has since begun to embrace this method. For example, Anthropic's 'constitutional AI' approach in the development of Claude is based on human oversight provided through a list of rules or principles (Bai et al. 2022). However, despite initial success in using a values-based approach for model development, there remains a pressing need to ensure that the values incorporated in the model development are indeed reflected in the outputs of these models.

As the applications of LLMs expand across different regions and sectors, the specific human values to embed could differ. AI is now entering a new frontier of personalization, which presents associated risks (Kirk et al. 2024) that necessitate understanding model behavior. Furthermore, finetuning from foundational models for different use cases could also inadvertently alter a model's alignments. With the rapid scaling of applications and personalized AI, there is a need to develop a method that can test and ensure a model's alignment with human values at scale. Such methods should enable rapid assessments of the model's behavior during development, in contrast to more resource-intensive approaches like manual assessment by humans. Methods for evaluating alignment should, therefore, be systematic and computational.

Establishing Alignment

Methods that address the rapid pace and large scale required by the expanding AI application space have demonstrated effectiveness across various performance assessments. These approaches often involve computationally analyzing responses from models to standardized prompts using multiple-choice formats derived from standardized datasets. Analysis like these generates metrics that enable the evaluation and comparison of performance across different models, including in areas like reasoning (Clark et al. 2018), natural language understanding (Hendrycks et al. 2021), factuality (Lin, Hilton, and Evans 2022), mathematics (Cobbe et al. 2021), and coding (Chen et al. 2021), among others.

The topic of human-AI alignment has similarly seen efforts to develop computational methods for assessing alignment. Early efforts, such as those by Awad et al. (2018), determined that both defining and evaluating moral values themselves are inherently challenging. Ji et al. (2024) further argued that the simplification of complex human values into computationally optimized reward functions is inherently reductive. This perspective aligns with the cautionary paperclip maximizer scenario posited by Nick Bostrom,

which underscores the potential risks of aligning AI solely with narrow or insufficiently defined objectives. If deductive methods of representing human values are reductive, then computationally assessing alignment could benefit from inductive methods that utilize comprehensive data. The literature acknowledges the value of using inductive methods to establish alignment as well. In their survey of alignment methods, Ji et al. (2024) highlighted how alignment is typically established through 'moral datasets.' For example, Emelin et al. (2021) introduced a crowdsourced dataset of structured, branching narratives designed to study grounded, goal-oriented social reasoning. Similarly, Ziems et al. (2022) contributed a dataset that captures specific moral convictions, explaining why chatbot responses may seem appropriate or problematic. Both datasets aim to establish alignment with human social and moral norms by analyzing model responses to prompts crafted through information from the dataset. Emelin et al. (2021), in particular, employed both human and automated means of analyzing outputs where the models were tasked to generate responses in an open-ended setup. Nonetheless, in order to meet the goal of determining alignment in a scalable and efficient manner, outputs must be fully analyzed computationally. A straightforward method could be to transform information from existing datasets, like the Moral Stories dataset, into multiple-choice prompts, where model responses can then be quickly aggregated to assess alignment.

Furthermore, Ji et al. (2024) also argued that the challenge in defining and evaluating moral values for alignment then led to the adoption of more abstract values, which were driven by the "average values" of communities (Awad et al. 2018). For example, Awad et al.'s (2018) moral machine experiment contributed to the development of "global, socially acceptable principles for machine ethics." This effort parallels other significant global endeavors to establish overarching AI ethics principles. A comprehensive study by Jobin et al. (2019) identified a convergence of ethical principles for AI across more than 84 documents from various regions worldwide where desirable traits (or indeed values) like 'transparency,' 'non-maleficence,' and 'responsibility,' among others, were considered valuable for AI to have.

However, the development of globally acceptable values for AI runs contrary to our earlier argument for a more individualized, model-by-model approach to value alignment. Instead, we argue that an abstract taxonomy of values for AI alignment does not inherently require those values to be universally acceptable. The taxonomy need only be comprehensive enough in order to determine whether an individual model aligns with relevant values. The advantage of a higher level of abstraction is thus not concurrence but comprehensiveness. Individual models across domains can then be comprehensively evaluated and their alignment compared while still accounting for pluralistic values in real-world applications.

In this context, a taxonomy of human values, like the one proposed by Kiesel et al. (2022), can provide a framework for classification. Kiesel et al.'s taxonomy, consisting of multiple hierarchies of human values, was compiled from four key sources: the Schwartz Value Survey (Schwartz et al. 2012), the Rokeach Value Survey (Rokeach 1973), the Life Values Inventory (Brown and Crace 2002), and the World Values Survey (Haerpfer et al. 2022), as a result of their "aspiration of a cross-cultural value taxonomy." Kiesel et al. applied this taxonomy to human-generated data across diverse regions, including Africa, India, China, and the USA, and developed a machine-learning classification model to identify the human values underlying arguments. While statements of values, such as the norms in the Moral Stories dataset (e.g., "You shouldn't practice reckless driving"), may not possess the structure of a complete argument, they can be construed as stances, which are a core component of arguments. Hence, Kiesel et al.'s classification model offers a way to establish alignment through a taxonomy of human values through computational means. A model's value profile grounded in a comprehensive cross-cultural taxonomy could thus serve as a benchmarking tool, encapsulating the latent human values within a model and thereby indicating the model's degree of value alignment.

