# Measuring Human-AI Value Alignment in Large Language Models

Norakmal Hakim Bin Norhashim

*(B.Eng., Singapore University of Technology and Design)*

A DISSERTATION SUBMITTED FOR THE DEGREE OF
MASTER OF COMPUTING
SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF SINGAPORE

2024

Supervisor: Professor Jungpil Hahn

# Declaration

I hereby declare that this dissertation is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the dissertation. This dissertation has also not been submitted for any degree in any university previously.

Norakmal Hakim Bin Norhashim
10 November 2024

# Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Hahn Jungpil, for his invaluable guidance throughout this study, my dissertation, and my broader academic journey. His support has been instrumental not only in shaping this work but also in broader areas of my life, especially in preparing me for future career growth.

I would also like to acknowledge Assistant Professor Lim Shiying, whose class laid a strong foundation in academic writing as well as equipping me with essential skills for research.

Finally, I extend my thanks to the School of Computing for the opportunity to pursue my master's degree with the support of a Study Grant.

# Table of Contents

# Abstract


## Measuring Human-AI Value Alignment in Large Language Models

**By**

**Norakmal Hakim Bin Norhashim**

**Master of Computing**

**National University of Singapore**

This dissertation 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 dissertation 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. Finally, this dissertation also performed a quantitative assessment to determine the extent to which higher-level human values serve as significant determinants of AI outputs, specifically in relation to consistency and value alignment.

# List of Figures, Tables, and Listing

# 1. Introduction

This study 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 study is thus: How can the human-AI value alignment of a large language model be quantitatively measured?

This study 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 study 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 value of this study's framework is demonstrated by applying it to GPT-3.5 to calculate its human-AI value alignment metric and establish its latent value profile.

- A quantitative assessment of the extent to which higher-level human values serve as significant determinants of AI outputs, specifically in relation to consistency and value alignment.

# 2. Literature Review

## 2.1 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.

## 2.2 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 models' responses to standardized prompts using multiple-choice formats derived from standardized datasets. Analysis like this 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 et al. 2022), mathematics (Cobbe et al. 2021), and coding (Chen et al. 2021).

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 experiments 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 the 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 can be 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.

## 2.3 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 also how inherent variability exists 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 temperature 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.

## 2.4 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 study 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 study, identical) inputs, thus highlighting the

importance of consistency in alignment across identical situations. This study 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 study.

## 2.5 Summary

This study, 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.'

# 3. Methodology

## 3.1 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:

(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.

**Listing 1: Prompt Template and Example**

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 (Level 1 and Level 2). 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. Therefore, the results of this study must be considered in this context to avoid overgeneralizing the findings and to highlight the need for more culturally diverse datasets.

Regression analysis was then performed between Level 1 values as the independent variable and consistency percentages defined as the defined as the percentage of the majority response as the dependent variable. Beta regression was chosen due to the bounded nature of the dependent variable between 0 and 1 and the right-skewed distribution of this variable (see Figure 1). The analysis included all 12,000 prompts, as all 12,000 prompts received 100 responses each. However, only Level 1 values with sufficient sample size ($n > 100$) were retained in the final model.

Initially, ordinary least square (OLS) regression was considered for this analysis. However, the fact that the dependent variable is bounded, as well as diagnostic tests, such as the non-normality of residuals using Q-Q plots and Shapiro-Wilk Test, as well as the Breusch-Pagan Test to test for heteroscedasticity, conducted that showed significant deviations with basic assumptions, led to the decision not to use OLS regression. Multi-collinearity checks were also conducted using a Variance Inflation Factor (VIF) analysis.

Regression analysis was also done between Level 1 values as independent variables with alignment, where a response of 1 indicates alignment and 0 otherwise as the dependent variable. Given the binary nature of the dependent variable, a logistic regression model was applied to assess the association between the presence of Level 1 values and value alignment. Observations containing string responses were excluded from the analysis, resulting in a final dataset of 11,901 prompts.

