Data & Policy (2026), 8: e19
doi:10.1017/dap.2026.10061

RESEARCH ARTICLE

LLM for social good: A value-driven LLM framework to embed social good values

Victor O. K. Li[1], Jacqueline C. K. Lam[1], Yang Han[1], Jon Crowcroft[2] and Lawrence Y. L. Cheung[3]

[1]Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong
[2]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[3]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong
Corresponding authors: Jacqueline C.K. Lam and Victor O.K. Li; Emails: jcklam@eee.hku.hk; vli@eee.hku.hk

Received: 10 January 2025; **Revised:** 18 November 2025; **Accepted:** 27 February 2026

Keywords: ethical AI; generative AI; LLM; social good; value realignment

Abbreviation: AI, artificial intelligence; GAI, generative artificial intelligence; LLM, large language model

Abstract

As large language models (LLMs) increasingly shape decision-making, public discourse, education, healthcare, and governance, a critical question emerges: *Whose values are these systems truly reflecting?* While today's generative AI systems demonstrate extraordinary capabilities, they also inherit hidden biases, ethical blind spots, and implicit assumptions embedded within their training data. Existing alignment approaches often remain opaque, resource-intensive, or insufficiently adaptive to diverse societal expectations.

This paper introduces a novel value-driven LLM framework designed to systematically uncover, quantify, and realign the implicit values embedded within LLMs toward socially desirable outcomes. Built upon the AI for Social Good (AIfSG) framework, our proposed methodology operationalizes ethical alignment across six critical domains: reasoning and interpretability, bias removal, transparency and accountability, security and privacy, moral and ethical observations, and public understanding. Using advanced embedding techniques, cosine-based value-difference metrics, and topic-weighted iterative fine-tuning, the framework transforms ethical alignment from an abstract aspiration into a measurable and actionable computational process.

To demonstrate adaptability across moral and regulatory paradigms, the framework evaluates two distinct reference value systems: the Ten Commandments and the General Data Protection Regulation (GDPR). Experimental results using open-source LLMs, including Llama 3.2 and Gemma 2, reveal substantial reductions in value misalignment, ranging from approximately 25% to 70% across ethical domains, while preserving model flexibility and scalability. Visualizations in value-embedding space further confirm significant convergence between original model outputs and socially aligned reference values after iterative realignment.

Beyond technical innovation, this work positions value alignment as a foundational challenge for the future governance of generative AI. The proposed framework functions both as a diagnostic instrument for identifying ethical gaps and as an intervention mechanism for adaptive value realignment, offering policymakers, developers, and institutions a scalable pathway toward transparent, accountable, and socially responsible AI systems. By bridging computational methods with moral, legal, and societal principles, this research advances a new paradigm for embedding human-centered values directly into the next generation of intelligent systems.

V.O.K.L., J.C.K.L. and Y.H. equally contributed to methodology and writing.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

Policy Significance Statement

This research introduces a framework to ensure large language models (LLMs), a type of advanced generative AI, align with values that promote societal good. As LLMs become increasingly influential in decision-making and public-facing applications, their potential to reflect harmful biases or unethical assumptions poses risks. The proposed approach systematically identifies and measures value misalignments in AI outputs and iteratively adjusts them to realign with a desirable set of social good values. For policymakers, this framework provides a practical and quantifiable instrument to promote responsible AI development, minimize unintended harm, and sustain public trust. As a diagnostic tool, it offers measurable indicators, such as value difference metrics and topic-specific analyses, which can be incorporated into AI-impact assessments, compliance benchmarking, and ethical audits to identify where alignment gaps may pose societal risks. As an intervention tool, it operationalizes topic-weighted realignment to support targeted ethical adjustments, continuous auditing, and adaptive oversight, enabling policymakers to track shifts in value alignment over time and ensure ongoing accountability in generative-AI systems. By embedding social good values, defined as moral, social, and policy-oriented principles that advance human well-being and societal welfare, directly into AI development and governance processes, this framework advances the creation of AI policies that prioritize ethical innovation, transparency, and equitable societal outcomes.