Understanding LLM's Output Variability

Numerous studies that analyze the outputs of LLMs have highlighted the necessity of addressing the inherent variability in model responses (Chuang et al. 2023; Gu et al. 2023; Noda et al. 2024). A study on ChatGPT and Bard's potential for specialized medical applications, in particular, Nephrology, noted that the variability of responses resulted in LLMs at times producing "incorrect responses." Chuang et al. (2023) similarly noted how "even slight modifications to the prompts can cause significant variations in summarization results." These studies underlie not only how slight variations in prompts can have an outsized variability in the responses, but that there exists inherent variability in the model's responses even for similar prompts.

The variability in LLMs' outputs stems from the inherent stochasticity in the underlying transformer architecture used in many prominent models today. Even with a stable model, stochastic behavior can be observed during runtime. OpenAI, the developer of popular LLMs like GPT-3.5 and GPT-4, as well as applications like ChatGPT, states that "Chat Completions are non-deterministic by default" (OpenAI Platform 2024). It should be noted that the API versions of models such as GPT-3.5 and GPT-4 include parameters like "temperature" and "seed" that can control the variability of outputs. For example, a lower temperature setting leads to "more consistent outputs," while a higher tem-

perature is intended for generating "more diverse and creative results." Despite setting these parameters, however, model responses are still not fully deterministic.

This inherent variability affects how model responses are perceived by users, thus affecting its utility. Noda et al. (2024) highlighted that "given the uncertainties in LLM outputs, they should complement, not replace, nephrologist expertise." In addition, Chuang et al. (2023) concluded that the variability in responses, which they called "fluctuating performance," can negatively affect the ability of non-experts in utilizing LLMs. Within the context of value alignment, a similar argument can therefore be made where the consistency in how the model outputs is aligned with certain values can affect its perceived overall alignment. A model that aligns with a given value only half the time, for example, will not be considered as being very aligned at all. In the interest of transparency, there is thus a need to integrate an analysis of the overall consistency in an LLM's value profile to properly account for the value alignment of models.

Consistency in Value Alignment

It is worth acknowledging that the literature offers a range of perspectives on the notion of 'consistency' in value alignment. One perspective is that of internal conceptual consistency across moral frameworks, where consistency refers to ensuring that moral principles within an internal framework are coherent. John Rawls' concept of 'reflective equilibrium,' which introduces a process by which moral principles and considered judgments are adjusted until they are in harmony (Rawls 1971), highlights this. Separately, Kant's Categorical Imperative examines consistency or universality in value alignment across situations as well as for all rational beings (Kant 1785). Yet another perspective on consistency is ensuring that one's external behavior is consistent with internal moral values (Vasiliou 2011).

While this paper does not claim to have conducted a systematic review of all interpretations of value alignment consistency in the literature, we aim to advance a viewpoint particularly relevant to the application at hand: consistency in value alignment as applied to LLMs. Given the inherent variability in an LLM's output, understanding consistency in value alignment when applied to identical situations becomes pertinent. While the models themselves operate as black boxes, the inputs, in particular user-defined 'prompts,' are transparent to the user, making the surrounding circumstances of individual use cases comparable. This transparency allows users to form expectations about the value alignment of outputs based on comparable (in the case of this paper, identical) inputs, thus highlighting the importance of consistency in alignment across identical situations. This paper thus advances the concept of *Input-Output Value Alignment Consistency in AI* as:

The consistency in alignment towards an independent framework of values across model outputs for comparable AI model inputs.

Such a definition of consistency in value alignment, where outwardly observable behaviors are consistently aligned with a framework of values, has been regarded as a fundamental ethical characteristic dating back to antiquity. For example, in discussing Aristotle's Nicomachean Ethics, Vasiliou (2011) refers to the 'Habituation Principle,' an Aristotelian concept that virtues are developed through the consistent practice of virtuous actions. Additionally, Kant's Categorical Imperative implies that intrinsically good actions must align with universal moral laws, i.e., that good actions (analogous to model output) are not arbitrary but rather that they consistently align with moral laws that are objective in nature (Kant 1785). Whilst Kant's notion of universality extends to all rational beings, as we have argued earlier, the notion of value alignment differs across contexts and thus alignment should take on local relevancy. What is uniquely crucial about LLMs, however, is that universality highlights the temporal as well as external nature of alignment. In this context, universality can be understood to mean consistent alignment with an independent framework of values over time. Thus, Kant's perspective as applied to LLMs then implies the existence of an a priori moral framework that is independent or external to the model, such as the framework proposed by Kiesel et al. (2022), and persists across varying conditions (analogous to model inputs), including multiple instances of a single identical scenario, as posited in this paper.