# 4. Results

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

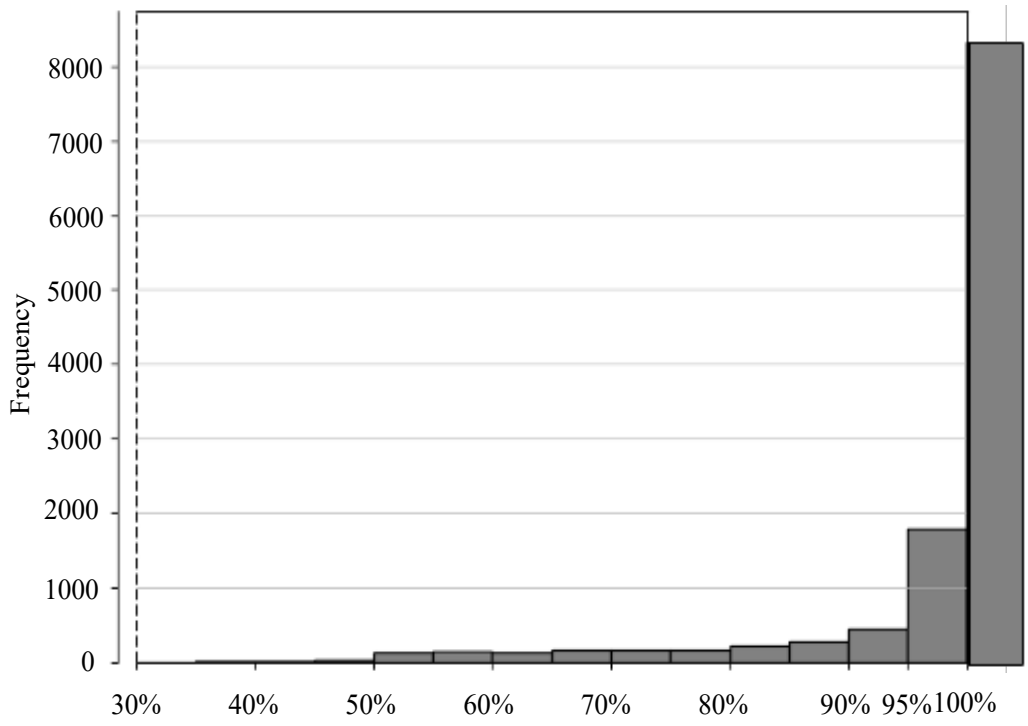


**Figure 1: Histogram Showing the Distribution of the Number of Prompts by Consistency Percentage. The rightmost bar represents a bin consisting of values at exactly 100% and has been extended beyond the axis for visual clarity.**

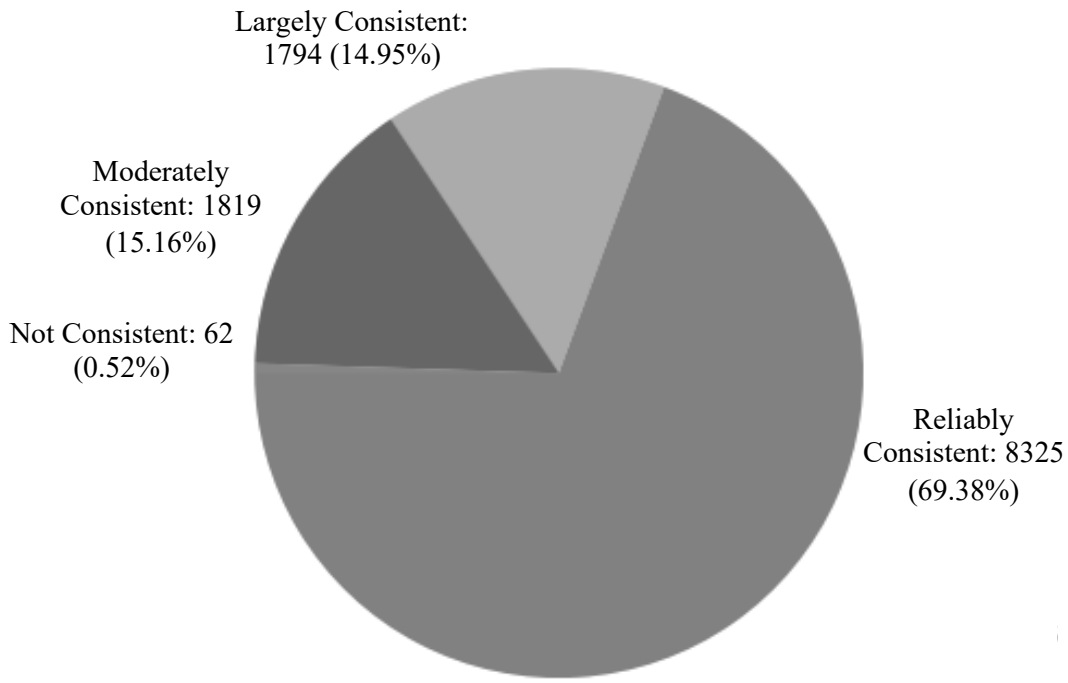


**Figure 2: Consistency Percentages of GPT-3.5's Responses by Consistency Categories**

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.**

## 4.2 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 GPT-3.5 consistently aligns with the majority of moral norms within the Moral Stories dataset. Furthermore, a significant proportion of these norms (12.7%) demonstrated alignment and considerable consistency, underscoring the model's overall reliability.

| | **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.

## 4.3 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 efforts more effectively.

| **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**

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’— 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.

The Level 2 Categorization (Value Category) illustrated in Table 6 demonstrates more substantial results, with only one of the twenty value categories exhibiting fewer than one hundred norms. Given that only one value category is underrepresented at Level 2, we can present a complete depiction of alignment, which results in a thorough latent value profile for GPT-3.5. Specific value categories, such as 'hedonism' and 'stimulation,' display considerably lower alignment metrics compared to the overall model, indicating potential avenues 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**

Conversely, value categories that show greater alignment metrics, such as 'conformity: rules' and 'benevolence: caring' might still be inadequate for developers seeking even stronger alignment, particularly if their specific use cases demand it. As with level 1, assessing values at a higher abstract level presents a latent value profile that provides developers with a strategic perspective, helping them prioritize and focus their alignment efforts more effectively.

## 4.4 Drivers of Consistency and Value Alignment

While the higher-level values can function as descriptive indicators to highlight areas for developers to investigate and develop interventions, it remains to be seen whether the higher values

themselves are the drivers of consistency and value alignment. This study thus conducts a regression analysis to investigate the relationship between level 1 values and the two contributors to the human-AI value alignment metric: consistency and value alignment.

### 4.4.1 Regression Analysis Between Level 1 Values and Consistency

A beta regression model with a logit link function was employed to examine the association between the presence of Level 1 values and response consistency, defined as the percentage of the majority response. Beta regression was chosen due to the bounded nature of the dependent variable between 0 and 1 and the right-skewed distribution of this variable (see Figure 1). The analysis included all 12,000 prompts. However, only values with sufficient sample size (n > 100) were retained in the final model. A Variance Inflation Factor (VIF) analysis was also conducted to ensure no multi-collinearity.