1. Introduction

Generative artificial intelligence (GAI), particularly large language models (LLMs) like OpenAI's GPT series, has transformed how we process, generate, and interact with information (Bharathi Mohan et al., 2024). These systems have demonstrated remarkable capabilities across a range of applications, from creative writing and education to automating business processes and scientific research (De Michele W et al., 2025; Ren et al., 2025; Spitsberg et al., 2025). Despite these advancements, the rapid proliferation of LLMs raises critical concerns regarding their ethical and societal impacts (Weidinger et al., 2021; Gabriel et al., 2024).

GAI systems are trained on vast datasets that inherently reflect the biases, assumptions, and value systems present in their source materials (Navigli et al., 2023). As a result, these models can inadvertently perpetuate stereotypes, reinforce inequities, and produce outputs that conflict with societal norms or ethical principles (Gallegos et al., 2024). For example, GAI applications have been criticized for producing biased job recommendations (Salinas et al., 2023), reinforcing harmful stereotypes (Schramowski et al., 2022), and generating content that prioritizes functionality over ethical considerations (Weidinger et al., 2021). Such issues highlight the urgent need for robust frameworks to guide AI systems toward socially beneficial outcomes.

In this study, we focus specifically on the normative content of LLM outputs and their alignment with explicit reference value sets. This is only one layer of a broader socio-technical problem. Many important ethical issues, such as environmental costs, data provenance, and intellectual property rights (Ong et al., 2024), lie outside the scope of the framework proposed here. Moreover, decisions to adopt LLMs are shaped by institutional and political-economic logics (e.g., drives toward efficiency, cost reduction, or data-driven decision-making in public administration) (Madan and Ashok, 2023). Our framework should, therefore, be seen as a complementary diagnostic and intervention tool at the model output layer, to be combined with structural and institutional analyses rather than as a stand-alone solution to AI ethics.

Within this broader ethical landscape, a central challenge lies in value alignment: How can we ensure that the outputs and behaviors of AI systems reflect widely accepted ethical standards and social good or socially desirable values? State-of-the-art approaches, such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), address this value gap in GAIs to some extent. However, these methods are resource-intensive, often relying on human evaluations that may not capture a wide range of perspectives and lack the right mechanisms to systematically address topic-specific moral or ethical

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

challenges. A further challenge concerns value pluralism and contestation. Ethical values are not fixed or universally agreed upon; they vary across cultures, domains, and political traditions and may conflict with one another even within a single context (Jobin et al., 2019). These tensions make value alignment a complex and context-dependent endeavor, requiring sensitivity to competing priorities and to the plural moral landscapes in which AI systems operate (Gabriel, 2020).

This study introduces a novel value-driven framework designed to systematically embed social-good values, defined as moral, social, and policy-oriented principles that advance human well-being and societal welfare, into large language models (LLMs). Built on the AI for Social Good (AIfSG) topical framework, the analysis organizes ethical alignment across six domains: Reasoning and Interpretability, Bias Removal, Transparency and Accountability, Security and Privacy, Moral and Ethical Observations, and Public Understanding (Li et al., 2021, 2024). By uncovering implicit value assumptions, quantifying deviations from reference values, and iteratively realigning AI outputs, the proposed framework functions both as a diagnostic tool for identifying value misalignment and as an intervention tool for targeted value realignment.

The proposed value-driven framework assumes that a reference value set has been established through prior ethical, legal, or political deliberation and provides a systematic means of assessing how closely LLM outputs conform to that set. The same methodology can also be used to compare multiple value sets to reveal where a model aligns more closely with one than another. Two illustrative case studies, using the Ten Commandments as a moral-philosophical reference set and the General Data Protection Regulation (GDPR) as a policy-based reference set, demonstrate the framework's adaptability across normative and regulatory contexts, offering insights into its effectiveness and potential for broader implementation.

The motivation for this research is rooted in the growing role of GAI systems in decision-making processes and public-facing applications. As these LLM technologies become integral to governance, education, healthcare, and other societal domains, their ethical implications must be rigorously examined. GAI systems that lack alignment with social good values risk eroding public trust, exacerbating inequities, and causing unintended harm. Conversely, embedding ethical principles into AI systems can enhance transparency, accountability, and inclusivity, fostering positive societal outcomes.

This study is guided by the following research questions:

1. What implicit values are encoded within LLMs, and how can they be systematically uncovered?
2. How can the differences between these implicit values and reference value sets be quantified?
3. Which topics (themes or subject matter; see Table 1, for examples, of topics) generated by LLMs exhibit significant value misalignments, and how can these discrepancies be explained?
4. How can topic-specific misalignments be resolved through iterative realignment processes?