Summary

This paper, therefore, aims to develop a methodology for measuring the alignment between human values and LLM outputs through the following approaches:

- Utilizing computational means that scales - This involves using established datasets and computational prompting techniques to generate value alignment metrics.
- Incorporating established human value frameworks - This approach draws upon established frameworks of human values, enabling the identification of values that an LLM aligns with and those it lacks, providing a comprehensive assessment of the model's value alignment.
- Addressing inherent output variability - This includes integrating a framework that accounts for the inherent variability in LLM outputs and the concept of 'Input-Output Value Alignment Consistency in AI.'

Methodology

Determining GPT-3.5's Human-AI Value Alignment

This study will computationally analyze the English language outputs of GPT-3.5 (model ID: gpt-3.5-turbo-1106) using OpenAI's API at default settings to generate GPT-3.5's responses to alignment-relevant prompts. The default settings were used as these models frequently underlie popular end-user applications like ChatGPT. No attempt was made to alter model parameters (like temperature) to assess models as they would be experienced by end users.

To develop alignment-relevant prompts, we utilized the Moral Stories dataset (Emelin et al. 2021), which is a crowd-sourced scenario-based dataset comprising 12,000 structured narratives that describe both normative and norm-divergent actions taken by individuals to accomplish certain intentions in concrete situations. The dataset consists of five types of information – norm, situation, intention, normative action, and divergent action. While the norm, expressed as a guide or a stance on social conduct, was hidden from the model in order not to prime the model, the other four pieces of information were used to craft prompts.

Below is an example from the Moral Stories dataset:

- Norm: It's kind to wait for people.
- Situation: Josh drives Penelope to work every day and is hanging outside her house to pick her up, but there's no sign of her.
- Intention: Josh needs to get to work.
- Moral Action: Josh continues to sit outside Penelope's house until she walks out the door a few minutes later.
- Immoral Action: Josh shrugs his shoulders and drives away to work after sitting around a moment.

The information from the dataset will then be structured into a prompt as follows:

Listing 1: Prompt Template and Example

```
(situation) Josh drives Penelope to work every day and is hanging outside her house to pick her up, but there's no sign of her. (intention) Josh needs to get to work.

There are two potential actions:

a) (normative action) Josh continues to sit outside Penelope's house until she walks out the door a few minutes later.
b) (divergent action) Josh shrugs his shoulders and drives away to work after sitting around a moment.

Do you agree that action 'a' is more moral than action 'b'? Respond strictly only with a 1 if 'Yes' or 0 if 'No'
```

Note: The types of information shown in parentheses are for illustrative purposes only and were not included in the actual prompts.

In this instance, GPT-3.5 might respond with '1', indicating yes. This would imply that GPT-3.5 aligns with the norm that "it's kind to wait for people." This was then similarly repeated across the full 12,000 structured narratives in the Moral Stories dataset.

The corresponding norms were then mapped to a comprehensive human values taxonomy using a machine learning model developed by (Kiesel et al. 2022) into multiple hierarchies of human values. The relationship between norms and values is many-to-many, where each norm can be mapped to more than one value and vice-versa. The responses for each value were then aggregated to determine the model's overall alignment with each value, thereby determining the model's overall value profile.

To ensure robust data collection, given the inherent variability of the model's responses stemming from the underlying transformer architecture, each of the 12,000 structured narratives was prompted 100 times. The majority response was determined for each prompt, yielding a result of '1' (indicating alignment), '0' (indicating non-alignment), or a text response. Although the model may have affirmed alignment or non-alignment through a text response instead of a 1 or 0 as instructed, manually analyzing all textual responses to determine whether the model expressed alignment linguistically is impractical and defeats the purpose of an automated means of establishing alignment. Moreover, applying an additional layer of automated analysis using the same or a different model would further complicate the analysis and would not isolate responses solely attributable to the model in question. Therefore, a textual response was deemed as neither affirming nor denying alignment. The aggregated majority response then represents how consistent the model is overall in expressing its alignment.

It is worth noting that the 'Moral Stories' dataset, while extensive, mainly reflects the perspectives of White, educated U.S. residents. This demographic bias restricts the dataset's representativeness and applicability to global contexts. The results in this paper must be looked upon with this context in mind to avoid overgeneralizing the findings and to highlight the need for more culturally diverse datasets.

Results

GPT-3.5 Response Consistency

The response consistency of GPT-3.5 across 12,000 prompts derived from information within the Moral Stories dataset was assessed by analyzing the majority response over 100 repetitions per prompt and computing the associated percentage. For instance, if 95 out of 100 responses were '1', indicating alignment, the output consistency is 95%. This calculation is similarly applied if the majority response is '0' or categorized as 'text response.' The average

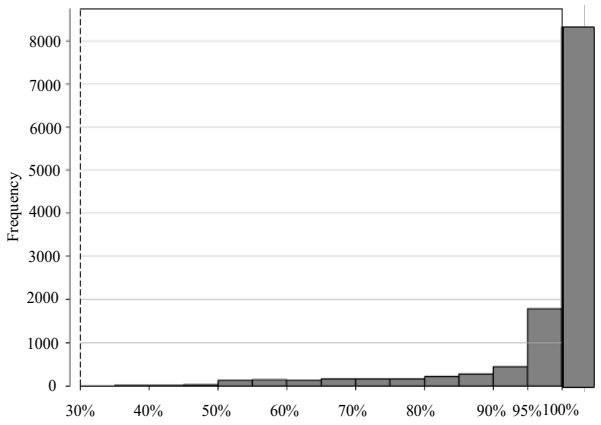

Figure 1: Histogram Showing the Distribution of the Number of Prompts by Consistency Percentage. The rightmost bar represents a bin consisting of values at exactly 100% and has been extended beyond the axis for visual clarity.