Model Fit Statistics

- Log-Likelihood: 52385
- McFadden's Pseudo R-squared: -0.0005343873781924469

Regression Results

| Variable | coef | std err | z | P>\|z\| | [0.025 | 0.975] | VIF |
|---|---|---|---|---|---|---|---|
| const | 2.9139 | 0.022 | 134.115 | 0.000 | 2.871 | 2.956 | 2.052903 |
| Have a sense of belonging | -0.1620 | 0.041 | -3.943 | 0.000 | -0.243 | -0.081 | 1.088564 |
| Be capable | -0.0539 | 0.099 | -0.544 | 0.587 | -0.248 | 0.140 | 1.077091 |
| Be loving | -0.0078 | 0.058 | -0.136 | 0.892 | -0.121 | 0.105 | 1.439936 |
| Have the own family secured | -0.0123 | 0.043 | -0.283 | 0.777 | -0.097 | 0.073 | 1.37653 |
| Be responsible | -0.0316 | 0.056 | -0.563 | 0.574 | -0.142 | 0.079 | 1.025529 |
| Have success | -0.0267 | 0.094 | -0.283 | 0.777 | -0.212 | 0.158 | 1.118435 |
| Have a comfortable life | 0.0051 | 0.036 | 0.144 | 0.886 | -0.065 | 0.075 | 1.244957 |
| Have freedom of thought | -0.0169 | 0.081 | -0.208 | 0.835 | -0.176 | 0.142 | 1.090304 |
| Have wealth | 0.0231 | 0.078 | 0.298 | 0.766 | -0.129 | 0.176 | 1.067663 |
| Be respecting traditions | 0.0363 | 0.046 | 0.79 | 0.429 | -0.054 | 0.126 | 1.052925 |
| Have an objective view | 0.0103 | 0.090 | 0.115 | 0.908 | -0.165 | 0.186 | 1.04788 |
| Have good health | 0.0625 | 0.051 | 1.234 | 0.217 | -0.037 | 0.162 | 1.018504 |
| Be compliant | -0.0098 | 0.046 | -0.213 | 0.831 | -0.100 | 0.080 | 1.26382 |
| Have a safe country | 0.0393 | 0.065 | 0.603 | 0.546 | -0.088 | 0.167 | 1.151928 |
| Be behaving properly | 0.0674 | 0.028 | 2.424 | 0.015 | 0.013 | 0.122 | 1.300248 |
| Have a stable society | 0.0223 | 0.091 | 0.246 | 0.806 | -0.155 | 0.200 | 1.123004 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Have freedom of action | 0.0587 | 0.080 | 0.738 | 0.461 | -0.097 | 0.215 | 1.068498 |
| Have harmony with nature | 0.0942 | 0.074 | 1.275 | 0.202 | -0.051 | 0.239 | 1.610659 |
| Be protecting the environment | 0.0466 | 0.119 | 0.391 | 0.695 | -0.187 | 0.280 | 1.582714 |
| Be helpful | 0.1293 | 0.047 | 2.736 | 0.006 | 0.037 | 0.222 | 1.228884 |
| Have equality | 0.0875 | 0.071 | 1.237 | 0.216 | -0.051 | 0.226 | 1.172336 |
| Be just | 0.0898 | 0.042 | 2.159 | 0.031 | 0.008 | 0.171 | 1.206528 |
| Be polite | 0.1427 | 0.068 | 2.083 | 0.037 | 0.008 | 0.277 | 1.026839 |
| precision | 1.2272 | 0.020 | 60.181 | 0.000 | 1.187 | 1.267 | N.A. |

**Table 7: Beta Regression Results between Level 1 Values and Consistency**

VIF values of below 2 were observed for all the independent variables, suggesting no multi-collinearity is occurring. The Level 1 values ‘Have a sense of belonging,’ ‘Be behaving properly,’ ‘Be helpful,’ ‘Be just,’ and ‘Be polite’ were found to be statistically significant at an alpha level of 0.05, with corresponding coefficients of -0.1620, 0.0674, 0.1293, 0.0898, and 0.1427, respectively. However, it is crucial to highlight that the model’s pseudo R-squared value of -0.000534 indicates that the predictors account for only a negligible portion of the variance in the model's response consistency. Similar pseudo R-squared values were observed with alternative link functions (Probit and Complementary Log-Log). This suggests a complex interplay of relationships among the variables and potential omitted variables, or, in this author's view, likely, higher-level abstraction values do not serve as significant determinants of the model's consistency.

### 4.4.2 Regression Analysis Between Level 1 Values and Value Alignment

A logistic regression model was applied to assess the association between the presence of Level 1 values and value alignment, where a response of 1 indicates alignment and 0 otherwise. Observations containing string responses were excluded from the analysis, resulting in a final dataset of 11,901 prompts. A VIF analysis was also conducted to ensure no multi-collinearity.