Table 1. *Topic and reference value set*

Topic	Ten Commandments	GDPR Principles (Article 5)
Reasoning and Interpretability	No. 9	Lawfulness, fairness, and transparency
Bias Removal	No. 9	Lawfulness, fairness, and transparency; Accuracy
Transparency and Accountability	No. 3, 9	Lawfulness, fairness, and transparency; Accountability
Security and Privacy	No. 8, 9, 10	Integrity and confidentiality
Moral and Ethical Observations	No. 4, 5, 8, 9, 10	Purpose limitation; data minimization; storage limitation
Public Understanding	No. 8	Lawfulness, fairness, and transparency

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

The study makes several key contributions to the field of value-driven GAI development:

- Uncovering implicit values: Developing methods to identify and analyze the assumptions encoded in LLM-generated outputs.
- Quantifying value differences: Introducing a systematic approach to measure moral and ethical misalignments using embedding techniques and similarity metrics.
- Iterative realignment: Proposing a novel topic-weighted fine-tuning process to address specific moral and ethical concerns while maintaining model utilities.
- Scalability and adaptability: Providing a modular framework that can be adapted to various ethical standards and societal contexts.

The remainder of this paper is organized as follows: Section 2 reviews related work in AI value alignment and ethical frameworks. Section 3 outlines the proposed methodology, including the steps for uncovering implicit values, quantifying value differences, explaining topic-specific value differences, and implementing iterative value realignment. Section 4 presents the experimental setup and results, highlighting key insights from the case study. Section 5 discusses the ethical implications, scalability, and limitations of the proposed methodology. Finally, Section 6 concludes with a summary of findings and directions for future research.

2. Related work

Value alignment, a critical area in AI ethics, seeks to ensure that GAI systems operate in accordance with human values. One of the most prominent methods in this domain is RLHF, which incorporates human-generated evaluations into the training process of LLMs (Ouyang et al., 2022). RLHF enables models to learn and respond to nuanced user expectations, aligning outputs with desired preferences (Ziegler et al., 2019). While RLHF has achieved strong behavioral alignment in models such as ChatGPT and Claude, it is resource-intensive and depends heavily on curated annotator pools whose feedback may not represent diverse moral or cultural perspectives (Casper et al., 2023; Wang et al., 2024). Recent advances such as Constitutional AI mitigate these challenges by replacing part of the human feedback loop with model-based critique and self-improvement guided by explicit written principles (Bai et al., 2022). However, these methods focus on general alignment, making it difficult to address specific topic-based misalignments within an ethical framework.

Other approaches to value alignment involve fine-tuning LLMs using domain-specific or value-targeted datasets. For example, the ETHICS dataset was developed to evaluate LLM's capability to understand basic shared human values, covering justice, well-being, duties, virtues, and commonsense morality (Hendrycks et al., 2020). Moreover, the Process for Adapting Language Models to Society (PALMS) framework fine-tunes GPT-3 on a curated set of questions and answers designed to reflect social and ethical values (Solaiman and Dennison, 2021). PALMS adopts an iterative process in which model outputs are evaluated for ethical adherence and additional training examples are incorporated based on observed shortcomings, resulting in progressively improved alignment without degrading core language capabilities. Similarly, principle-driven self-alignment techniques leverage synthetic data guided by human-written principles, reducing the need for extensive human input during training (Sun et al., 2024). These methods allow targeted adjustments to LLM outputs but often rely on pre-specified datasets, which may not capture evolving ethical requirements or address localized context-specific ethical issues (Weidinger et al., 2021).