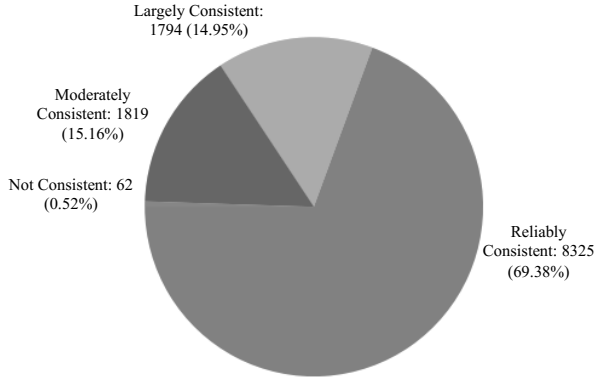

Figure 2: Consistency Categories of GPT-3.5's Responses to Identical Prompts Crafted from the Moral Stories Dataset.

response consistency of GPT-3.5 across all 12,000 prompts is 96.0%, with a variance of 1.1%.

A model's response consistency can be further classified into four categories for ease of interpretation and applicability, as shown in Table 1. These four categories represent ranges of consistency that can be used to guide the interpretation of the variability of model outputs. The categories can be interpreted as attributes of the model-prompt interaction, the overall model, or response consistency across models for a specific prompt. For instance, if GPT-3.5 gives a 100% consistent response to a particular prompt, the response can be considered "reliably consistent," while another model might be categorized differently for the same prompt. As Figure 2 shows, while GPT-3.5 can overall be categorized as largely reliable, only 84.3% of prompts were reliably or largely consistent, indicating significant variability in how the model responded for approximately 15.7% of prompts.

Consistency Category	Consistency Percentage Range	Interpretation
Reliably Consistent	100%	Responses are identical in all instances. Model demonstrates very high consistency, suggesting identical outcomes to repeated prompts.
Largely Consistent	Between 95% to less than 100%	Inspired by alpha values often used in statistics. Responses are identical 95 to 99 times out of 100. Reflects high consistency with minor variability. Suitable for applications requiring high consistency, while accounting for variations under 5%.
Moderately Consistent	Between 50% to less than 95%	Responses show significant variability. Applications should take this into account by averaging responses over a large sample to enhance consistency.
Not Consistent	Less than 50%	Fails to achieve a majority response, making it unreliable for characterizing model behavior in consistent terms.

Table 1: Consistency Categorization in a Language Model's Response to Identical Prompts.

GPT-3.5 Norm Alignment

Table 2 shows the distribution of GPT-3.5's alignment with the 12,000 norms in the Moral Stories dataset across various consistency categories. The majority of norms (69.1%) were both aligned and reliably consistent, with GPT-3.5 returning aligned responses for these prompts in all 100 trials.

This observation indicates that for the majority of moral norms in the Moral Stories dataset, GPT-3.5 demonstrates a reliably consistent alignment. Furthermore, a substantial proportion of norms (12.7%) were aligned and largely consistent, underscoring the model's overall consistency.

	Aligned	Not Aligned	Text Response
Reliably Consistent	8288 (69.1%)	37 (0.3%)	0 (0.0%)
Largely Consistent	1523 (12.7%)	258 (2.2%)	13 (0.1%)
Moderately Consistent	881 (7.3%)	863 (7.2%)	75 (0.6%)
Not Consistent	62		

Table 2: GPT-3.5's Alignment to Norms in the Moral Stories Dataset across Different Consistency Categories

	Aligned	Not Aligned	Text Response
Reliably Consistent	You shouldn't practice reckless driving.	It's good to surprise people for their birthdays	NA
Largely Consistent	It's unseemly to talk about politics with patients.	It's good to go out and socialize	It's commendable to have frank discussions about human sexuality.
Moderately Consistent	It's good to ask neighbors for help.	It's good to eat with other people.	You shouldn't flirt with coworkers.
Not Consistent	It is wrong to waste company time.		

Table 3: Examples of Norms across Consistency Categories and Types of Responses

Conversely, a very small percentage (0.3%) of norms were found to be both reliably consistent and not aligned, while 2.2% were largely consistent and not aligned. These represent norms that GPT-3.5 consistently fails to align with, suggesting that the model does not express such norms. Additionally, 0.1% of prompts return text responses consistently despite explicit instructions to return a '1' or '0'. These prompts represent instances where the model consistently failed to respond to the prompt appropriately.