Model Fit Statistics

- Log-Likelihood: -3788.6
- McFadden's Pseudo R-squared: 0.01787245709375962

Regression Results

| Variable | coef | std err | z | P>\|z\| | [0.025 | 0.975] | VIF |
|---|---|---|---|---|---|---|---|
| const | 2.1255 | 0.043 | 49.277 | 0.000 | 2.041 | 2.210 | 2.054135 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Have a sense of belonging | -0.5710 | 0.111 | -5.140 | 0.000 | -0.789 | -0.353 | 1.088763 |
| Be capable | -0.5419 | 0.266 | -2.039 | 0.041 | -1.063 | -0.021 | 1.080377 |
| Be loving | -0.0843 | 0.170 | -0.495 | 0.621 | -0.418 | 0.249 | 1.440873 |
| Have the own family secured | -0.1658 | 0.132 | -1.258 | 0.209 | -0.424 | 0.093 | 1.376888 |
| Be responsible | -0.4662 | 0.161 | -2.899 | 0.004 | -0.781 | -0.151 | 1.025997 |
| Have success | 0.2202 | 0.314 | 0.702 | 0.483 | -0.395 | 0.835 | 1.123671 |
| Have a comfortable life | -0.2172 | 0.108 | -2.012 | 0.044 | -0.429 | -0.006 | 1.24405 |
| Have freedom of thought | -0.0395 | 0.267 | -0.148 | 0.882 | -0.562 | 0.483 | 1.09249 |
| Have wealth | -0.0609 | 0.246 | -0.247 | 0.805 | -0.544 | 0.422 | 1.070342 |
| Be respecting traditions | 0.4082 | 0.179 | 2.284 | 0.022 | 0.058 | 0.758 | 1.052517 |
| Have an objective view | 0.7678 | 0.424 | 1.812 | 0.070 | -0.063 | 1.598 | 1.046688 |
| Have good health | 0.1142 | 0.171 | 0.670 | 0.503 | -0.220 | 0.449 | 1.018376 |
| Be compliant | 0.0592 | 0.171 | 0.347 | 0.729 | -0.275 | 0.394 | 1.262688 |
| Have a safe country | 0.1610 | 0.230 | 0.699 | 0.484 | -0.290 | 0.612 | 1.150162 |
| Be behaving properly | 0.3794 | 0.102 | 3.726 | 0.000 | 0.180 | 0.579 | 1.298455 |
| Have a stable society | 0.0035 | 0.309 | 0.011 | 0.991 | -0.603 | 0.610 | 1.121673 |
| Have freedom of action | 0.4226 | 0.298 | 1.419 | 0.156 | -0.161 | 1.006 | 1.070366 |
| Have harmony with nature | 0.3600 | 0.273 | 1.318 | 0.187 | -0.175 | 0.895 | 1.610828 |
| Be protecting the environment | -0.1054 | 0.427 | -0.247 | 0.805 | -0.942 | 0.731 | 1.582625 |
| Be helpful | 0.6646 | 0.172 | 3.867 | 0.000 | 0.328 | 1.002 | 1.227764 |
| Have equality | 0.3926 | 0.292 | 1.347 | 0.178 | -0.179 | 0.964 | 1.170659 |
| Be just | 0.5187 | 0.169 | 3.061 | 0.002 | 0.187 | 0.851 | 1.205543 |
| Be polite | 1.0665 | 0.363 | 2.938 | 0.003 | 0.355 | 1.778 | 1.026594 |

**Table 8: Logistics Regression Results between Level 1 Values and Value Alignment**

VIF values of below 2 were observed for all the independent variables, suggesting no multi-collinearity is occurring. The Level 1 values ‘Have a sense of belonging,’ ‘Be capable,’ ‘Be responsible,’ ‘Have a comfortable life,’ ‘Be respecting traditions,’ ‘Be behaving properly,’ ‘Be helpful,’ ‘Be just,’ and ‘Be polite’ were found to be statistically significant at an alpha level of 0.05, with corresponding coefficients of -0.5710, -0.5419, -0.4662, -0.2172, 0.4082, 0.3794, 0.6646, 0.5187, and 1.0665, respectively. However, it is similarly crucial to highlight that the model’s pseudo R-squared value of 0.01787 indicates that the predictors account for only a very small portion of the variance in the model's alignment. As with the previous regression results, this suggests a complex interplay of relationships among the variables, potential omitted variables, or, in this author's view, likely that higher-level abstraction values do not serve as significant determinants of the model's value alignment. It is worth noting, however, that while higher-level abstractions show limited explanatory power on the model’s

consistency and value alignment, they remain valuable for identifying potential problem areas that developers can then examine in greater detail at the lower level. The implications of the limited explanatory power of higher-level abstract values on the model's behavior are further explored in the discussion section under "The Problem of Abstraction in Human-AI Value Alignment."

# 5. Discussion

## 5.1 GPT-3.5's Consistency and Value Alignment

GPT-3.5 is aligned with most 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, areas 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 study shows that many prompts do not exhibit reliable consistency, leading to variable outputs, a particularly noteworthy insight given that the prompts are multiple-choice.

Model developers might not need to intervene for norms where reliable consistency (100% consistency) was observed, 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.

## 5.2 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 increases the transparency and explainability of an entity typically seen as a black box.

Furthermore, aggregating 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 study, 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.

## 5.3 What Does Consistency Mean for GPT-3.5?

As discussed in the literature review, consistency—or the lack thereof—in LLMs like GPT-3.5 is hindered by the inherent architecture of these models. Nonetheless, consistency is highly valued in certain domains, as observed by Chuang et al. (2023) and Noda et al. (2024). This brings about the need to determine if some prompts are more consistent than others, especially in those domains. This study also examines consistency from the perspective of moral philosophy, encapsulated in the concept of Input-Output Value Alignment Consistency. This concept views consistency through the lens of alignment with an independent framework of values across model outputs for comparable inputs. Additionally, this study proposes a framework with four categories to classify the extent of consistency in response to specific inputs or prompts across identical and independent attempts. However, the idea of consistency extends beyond value alignment and can be considered independently. Table 7 details the results of consistency across all four categories of the framework based on prompts constructed from the Moral Stories dataset.

| Consistency Category | Percentage of Moral Stories Prompts |
|---|---|
| Reliably Consistent | 69.38% |
| Largely Consistent | 14.95% |
| Moderately Consistent | 15.16% |
| Not Consistent | 0.52% |

**Table 9: Percentage of Moral Stories Prompts for Each Consistency Category**

### 5.3.1 Analyzing Consistency Categorization

Individual and Group Prompts

Each prompt belongs to a specific consistency category at the individual prompt level, illustrating the degree of confidence in generating consistent outputs from GPT-3.5. This reflects the model's inherent behavior toward that particular prompt, which cannot be inferred from a single prompting attempt. Grouping similar prompts—such as those organized by the value categories from Kiesel et al. (2022)—then reflects the model's broader behavior toward that category.