Recent research has explored the use of high-dimensional embeddings to model human values in AI outputs. Specifically, text embeddings have become widely used to analyze the latent representations of underlying language models. They are numerical encodings of textual information, capturing semantic and contextual meaning in a high-dimensional embedding (vector) space. Early embedding techniques, such as Word2Vec (Mikolov, 2013) and GloVe (Pennington et al., 2014), laid the groundwork for calculating text-based vectors with semantic meaning. More recent advancements in transformer-based

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

deep learning models, such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2019), have enabled more sophisticated text embeddings by considering sentence contexts and capturing semantic relationships (Li et al., 2020). These text embeddings have been widely used to analyze the ethical dimensions of LLMs (Jentzsch et al., 2019; Bender et al., 2021; Schramowski et al., 2022; Salinas et al., 2023; Preniqi et al., 2024). These high-dimensional representations analyze value-related responses generated by LLMs, enabling comparisons across different languages and cultures (Cahyawijaya et al., 2024). By visualizing value similarities and differences using dimension reduction techniques, such as Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), these methods provide insights into the implicit assumptions encoded in GAI systems. However, while embeddings are effective for identifying value disparities, they do not directly provide mechanisms for correcting value misalignments.

Aligning GAI systems with specific ethical benchmarks, such as religious, philosophical, or legal frameworks, has also gained traction. Reference value sets, such as the Ten Commandments or the Universal Declaration of Human Rights, offer clear moral and ethical standards against which GAI outputs can be evaluated. These benchmarks are especially useful for realigning GAI systems in culturally sensitive or morally contentious areas, providing clear criteria for evaluating alignment (Gabriel, 2020). However, challenges remain in adapting these frameworks to computational processes and balancing competing priorities among different topic-specific ethical principles. Complementing these framework-based approaches, recent AI governance-oriented work on advanced AI assistants conceptualizes value alignment as a tetradic relationship between the AI system, the user, the developer, and society, identifying distinct varieties of misalignment and underscoring the need to balance these competing claims at the sociotechnical level (Gabriel et al., 2024).

Recent debates in AI ethics have highlighted tensions between the rhetoric and practice of responsible AI. Critics warn against "ethics washing," where organizations adopt ethical principles or advisory boards primarily for reputational purposes without substantive changes to design or governance (Wagner, 2019). Conversely, "ethics bashing" reflects the growing cynicism toward ethical initiatives, dismissing them as ineffectual or obstructive to innovation (Bietti, 2021). These debates underscore the need for value-driven frameworks that move beyond abstract principles toward measurable and topic-specific mechanisms of ethical alignment.

Amid recent advances and ongoing debates in AI ethics, value alignment methods still face key limitations. Methods such as RLHF and Constitutional AI improve general human preference alignment but remain resource-intensive and lack mechanisms to quantify or correct topic-specific ethical gaps (Bai et al., 2022; Casper et al., 2023; Wang et al., 2024). Frameworks like PALMS (Solaiman and Dennison, 2021) introduce iterative refinement but depend on manual data curation and lack a structured, topic-weighted process to prioritize areas of misalignment. Building on these foundational works, this study introduces a scalable and systematic value-driven framework for embedding social good values into LLMs. By (1) uncovering implicit values, (2) quantifying value differences, (3) analyzing topic-specific differences, and (4) implementing iterative value realignment, the proposed framework provides a novel, data-driven approach to ethical value alignment in LLMs. It is adaptable to diverse ethical standards, enhancing cross-domain applicability and enabling developers and policymakers to foster responsible AI innovation.

3. Methodology

This section describes the proposed value-driven framework for embedding social good values into LLMs. The methodology is structured around four key steps: (1) uncovering implicit values, (2) quantifying value differences, (3) analyzing topic-specific value differences, and (4) implementing an iterative value realignment process. These four steps work together to systematically realign LLM outputs with a desirable set of social good values while maintaining model performance and flexibility. Figure 1 summarizes our novel value-driven LLM methodology for social good value embedding. The proposed framework is evaluated through computational experiments using multiple LLM models and