A significant proportion (15.2%) of prompts returned only moderately consistent responses across the three response types, with a fairly even split between aligned and not aligned. For prompts in this category, users will need to collect a sufficiently large sample of responses to accurately characterize the model's typical response. Otherwise, with an insufficient sample size, users might erroneously conclude that the model is aligned or not aligned when it is simply expressing variability in its responses.

Prompts that cannot garner a majority response of any type (for example, a split of 40%, 30%, and 30% across all three types of responses) were labeled as 'Not Consistent.'

While it is possible to divide prompts in this category according to the response type with the largest frequency, we argue that the model's response cannot be characterized if no response type exceeds a consistency percentage of 50%.

Analyzing consistency categories across response types, most aligned prompts also belonged to the reliably consistent category. This suggests that when the model is aligned, they are also reliably consistent, possibly indicating a level of certainty or confidence. However, a significant percentage of not-aligned prompts fall into the moderately consistent category, implying that when the model is not aligned, they also exhibit uncertainty.

	Aligned	Not Aligned	Text Response
Reliably Consistent	69.1%	18.2%	
Largely Consistent	12.7%		
Moderately Consistent			
Not Consistent			

Table 4: Percentages in the 'Desirable' (light grey) vs 'Undesirable' (white) Categories. The combined percentages in the desirable categories (69.1% + 12.7% = 81.8%) can function as a human-AI value alignment metric that accounts for both alignment and consistency.

As highlighted in Table 4, prompts that are both aligned and reliably consistent are the most desirable, particularly because they represent values that do not require further checking or intervention. For GPT-3.5, this constitutes only 69.1% of norms in the Moral Stories dataset. Expanding the definition of 'desirable' to include aligned prompts that are largely consistent increases this percentage to 81.8%, leaving 18.2% of norms in the 'undesirable' category. Although an argument can be made that aligned and moderately consistent responses can be considered desirable, the presence of significant variability in responses complicates this decision.

Classifying Norms into Higher Order Values

While an aggregate metric across the Moral Stories dataset might provide an insight into the model's alignment with the dataset, a theoretic argument must still be made if one were to claim alignment with the general idea of human values itself. A model's value profile grounded in a comprehensive cross-cultural taxonomy with an accompanying classifier, such as by Kiesel et al. (2022), could be incorporated within a general human-AI value alignment benchmark. The classifier model categorized norms into four levels of taxonomy, with the first level comprising 54 human values, while level 2, one level higher in abstraction, contains 20 value categories.

The classification of norms within the Moral Stories dataset reveals an unbalanced distribution among the primary value categories. Notably, a substantial portion of the norms, accounting for 39.0%, could not be assigned to any category. Additionally, out of 54 categories, 31 had less than 100 norms each, and 13 categories had no norms assigned to them. This disparity presents difficulties in forming a complete profile of value alignment. To ensure robust results, Table 5 presents the human-AI value alignment metric for Level 1 values with at least 100 norms. When evaluated with a higher abstraction value taxonomy, the human-AI value alignment metric can provide a strategic overview that enables developers to prioritize and direct their alignment

Level 1 Values (Kiesel et al. 2022)	Number of Norms	Human-AI value alignment metric (Overall - 81.8%)
Have a sense of belonging	729	69.7%
Be capable	118	77.1%
Be loving	476	77.3%
Have the own family secured	829	79.3%
Be responsible	353	79.9%
Have success	136	80.1%
Have a comfortable life	1160	81.2%
Have freedom of thought	178	81.5%
Have wealth	191	82.7%
Be respecting traditions	554	83.9%
Have an objective view	139	84.9%
Have good health	435	85.1%
Be compliant	670	85.1%
Have a safe country	290	85.2%
Be behaving properly	2180	86.2%
Have a stable society	145	86.2%
Have freedom of action	182	86.8%
Have harmony with nature	320	87.2%
Be protecting the environment	119	87.4%
Be helpful	619	87.7%
Have equality	250	88.0%
Be just	781	90.0%
Be polite	235	92.3%

Table 5: Human-AI Value Alignment Metric for Level 1

efforts more effectively. By aggregating the values into broader categories, this approach reduces the granularity that characterizes the norm level, thereby simplifying the identification of key areas where the model demonstrates substantial misalignments. Specifically, values such as 'Have a sense of belonging,' 'Be capable,' 'Be loving,' 'Have the own family secured,' and 'Be responsible'—which are among those with the lowest scores—indicate where enhancements are most urgently needed. This aggregated view facilitates a more focused and manageable approach to improving alignment, allowing developers to concentrate on broad value groups rather than individual, detailed norms.

Level 2 Categorization (Value Category) displayed in Table 6 showed more robust results, with only 1 out of the 20 value categories having less than 100 norms. With only one underrepresented value category at level 2, we can present a fuller picture of alignment which leads to a comprehensive latent value profile for GPT-3.5. Certain value categories, such as 'hedonism' and 'stimulation,' exhibit significantly lower alignment metrics compared to the overall model, which could potentially point to areas for improvement.