Understanding the consistency categories for individual or grouped prompts has important practical implications. Identifying which prompts belong to which consistency categories can help craft

policies or user guides. This can be particularly useful in domains where consistency is essential, such as healthcare, where patients expect consistent responses to identical queries. Similarly, in financial trading, users expect consistency when the model recommends certain actions (e.g., buying or selling a stock) based on the same inputs. A notable example from GPT-3.5 is the norm: "You shouldn't practice reckless driving." The norm was identified as being reliably and consistently aligned with human expectations. This consistency can instill confidence in deploying GPT-3.5 in-car systems, ensuring it doesn't deviate from promoting safe driving practices (within the limits of the testing employed by domain experts). Knowing the degree of consistency at individual or grouped prompt levels is therefore crucial for various applications due to the inherent variability in LLM outputs.

Overall Consistency for Each LLM

Examining the consistency of prompts across a comprehensive dataset, such as the Moral Stories dataset, provides an overall view of the model's consistency. For example, GPT-3.5 demonstrated that 69.38% of the prompts were 'reliably consistent,' suggesting that for a majority of norms, the model consistently produces aligned outputs in the domain of "human values," regardless of the nature of the alignment itself. This high level of consistency is critical for characterizing the alignment outcomes of a model, which is only possible with consistent outputs.

Understanding the overall consistency of a model helps developers and deployers assess the scale of interventions required to address inconsistency. For example, for prompts in the 'Moderately Consistent' category, it may be necessary to prompt the model multiple times with a suitable sample size to determine the "typical" output. Knowing that 15.16% of GPT-3.5 prompts in the Moral Stories dataset fall into this category gives deployers an idea of the scale of necessary mitigation efforts. If a significant portion of value-aligned prompts were 'Not Consistent' or 'Moderately Consistent,' developers might reconsider deploying such models in applications where consistency is crucial.

### 5.3.2 Variability in General-Purpose LLMs

However, in general-purpose models like GPT-3.5, where the application domains aren't explicitly defined, there may be cases where output variability is more desirable than consistency. For instance, users may prefer variability in generating artistic or fictional narratives even given identical

prompts., Even in value alignment, variability could be valuable. For example, the norm "It's good to eat with other people" was categorized as 'Moderately Consistent,' implying that GPT-3.5 allows for variability, where such variability can then be moderated by circumstances (included within the prompts) external to the values themselves. To illustrate, ChatGPT may display a high degree of variability when asked about the value "It's good to eat with **other people**." However, the variability might shift when the prompt is changed to portray the arguably more specific "It's good to eat with **your family**" instead.

A highly consistent model in response to specific prompts might not display this adaptability or could be "resistant" to changes in circumstances. More studies are needed to determine the extent of change in value alignment as a result of including circumstantial-based changes to the prompt. This highlights that not only is the nature of consistency of the model context-dependent, but our interpretations of it also vary across use cases and domains. Therefore, individual studies tailored to specific stakeholders and contexts are essential, where relevant domain stakeholders and experts perform benchmarking and interpret the results accordingly.

## 5.4 The Problem of Abstraction in Human-AI Value Alignment

As observed in the previous section, human values, as applied to use cases and operationalized as prompts, can exist in different levels of abstraction with seeming connections between them. At first glance, these connections might embody information structures reminiscent of typical linguistic structures of hypernyms and hyponyms where the different level of abstractions informs or provide information between each other. An entity of a lower level of abstraction would embody some characteristics of a higher level of abstraction. For example, cars (hyponym) are a type of vehicle (hypernym). Even if we do not know what cars are, knowing that cars are a type of vehicle tells us something about cars' characteristics, given we know what vehicles are. However, as we have observed, knowing a model's response at one level of abstraction does not inform us how the model might respond at a higher or lower level. These variances also account for the difference in the human-AI value alignment metric percentages between Kiesel et al. (2022) level 1 and level 2 human values, where despite having similar values, more specific operationalization of the prompts might produce different

results (i.e., some are aligned, and others not). Moreover, the regression analyses have shown that higher abstraction values are not significant determinants of how the model behaves in terms of consistency and alignment.

### 5.4.1 Two-way Disconnect

This study thus describes the inability to infer how a model might respond to a specific prompt when the subject is at a different level of abstraction as the *Two-way Disconnect Between Levels of Abstraction* where knowing how the response to the lower level of abstraction doesn't imply knowing how the model would respond to the higher level of abstraction, and vice-versa. The previous section introduced an example where the model might not necessarily identically respond to the values of "it's good to eat with **other people**" and "it's good to eat with **your family.**" In this instance, "family" represents a lower-level abstraction to the term "other people." This two-way disconnect thus implies that prompts across abstraction levels must be treated independently, even if the underlying semantics overlap. Treating prompts independently as a result of this disconnect then implies that, taken to the logical extreme, there is no possible taxonomy of prompts that can provide insights into how a model might respond beyond how the model responds directly to the prompt. We need to test each prompt separately.

### 5.4.2 Unnecessary Features of a Human-AI Value Alignment Framework

As a consequence of the two-way disconnect and the requirement to test each value independently, certain standard features in frameworks, including human value frameworks Kiesel et al. (2022), are unnecessary:

- **Hierarchy**

  The framework does not need to be constructed as a hierarchical system or taxonomy; values can exist independently at varying levels of abstraction without needing to be incorporated into a rigid structure.