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

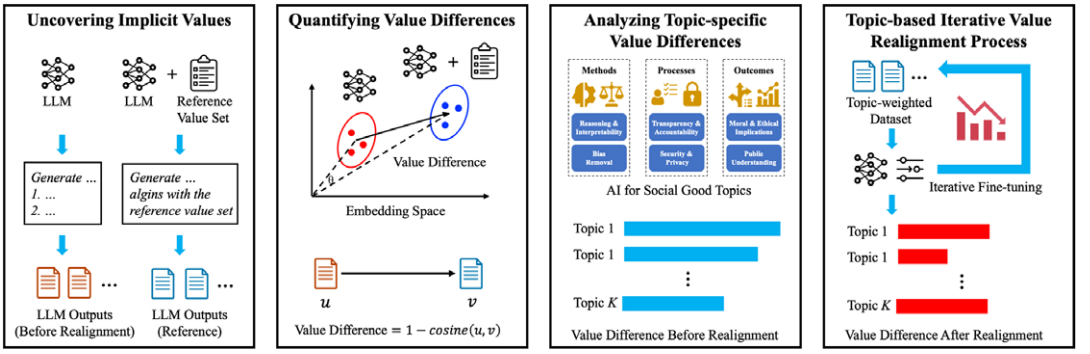


Figure 1. *A value-driven LLM framework to embed social good values into LLMs.*

reference value sets, where its effectiveness is assessed based on the framework's ability to identify topic-specific value misalignments and reduce value differences through iterative value realignment.

3.1. Uncovering implicit values

LLMs are trained on extensive datasets that encapsulate the diversity of human language, culture, and knowledge. However, this diversity also embeds implicit biases, assumptions, and value systems that may not align with universally accepted ethical standards or principles of social good. The process of uncovering these implicit values is essential for identifying misalignments and serves as the foundation for subsequent steps in the proposed framework. This section elaborates on the methodology for identifying the implicit values underlying an LLM model, focusing on topic selection, reference value set selection, and prompt design.

To systematically uncover the implicit values embedded in LLMs, it is necessary to define the domains where ethical considerations are critical. Based on prior research and existing frameworks for AI ethics, this study selected the six core topics representing key pillars of AIfSG (Li et al., 2021, 2024), including (1) Reasoning and Interpretability, (2) Bias Removal, (3) Transparency and Accountability, (4) Security and Privacy, (5) Moral and Ethical Observations, and (6) Public Understanding.

A core component of uncovering implicit values is selecting an appropriate reference value set that represents a desirable set of moral and ethical standards. Using a fixed, absolute value set allows systematic comparison of LLM outputs against pre-defined benchmarks, offering a consistent and interpretable method for assessing value alignment. The proposed value-driven framework assumes the reference value set to be an input defined by relevant stakeholders, such as domain experts, policymakers, or affected communities. To demonstrate the framework's versatility across different ethical paradigms, we selected two illustrative reference sets: the Ten Commandments, representing a concise and historically influential moral code often used in philosophical discussions of universal ethics (Irwin, 2011), and the GDPR, reflecting a contemporary, legally grounded framework that operationalizes core principles such as fairness, accountability, and respect for individual rights (Kasirzadeh and Clifford, 2021). Together, these two sets capture both moral-philosophical and policy-based dimensions of value alignment, showcasing the framework's adaptability to diverse normative contexts.

The Ten Commandments, inherently rooted in religious and moral traditions, offer a concise and universally recognizable foundation widely accepted by a considerable portion of the global population (Green, 2000; Friedman, 2025). This study adopts selected Commandments as a reference value set, with each one mapped to the corresponding AIfSG topics (see Table 1). The correspondences between the Ten Commandments and contemporary ethical constructs (e.g., privacy, fairness, accountability) are illustrative and heuristic rather than literal. They are used to demonstrate how the proposed framework can flexibly translate abstract moral ideas into policy-relevant domains for the purpose of quantifying and comparing value differences before and after realignment. Specifically, Commandment 9 ("You shall not

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

bear false witness against your neighbor") underpins reasoning, interpretability, and bias removal by emphasizing truthfulness and fairness, key principles for reducing algorithmic bias and ensuring explainable AI outputs (Schlender and Spanakis, 2020; Voorhees, 2021). Commandments 9 and 3 ("You shall not take the name of the Lord your God in vain") relate to transparency and accountability by promoting honesty in AI disclosures and discouraging misuse of authority (Voorhees, 2021; Umbrello, 2025). Security and privacy concerns align with Commandments 8 ("You shall not steal"), 9, and 10 ("You shall not covet"), which condemn theft, deception, and covetous data practices, grounding AI governance in respect for data ownership and informed consent (Thacker, 2020; Sangwa and Mutabazi, 2025). Broader ethical observance draws from Commandments 4 ("Remember the Sabbath day, to keep it holy"), 5 ("Honor your father and your mother"), 8, 9, and 10 means designing AI systems that promote human flourishing through respect for rest, family, and community and that refrain from cheating, deceiving, or exploiting others (DeLashmutt, 2025; Sangwa and Mutabazi, 2025). Finally, public understanding is supported by the principle in Commandment 8, ensuring an informed public guards against exploitation and affirms the dignity and rights of all, which is consistent with the spirit of the commandment against theft (Heritage History, 2025).