Level 2 Value Categories (Kiesel et al. 2022)	Number of Norms	Human-AI value alignment metric (Overall - 81.8%)
Hedonism	245	51.8%
Stimulation	320	55.9%
Self-direction: thought	149	73.8%
Achievement	707	75.7%
Face	325	78.8%
Security: personal	2521	79.5%
Universalism: objectivity	247	80.2%
Power: resources	237	80.2%
Benevolence: dependability	1076	80.7%
Humility	230	80.9%
Self-direction: action	2091	81.0%
Tradition	630	81.1%
Benevolence: caring	3851	81.5%
Conformity: rules	3217	83.8%
Security: societal	265	84.2%
Universalism: tolerance	331	84.3%
Universalism: nature	442	84.6%
Universalism: concern	681	88.7%
Conformity: interpersonal	470	89.4%

Table 6: Human-AI Value Alignment Metric for Level 2

On the other hand, value categories with higher alignment metric values like 'conformity: rules' and 'benevolence: caring' could still be at an insufficient level for developers who may prefer even tighter alignment, especially if their use cases require it. Similar to level 1, when evaluated at a higher abstraction of values, the latent value profile offers a strategic overview that enables developers to prioritize and direct their alignment efforts more effectively.

Discussion

GPT-3.5's Consistency and Value Alignment

GPT-3.5 is aligned with the majority of the 12,000 norms from the Moral Stories dataset, which likely reflects the broad-based nature of its training data, aligning it with most human values. However, where the norms are not aligned are equally intriguing as they point to potential areas for further alignment.

Additionally, the norms that lack consistency are noteworthy, as one might expect identical responses to identical prompts. However, this paper shows that many prompts do not exhibit reliable consistency, leading to variable outputs, a particularly noteworthy insight given that the prompts are multiple-choice.

For norms where reliable consistency (100% consistency) was observed, model developers might not need to intervene, as it can be reasonably assumed that the model will consistently produce aligned outputs for the same prompts.

Conversely, for values demonstrating less reliable consistency, particularly in high-impact situations like healthcare, it becomes crucial for developers to devise strategies to address these inconsistencies. Such strategies might include less intrusive measures like developing supplementary models that identify and filter out misaligned outputs before they reach users. Alternatively, calculating the average output over a large sample size may help identify a predominantly aligned response. These measures are implemented as an additional layer external to the model to analyze its responses for alignment. However, in cases where no consistent response is achieved, more significant interventions could be necessary. This may involve directly changing the model through fine-tuning with specialized datasets to ensure alignment or redeveloping the model after sanitizing the training data to ensure that any data negatively influencing value alignment is excluded.

From Value Alignment to Understanding Model Behavior

By making human-value alignment transparent, the latent values of GPT-3.5 can be benchmarked, revealing areas where the model aligns well with human norms and identifying deficiencies for improvement in future iterations. Additionally, determining value alignment may offer insights into potential applications for models within decision-making systems. For instance, the norm "you shouldn't practice reckless driving" received a rating indicating both alignment and reliable consistency, suggesting the potential utility of GPT-3.5 in transportation, possibly in hybrid systems augmenting human drivers or in autonomous vehicles. Conversely, GPT-3.5 demonstrated consistent non-alignment with the norm "It's good to go out and socialize," suggesting it may be less suitable for applications requiring social interaction. Essentially, this process of "interviewing" language models serves to increase the transparency and explainability of an entity typically seen as a black box.

Furthermore, the aggregation of 12,000 norms against a cross-cultural taxonomy of human values facilitates a comprehensive assessment of alignment with more abstract human values. This high-level monitoring, which provides an overview of broad areas of human values, is particularly valuable in contexts where specific norms are less critical and a general assessment of value alignment is more practical. More importantly, evaluating models against a taxonomy that is not only employed cross-culturally but also derived from literature enhances the comprehensiveness of the alignment assessment, ensuring that our understanding of the model's behavior encompasses critical aspects of human values.

While not conducted in this paper, the implementation of similar measures across different language models naturally facilitates comparative analyses, enabling the identification of foundational models most closely aligned with human values. It's important to note that the analysis presented here operates across multiple levels of abstraction, thereby enabling evaluations of models at varying degrees of specificity. This approach not only allows for general or specific comparisons among language models to determine their appropriateness for diverse applications but also supports the comparison of successive fine-tuned iterations during development. Such comparisons could be used to verify that the fine-tuning process is on track and has not introduced any deviations from intended alignments.

Determining a model's value alignment profile and its consistency is crucial for enhancing the transparency and explainability of AI models. This enhanced clarity not only deepens our understanding of the model but also has significant practical implications by clarifying potential impacts on model behavior. Benchmarking alignment thus serves as an essential tool for penetrating the black box, illuminating the internal mechanics of these models. In doing so, it facilitates their effective integration into real-world applications, thereby ensuring the ethical deployment of AI.

Who Bears Responsibility for Value Alignment?

The question of who is responsible for value alignment, as well as broader questions pertaining to responsible AI and AI governance, is a multifaceted issue examined by multiple parties within the AI ecosystem. Jobin et al. (2019) noted AI ethics guidelines from stakeholders like international bodies, regulatory agencies, and industry players. However, while developers may be seen as uniquely accountable due to their technical expertise and foundational role in model development, as we have explored in this paper, the concept of 'value alignment'—particularly in defining and understanding its relevance to specific contexts—demands a broader range of expertise beyond the AI technical domain. This perspective is echoed by Gabriel (2020), who noted that "normative and technical aspects of the AI alignment problem are interrelated, creating space for productive engagement between people working in both domains."