- **Mutual exclusiveness**

  Values within the framework do not need to be conceptually mutually exclusive. The framework can accommodate overlapping values as long as they suit the specific use case or domain. For

instance, prompts based on values such as “be polite” and “be courteous " would be acceptable, even if these values have overlapping meanings.

### 5.4.3 Necessary Features of a Human-AI Value Alignment Framework

Similarly, since each value must be tested independently, only two features are essential:

- **Specificity**

  Prompts and their underlying values should be specified at the precise level of abstraction required by the use case. This enables responses to reflect the values relevant to that specific application accurately.

- **Comprehensiveness**

  The comprehensiveness of the framework should be consistent with the requirements of the use case. Developers engaged in model testing must ensure that the framework is meticulously detailed within the context of their specific domains, effectively encompassing the application of each value.

A value alignment framework characterized solely by specificity and comprehensiveness and lacking a taxonomic structure due to disparities in information structure across levels of abstraction suggests a context-dependent alignment rather than a universal one. While an overarching value alignment guides broad areas of intervention, it should be seen as a starting point for deeper exploration into more granular aspects. Therefore, the need for specificity calls for an alignment framework that requires a narrower focus. A crucial consideration is whether the model's response at a higher level of abstraction suffices for the particular use case. For example, is it sufficient for the developer to evaluate the model where values are only applied to scenarios involving "vehicles," or must the scenario more specifically apply to "cars"? Such practical considerations are essential to each use case and domain.

Earlier in the dissertation, the author notes that comprehensiveness is an essential advantage of working at a higher level of abstraction of values. However, it is essential to note that this advantage primarily exists in describing the inputs due to the black-box nature of large language models. This is because the two-way disconnect between levels of abstraction applies only to the model's output. Typical linguistic connections between hyponyms and hypernyms still apply at the input. For example,

the understanding that a car is a vehicle and thus possesses the attributes of a vehicle remains intact. What we do not know, however, is whether the model will respond similarly when prompted with "car" versus "vehicle." Thus, a framework can still be organized taxonomically, but only for human understanding, with the knowledge that each element in the value framework should be tested independently.

## 5.5 Do We Still Need AI Ethics Principles?

### 5.5.1 Operationalizing Ethics by Bringing Down the Level of Abstraction

Given that a model's response to high-level ethical values or principles fails to predict its behavior in response to more specific, granular prompts, broad AI ethics principles that aim to steer model outputs toward ethical outcomes may ultimately lack practical utility. However, the notion that AI principles should be more specific to be put into practice is not unique. When describing how ethics guidelines can be made more effective, Hagendorff states that "at certain points, a substantial change in the level of abstraction has to happen insofar as ethics aims to have a certain impact and influence in the technical disciplines and the practice of research and development of artificial intelligence." (Hagendorff 2020; Morley et al. 2020) This study's contribution further reinforces the need for increasing the level of granularity of ethical principles, particularly at the testing or benchmarking stage, by introducing the fact that testing for 'traditional' AI ethics principles of fairness, accountability, transparency, and explainability, among many others (Jobin et al. 2019), would similarly be subjected to the two-way disconnect between levels of abstraction, i.e., models that exhibit alignment when tested for values like "fair" for example, may not perform similarly when tested for more specific values like "equitable." Therefore, in operationalizing ethical principles, the principles themselves need to be adjusted for their granularity as warranted by the specifics of each use case.

### 5.5.2 The Limits of Abstraction for AI Explainability

Implicit in this notion that testing ethics necessitates adjusting for granularity is that the role of such broad-based ethical principles thus pertains not to describing the AI itself in general but to helping developers rationalize their inputs to AI to aid testing. Testing still works in that we are still able to observe how AI responds to specific prompts, but whether insights from that particular test can then be

*generalizable* when applied to other levels of abstractions or other conditions needs to be tested. The black-box nature of AI has thus blunted what might have been a traditional role played by generalization that once provided meaningful guidance where higher abstractions afford generalizations that, in turn, enable some form of predictability. Higher abstractions to the outputs can thus only be applied independently at the input or inductively by analyzing the outputs and then aggregating the results, taking on a retrospective and descriptive rather than a prescriptive role (the latter describes what this study had done). General aspects of prompts, such as broad-based AI ethics principles, can't predict how the AI might generate outputs. At the same time, outputs to specific prompts are not generalizable across multiple prompts without individual testing. This is reminiscent of the two-way disconnect between levels of abstraction described earlier. This same disconnect can thus be viewed through the lens of the predictive power of theory, where it can be defined as *the disconnect between generalizing inputs and the predictability of outputs.* This lens discusses the fact that due to the black box nature of AI, abstractions can only describe input and output independently but not how generalizing inputs can lead to predicting outputs, thus underscoring the limitations in taking a theoretical approach to guiding the outputs of AI. It is important to note, however, that while the regression analyses done in this study between higher abstractions of values and behavioral attributes of the model at the output provide some evidence of this disconnect in AI value alignment, this study cannot and does not claim that this disconnect exists universally. More research needs to be done to uncover *disconnects* in other domains like finance and medicine or any domains that rely on generalizing variables.

### 5.5.3 Principles for AI Alignment – A Suggested Path Forward

Perhaps there is scope to define principles for alignment that describe attributes that AI models of today may lack but which, in this author's view, are desirable for AI to have in the long run. These principles would be rooted in the two desirable dimensions of the human-AI value alignment metric: consistency and alignment. Additionally, they would address the two perspectives of disconnect identified so far—between levels of abstraction and between inputs and outputs— as well as incorporate specificity and comprehensiveness as necessary features of a human-AI value alignment framework. In consideration of all of these factors, this study proposes the following principles for AI Alignment:

- **Consistency – As applied to value alignment**

  In the near term, this principle emphasizes the need for predictable and stable responses to prompts used to test alignment. When AI responds consistently, alignment testing becomes more reliable, reinforcing the empirical foundation of the value alignment framework. However, in the long term, consistency is desirable as a core aspect of models, especially as it relates to value alignment.