The GDPR establishes a legal-ethical framework for responsible data processing and governance. Building on Article 5 of the GDPR, this study adopts its seven core principles as a reference value set, including (1) lawfulness, fairness, and transparency; (2) purpose limitation; (3) data minimization; (4) accuracy; (5) storage limitation; (6) integrity and confidentiality; and (7) accountability. Each principle is mapped to the corresponding AIfSG topics (see Table 1). Specifically, lawfulness, fairness, and transparency underpin reasoning, interpretability, and public understanding through explainable and trustworthy AI design (Roundtree, 2023; Cheong, 2024; Singhal et al., 2024). Accuracy and fairness guide bias detection and correction (Alvarez et al., 2024). Transparency and accountability ensure traceability and auditability in decision processes (Wieringa, 2020). Integrity and confidentiality protect data security and privacy (Lemieux and Werner, 2024). Purpose limitation, data minimization, and storage limitation govern moral and ethical data use and retention (Mühlhoff and Ruschemeier, 2025).

While the Ten Commandments and the GDPR serve as a starting point, the proposed value-driven framework can be adapted to incorporate other moral and ethical standards, such as the Universal Declaration of Human Rights (UDHR) (United Nations, 1948), or professional and policy guidelines, such as the IEEE Ethically Aligned Design Principles (Shahriari and Shahriari, 2017), the OECD AI Principles (OECD, 2024), and the EU AI Act (European Commission, 2024). This adaptability ensures that the proposed value-driven framework can accommodate diverse ethical, cultural, and regulatory contexts.

To extract implicit values and ensure consistent value realignment, a structured prompting strategy was developed to systematically engage the LLM across all six AIfSG topics. The prompting process comprised three stages: (1) Before Value Realignment: generating original principles to ensure a selected AIfSG topic, capturing the LLM's implicit value assumptions before realignment, (2) Aligning with the Reference Value Set: generating equivalent principles aligned with the selected reference value set, and (3) After Value Realignment: generating equivalent principles that reflect ethical awareness without explicit reference to any reference value set. When aligning the original principles to the reference value set, the prompt design explicitly incorporated considerations of translation and interpretive nuances, particularly for historically derived frameworks such as the Ten Commandments, to mitigate semantic ambiguity and ensure contextually faithful alignment. Detailed prompt templates are provided in Supplementary Table S1 in the Supplementary Information.

3.2. Quantifying value difference

Once the implicit values embedded in LLMs have been uncovered, the next step involves quantifying the extent of misalignment between these values and the reference value set. This quantification provides a measurable basis for evaluating the LLM's ethical consistency and identifying specific areas requiring

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

further intervention. The proposed value-driven framework employs high-dimensional embeddings, similarity metrics, and visualization techniques to capture and analyze value differences systematically.

To enable quantitative analysis, both the LLM-generated outputs before realignment and the LLM-generated outputs aligned with the reference value set were transformed into high-dimensional vector representations (i.e., embeddings). Value embeddings provide a dense numerical representation of semantic meaning, enabling comparisons across diverse textual inputs. A pre-trained sentence embedding model was used to generate embeddings for each text sample. This embedding model encodes the linguistic and semantic content of a sentence into a fixed multi-dimensional vector space, facilitating quantitative comparisons. For each topic, the LLM-generated responses and their corresponding responses aligning with the reference value set were converted into value embeddings. The same embedding model was used across all topics to ensure consistency and comparability.

The value alignment between LLM outputs and the reference value set was quantified using cosine similarity ranging from -1 to 1, a metric widely used to compare the similarity of vectors in high-dimensional spaces. The value difference was calculated as cosine distance, that is, one minus the cosine similarity (see Equation (1)).

$$\text{Value Difference} = 1 - cosine(u, v) \tag{1}$$

where u represents the embedding of one LLM-generated response before realignment and v represents the embedding of the corresponding LLM-generated response aligned with the reference value set. As a result, the value difference ranges from 0 to 2. Alignment refers to cases where the value difference approaches zero, indicating strong correspondence with the reference value set, while misalignment denotes higher value differences that signify divergence from the reference value set. For example, a value difference of 0.1 indicates that the LLM output closely reflects the reference value. In contrast, a value difference of 0.9 highlights a substantial divergence, warranting realignment. For each topic, the average cosine distance across all sample pairs (LLM-generated outputs before realignment and LLM-generated outputs aligned with the reference value set) was calculated, providing an overall value misalignment score for each topic and enabling comparative analysis across different ethical dimensions.