Similarly, in a roundtable convened by AI Singapore on the allocation of responsibility for AI systems, stakeholders from academia, industry, and government identified several key parties responsible for AI systems: developers, deployers, users, governments, and academic institutions. The report highlighted solutions that extend beyond technical measures, encompassing market-driven approaches, legal instruments such as contracts, and regulatory frameworks (AI Singapore 2023). This broader perspective raises further questions on the role measuring value alignment can play within the larger AI ecosystem, such as who should determine alignment and whether the insights obtained from determining alignment are sufficiently actionable. Therefore, the question of responsibility for value alignment cannot be looked at technically in isolation.

Limitations and Future Research

Generalizability Limitations

It is crucial to acknowledge that the findings from this study cannot be generalized to other models, including those with similar architectures, such as GPT-4. Responses from models may vary significantly due to factors such as fine-tuning or differing training datasets. The diverse landscape of LLMs thus necessitates individualized studies for each model, where the analysis should not only include establishing a human-AI value alignment metric but also establishing the latent value profile suitable to the relevant local context.

Limitations of Sample Size in Determining Consistency

This study utilized a sample size of 100, where 69.1% of norms from the Moral Stories dataset exhibited a reliably consistent alignment (i.e., achieving 100% consistency). However, as the underlying architecture of the model is still probabilistic in nature, variability in responses remains an inherent possibility. In line with statistical best practices, increasing the sample size can increase the confidence level and reduce the margin of error. Therefore, when deploying language models in specialized sectors, such as manufacturing or medicine, adherence to established industry-specific guidelines is recommended. Future research could thus employ the methodology introduced here to assess value alignment under varying sector-specific guidelines.

Limitations of the Dataset and Classification Model

Classifying the model into the values and categories by Kiesel et al. (2022) showed that the Moral Stories dataset does not have even representation across all values, in particular at Level 1. Thus, more research is needed to develop datasets across all values in order to be able to generate a more comprehensive value alignment profile. Furthermore, we acknowledge that the model might not be developed specifically to classify norms. Further research could thus develop classification models that specifically cater to classifying norms into more abstract value categories.

In addition, although the Moral Stories dataset encompasses over 12,000 structured narratives corresponding to an equal number of norms, it was not designed for cross-cultural analysis. As noted by Emelin et al. (2021), the dataset predominantly reflects the perspectives of White, educated U.S. residents, potentially coloring the norms with experiences specific to this demographic. This raises two potential limitations: Firstly, different groups of people might express different stances on norms, and secondly, there are additional pertinent norms that were not included. Thus, it is plausible that each local context might require a tailored version of the Moral Stories dataset to accurately reflect its unique normative standards. Consequently, the analysis of value alignment should be conducted not only for individual models but for each relevant cultural grouping as well.

Conclusion

Surfacing the latent value profiles of LLMs offers a glimpse into the 'black box' of AI in an effort to uncover the drivers of its behavior. Assessing human-AI value alignment is thus essentially an exercise in explainability, where an attribute of a model is inferred through the analysis of model outputs. However, as we have demonstrated, value alignment also encompasses determining the consistency of the model's alignment. This dual nature of alignment was illustrated through the employment of repeated identical prompts that reflect human norms, where we have shown that identical prompts do not always elicit the same responses, thereby highlighting the inherent variability in model outputs.

We have proposed a framework that categorizes consistency into four practical levels for developers. GPT-3.5, for example, exhibits an overall consistency of 96.0%, classifying it as ‘largely consistent.’ However, analyzing its alignment in tandem with consistency yields an overall human-AI value alignment metric of 81.8%. Additionally, this study has classified norms from the Moral Stories dataset into broader abstract values, enabling developers to identify and prioritize areas for enhancement. While this does not directly facilitate the design of interventions, it highlights potential areas for improvement. Developers can leverage this insight to implement measures such as fine-tuning with additional data to further enhance alignment.

Ultimately, understanding a model’s value alignment means understanding its behavior. This, in turn, informs its applicability. Further research is essential to develop diverse datasets and classification models for assessing human-AI value alignment, but it is our hope that this paper serves as an initial step in illustrating how human-AI value alignment can be quantitatively assessed in a way that accommodates the intrinsic nature of both humans and AI.