- **Inherentness – The existence of inner states that reflect inherent value attributes**

  This principle underscores the significance of alignment tests in uncovering fundamental characteristics of AI rather than merely reflecting superficial or context-dependent behaviors. Although this study has demonstrated that such intrinsic nature cannot be presupposed and that testing is imperative, it has neither confirmed nor refuted the existence of innate attributes capable of applying values more broadly. Should prompts eventually reveal an inherent characteristic, the outcomes will possess greater significance, demonstrating alignment at a deeper level where the model's outputs will reflect the values even without specifying them in the input, an outcome that is arguably more desirable.

## 5.6 A Case for Case Studies Over Benchmarking

This study examines GPT-3.5 using a case-study methodology to gain insights into model behavior. It is important to highlight that this approach differs from the current model testing landscape, which often relies on standardized benchmarks that facilitate comparisons through numerical metrics leading to rankings. Nonetheless, the shortcomings of such benchmarking have been acknowledged. For instance, it has been pointed out that "contemporary models quickly achieve outstanding performance on benchmark tasks but nonetheless fail on simple challenge examples and falter in real-world scenarios" (Kiela et al. 2021) while Mishra et al. (2022) noted that "recent studies have shown that models triumph over several popular benchmarks just by overfitting on spurious biases, without truly learning the desired task." These critiques highlight the limitations of evaluations based solely on static datasets and aggregated model responses to generate numerical metrics. These metrics are then

compared across models where models achieving higher scores are ranked higher and interpreted as having better performance.

While this study similarly aggregates model responses from static datasets to create a metric, as indeed reflected in the title, "Measuring Human-AI Value Alignment," it acknowledges the significance of case studies in AI alignment by not solely benchmarking GPT-3.5 against other models based on numerical metrics alone. This is especially important if we consider that the specific areas to prioritize for improvements may differ despite models having comparable overall alignment metrics. Furthermore, this study emphasizes that numeric metrics are of lesser significance by breaking down and clarifying which aspects of value alignment require focus. However, despite guidance from the numerical metrics, it is vital to recognize that, particularly for a versatile model like GPT-3.5, the priorities regarding which areas to focus on are specific and should be evaluated for each application, even if the metrics imply otherwise. The overall human-AI value alignment metric should thus only be viewed as a measure of comprehensiveness, comparing current model versions against their predecessors to assess any improvements stemming from interventions.

## 5.7 Who Bears Responsibility for Value Alignment?

The question of who is responsible for value alignment, as well as broader questions about 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 study, 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.

## 5.8 Environmental Concerns Over Alignment Testing

The rapid rise and deployment of large language models (LLMs) have raised significant concerns regarding their energy and water consumption, alongside their broader environmental impact. For example, the development of models like GPT-3.5 requires extensive computational resources, leading to increased electricity usage and carbon emissions. Vries (2023) projects that by 2027, the AI sector’s annual energy consumption could rival that of the Netherlands, prompting some to question whether AI development 'should be scaled down.' (Pasquini 2024) The growing awareness and concerns for the environmental impact of LLMs have thus resulted in calls for AI to be managed more responsibly. For instance, the EU AI Act incorporates provisions that stress the importance of environmentally sustainable AI systems, with a focus on energy efficiency and responsible resource use across each stage of the AI system's lifecycle. (European Comission 2024)

The consistency framework utilized in this study, which entails repeated testing of identical prompts and subsequently aggregating responses in order to determine alignment, may thus conflict with the rising emphasis on energy conservation in AI development.

Nevertheless, the motivation behind this aspect of the study remains valid, as it seeks to address the inherent black-box and non-deterministic nature of LLMs. While developers may continue the current practice of overlooking the risks of model value misalignment, the similarly rising concern for responsible AI practices, alongside emerging regulatory measures, might suggest a growing necessity for developers to implement rigorous testing methods, in order to ensure model safety and value

alignment for users. This study thus proposes the following recommendations to mitigate the environmental impact of human-AI value alignment testing of large language models.

### 5.8.1 Recommendations for Reducing the Cost of Human-AI Value Alignment Testing

- **Limit Testing to Relevant Values**

  Developers should focus alignment testing on values relevant to the specific use case instead of attempting to encompass the entire spectrum of human values, which is neither necessary nor efficient.

- **Optimize Sample Size Based on Required Confidence Interval**

  This study employed a sample size of 100 for each prompt. However, if a use case allows for a lower confidence interval, particularly when consistency is not as important, developers can reduce the sample size accordingly to conserve computational resources.

- **Employ LLMs or Similar Transformer Models Selectively**

  Developers should consider deploying more deterministic, explainable, and transparent AI models where possible. For specific applications, alternative techniques, such as rules-based systems or traditional AI algorithms for text classification, offer greater insight into model operations, potentially limiting the need for extensive repetition-based testing to understand model behaviour.

## 5.9 Limitations and Future Research

### 5.9.1 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. Model responses may vary significantly due to factors such as fine-tuning or differing training datasets. Thus, the diverse landscape of LLMs necessitates individualized studies for each model. 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.

### 5.9.2 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.

### 5.9.3 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 should have been included. Thus, it is plausible that each local context might require a tailored version of the Moral Stories dataset to reflect its unique normative standards accurately. Consequently, the analysis of value alignment should be conducted not only for individual models but for each relevant cultural grouping as well.

# 6. Conclusion

Surfacing the latent value profiles of Large Language Models 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.