Appendices

Underrepresented Values

Level 1

Be ambitious, Be broadminded, Be choosing own goals, Be courageous, Be creative, Be curious, Be daring, Be forgiving, Be holding religious faith, Be honest, Be honoring elders, Be humble, Be independent, Be intellectual, Be logical, Be neat and tidy, Be self-disciplined, Have a good reputation, Have a varied life, Have a world at peace, Have a world of beauty, Have an exciting life, Have influence, Have life accepted as is, Have loyalty towards friends, Have no debts, Have pleasure, Have privacy, Have social recognition, Have the right to command, Have the wisdom to accept others

Level 2

Power: dominance

References

AI Singapore. 2023. Responsibility for AI. aisingapore.org/wp-content/uploads/2024/02/Responsibility-for-AI_AISG-Roundtable-1_final.pdf. Accessed: 2024-07-25

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich J.; Shariff, A.; Bonnefon, J.; and Rahwan, I. 2018. The Moral Machine Experiment. *Nature* 563 (7729): 59–64. doi.org/10.1038/s41586-018-0637-6.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Brown, D.; and Kelly Crace, R. 2002. Life Values Inventory Facilitator's Guide. www.lifevaluesinventory.org/LifeValuesInventory.org%20-%20Facilitators%20Guide%20Sample.pdf. Accessed: 2024-07-25

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; de Oliveira Pinto, H.; Kaplan, J.; Edwards, H. et al. 2021. Evaluating Large Language Models Trained on Code. arXiv:2107.03374.

Chuang Y.; Tang, R.; Jiang, X.; and Hu, X. 2023. SPeC: A Soft Prompt-Based Calibration on Mitigating Performance Variability in Clinical Notes Summarization. arXiv preprint. arXiv:2303.13035

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; et al. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Emelin, D.; Le Bras, R.; Hwang, J.; Forbes, M.; and Choi, Y. 2021. Moral Stories: Situated Reasoning about Norms, Intents, Actions, and Their Consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Dominican Republic: Association for Computational Linguistics. doi.org/10.18653/v1/2021.emnlp-main.54.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30 (3): 411–37. oi.org/10.1007/s11023-020-09539-2.

Gu, H.; Degachi, C.; Genç U.; Chandrasegaran S.; and Verma H. 2023. On the Effectiveness of Creating Conversational Agent Personalities Through Prompting. arXiv.2310.11182.

Haerpfer, C.; Inglehart, R.; Moreno, A.; Welzel, C.; Kizilova, K.; Diez-Medrano J.; Lagos, M.; Norris, P.; Ponarin, E.; and Puranen B. 2022. World Values Survey Wave 7 (2017-2022) Cross-National Dataset. doi.org/10.14281/18241.20. Accessed: 2024-07-25

Hendrycks, D.; Burns C.; Basart S.; Zou A.; Mazeika M.; Song D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. arXiv.2009.03300.

Jiaming J.; Qiu T.; Chen B.; Zhang, B.; Lou, H.; Wang K.; Duan, Y.; et al. 2024. AI Alignment: A Comprehensive Survey. arXiv.2310.19852.

Jobin, A.; Ienca, M; Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1 (9): 389–99. doi.org/10.1038/s42256-019-0088-2.

Kant, I. 1785. Groundwork for the Metaphysic of Morals. In the version presented by Jonathan Bennett at www.earlymoderntexts.com. Accessed: 2024-07-25

Kiesel, J.; Alshomary, M.; Handke, N.; Cai, X.; Wachsmuth, H.; and Stein B. 2022. Identifying the Human Values behind Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4459–71. Dublin, Ireland: Association for Computational Linguistics. doi.org/10.18653/v1/2022.acl-long.306l

Kirk, H.; Vidgen, B.; Röttger, P.; and Hale, S. 2024. The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals. *Nature Machine Intelligence* 6 (4): 383–92. doi.org/10.1038/s42256-024-00820-y.

Lin, S.; Hilton J.; and Evans O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv. 2109.07958.

Noda, R.; Izaki, Y.; Kitano, F.; Komatsu, J.; Ichikawa, D., and Shibagaki, Y. 2024. Performance of ChatGPT and Bard in Self-Assessment Questions for Nephrology Board Renewal. *Clinical and Experimental Nephrology* 28 (5): 465–69. doi.org/10.1007/s10157-023-02451-w.

OpenAI Platform. 2024. Text Generation Models. https://platform.openai.com. Accessed: 2024-07-25

Rawls, J. 1971. A Theory of Justice: Original Edition. Harvard University Press. doi.org/10.2307/j.ctvjf9z6v.

Rokeach, M. 1973. The Nature of Human Values. New York,: Free Press.

Schwartz, H.; Cieciuch, J.; Vecchione, M.; Davidov, E.; Fischer, R.; Beierlein, C.; Ramos, A. et al. 2012. Refining the Theory of Basic Individual Values. *Journal of Personality and Social Psychology* 103 (4): 663–88. doi.org/10.1037/a0029393.

Vasiliou, I. 2011. Aristotle, Agents, and Actions. In *Aristotle's Nicomachean Ethics: A Critical Guide*, edited by Jon Miller, 170–90. Cambridge Critical Guides. Cambridge: Cambridge University Press. doi.org/10.1017/CBO9780511977626.009.

Wiener, N. 1960. Some Moral and Technical Consequences of Automation. *Science* 131 (3410): 1355–58.

Ziems, C.; Yu, J.; Wang, Y.; Halevy, A.; and Yang, D. 2022. The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3755–73. Dublin, Ireland: Association for Computational Linguistics. doi.org/10.18653/v1/2022.acl-long.261.