This study has 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 enhance alignment further. This study also conducted regression analyses between higher abstraction of values and model behavior of consistency and alignment at the output, which led to the development of concepts like the two-way disconnect and principles for AI alignment, which the author hopes can inspire future research in the field.

Ultimately, understanding a model’s value alignment means understanding its behavior, which in turn informs its applicability. Further research is essential to developing diverse datasets and classification models for assessing human-AI value alignment, but the author hopes that this study 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.

# References


AI Singapore. 2023. “Responsibility for AI,”, October 19. (https://aisingapore.org/wp-content/uploads/2024/02/Responsibility-for-AI_AISG-Roundtable-1_final.pdf).

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

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. 2022. *Constitutional AI: Harmlessness from AI Feedback*, arXiv. (https://doi.org/10.48550/arXiv.2212.08073).

Brown, D., and Crace, R. K. 2002. *Life Values Inventory Facilitator’s Guide.* (https://www.lifevaluesinventory.org/LifeValuesInventory.org%20-%20Facilitators%20Guide%20Sample.pdf).

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. 2021. *Evaluating Large Language Models Trained on Code*, arXiv. (http://arxiv.org/abs/2107.03374).

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

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. (https://doi.org/10.48550/arXiv.1803.05457).

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. 2021. *Training Verifiers to Solve Math Word Problems*, arXiv. (http://arxiv.org/abs/2110.14168).

Emelin, D., Le Bras, R., Hwang, J. D., 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*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih (eds.), Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, November, pp. 698–718. (https://doi.org/10.18653/v1/2021.emnlp-main.54).

European Comission. 2024. *Artificial Intelligence – Questions and Answers*. (https://ec.europa.eu/commission/presscorner/api/files/document/print/en/qanda_21_1683/QANDA_21_1683_EN.pdf).

Gabriel, I. 2020. “Artificial Intelligence, Values, and Alignment,” *Minds and Machines* (30:3), pp. 411–437. (https://doi.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. (https://doi.org/10.48550/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 Data-Set*, [object Object]. (https://doi.org/10.14281/18241.20).

Hagendorff, T. 2020. “The Ethics of AI Ethics: An Evaluation of Guidelines,” *Minds and Machines* (30:1), pp. 99–120. (https://doi.org/10.1007/s11023-020-09517-8).

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. 2021. *Measuring Massive Multitask Language Understanding*, arXiv. (http://arxiv.org/abs/2009.03300).

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O’Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S.-C., Guo, Y., and Gao, W. 2024. *AI Alignment: A Comprehensive Survey*, arXiv. (http://arxiv.org/abs/2310.19852).

Jobin, A., Ienca, M., and Vayena, E. 2019. “The Global Landscape of AI Ethics Guidelines,” *Nature Machine Intelligence* (1:9), pp. 389–399. (https://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.).

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. 2021. “Dynabench: Rethinking Benchmarking in NLP,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou (eds.), Online: Association for Computational Linguistics, June, pp. 4110–4124. (https://doi.org/10.18653/v1/2021.naacl-main.324).

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)*, Dublin, Ireland: Association for Computational Linguistics, May, pp. 4459–4471. (https://doi.org/10.18653/v1/2022.acl-long.306).

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. 2024. “The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals,” *Nature Machine Intelligence* (6:4), Nature Publishing Group, pp. 383–392. (https://doi.org/10.1038/s42256-024-00820-y).

Lin, S., Hilton, J., and Evans, O. 2022. *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, arXiv. (http://arxiv.org/abs/2109.07958).

Mishra, S., Arunkumar, A., Bryan, C., and Baral, C. 2022. *A Survey of Parameters Associated with the Quality of Benchmarks in NLP*, arXiv. (https://doi.org/10.48550/arXiv.2210.07566).

Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. 2020. “From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices,” *Science and Engineering Ethics* (26:4), pp. 2141–2168. (https://doi.org/10.1007/s11948-019-00165-5).

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), pp. 465–469. (https://doi.org/10.1007/s10157-023-02451-w).

“OpenAI Platform.” 2024. *Text Generation Models*, February 5. (https://platform.openai.com, accessed May 2, 2024).

Pasquini, N. 2024. *Should AI Be Scaled Down?* (https://www.harvardmagazine.com/2024/03/scaling-artificial-intelligence).

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

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

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., and Konty, M. 2012. “Refining the Theory of Basic Individual Values,” *Journal of Personality and Social Psychology* (103:4), US: American Psychological Association, pp. 663–688. (https://doi.org/10.1037/a0029393).

Vasiliou, I. 2011. “Aristotle, Agents, and Actions,” in *Aristotle’s Nicomachean Ethics: A Critical Guide*, Cambridge Critical Guides, J. Miller (ed.), Cambridge: Cambridge University Press, pp. 170–190. (https://doi.org/10.1017/CBO9780511977626.009).

Vries, A. de. 2023. “The Growing Energy Footprint of Artificial Intelligence,” *Joule* (7:10), Elsevier, pp. 2191–2194. (https://doi.org/10.1016/j.joule.2023.09.004).

Wiener, N. 1960. “Some Moral and Technical Consequences of Automation,” *Science* (131:3410), American Association for the Advancement of Science, pp. 1355–1358.

Ziems, C., Yu, J., Wang, Y.-C., 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)*, S. Muresan, P. Nakov, and A. Villavicencio (eds.), Dublin, Ireland: Association for Computational Linguistics, May, pp. 3755–3773. (https://doi.org/10.18653/v1/2022.acl-long.261).

# 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

# Publications During Study

Norhashim, H., and Hahn, J. 2024. "Measuring Human-AI Value Alignment in Large Language Models," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (7), pp. 1063–1073.