Moreover, to facilitate the interpretation of high-dimensional value embeddings and their differences, dimensionality reduction analysis was performed to project value embeddings into a two-dimensional space. Specifically, the mean value embeddings were calculated for each topic. Multidimensional Scaling (MDS) (Hout et al., 2013) was employed to reduce the dimensionality of the mean value embeddings, preserving the pairwise distance relationships between them in a lower-dimensional space. This approach allows the visualization to better reflect the value differences in the value embedding space, providing an interpretable representation of the relationship between the LLM outputs and the reference values across topics.

3.3. Analyzing topic-specific value difference

Once value difference has been quantified and visualized, the proposed value-driven framework identifies and prioritizes topics based on the difference in values. Topics were ranked by the average value distance of their respective outputs. Topics with the smallest and largest deviations from the reference value set were identified, highlighting the topics with the least and most significant ethical gaps. Topics with larger average value differences are deemed to have more significant misalignments and are prioritized for further intervention. Conversely, topics with smaller average differences are noted for their closer alignment, requiring minimal adjustment and further intervention.

3.4. Topic-based iterative value realignment

The iterative realignment process serves as the central mechanism for addressing topic-specific value disparities identified during the previous step. This method relies on a systematic, iterative approach to

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

refine the behavior of the LLM, ensuring that its outputs align closely with pre-defined reference values across diverse topics. By iteratively making targeted adjustment, the model can eventually achieve a balance between adhering to moral and ethical standards and preserving its wide-ranging utilities. The iterative framework is particularly effective in handling complex, multi-dimensional value difference, as it allows for incremental, data-driven improvements over successive iterations. The process of iterative realignment systematically reduces value differences through the following stages:

(1) ***Dataset Construction:*** For each AIfSG topic, pairs of the original principle (LLM-generated output before realignment) and the reference-aligned principle (LLM-generated output aligned with the reference value set) were first obtained from the value-uncovering stage (see Section 3.1). These pairs were combined to form a dataset in which each input (user message) was a prompt instructing the LLM to generate an equivalent principle so that it reflects ethical awareness without explicitly referencing any ethical framework, and each output (assistant message) was the corresponding principle aligned with the reference value set. A topic label was retained for each pair to enable topic-specific weighting during fine-tuning. The dataset was then divided into training, validation, and test subsets, stratified by topic, to ensure balanced representation across all topics. This dataset enables the model to learn how to transform a generic principle into one that aligns with a specific reference value set. By mapping each principle to its corresponding reference-aligned counterpart, the LLM acquires an implicit understanding of how its internal value representations should be adjusted to align with the reference value set.

(2) ***Topic Prioritization Based on Value Difference:*** To identify and prioritize topics requiring stronger ethical realignment, the mean value difference for each topic was aggregated and normalized to the range [0, 1], yielding a topic-specific loss weight. These weights were used to scale the training loss during fine-tuning so that the total loss for each batch was computed as the weighted mean of individual cross-entropy losses across examples. Topics with larger mean value differences, therefore, contributed more strongly to loss optimization, directing learning toward areas requiring greater value realignment while maintaining balanced representation across all topics.

(3) ***LLM Fine-tuning:*** Fine-tuning was performed using a topic-weighted supervised fine-tuning approach implemented with parameter-efficient adaptation (LoRA) (Hu et al., 2022). Each input–output pair was encoded as a conversational message, where the user message contained the prompt to generate an equivalent principle without specifying the reference value set, and the assistant message contained the reference-aligned principle. Training optimized the token-level cross-entropy loss, the standard objective for causal language modeling, which measures how well the predicted token distribution matches the ground-truth aligned output. The loss for each example was multiplied by its topic-specific weight, ensuring that topics requiring greater value realignment exerted proportionally stronger influence on parameter updates. This formulation enables fine-grained, topic-based optimization of value realignment while seamlessly integrating with LoRA-based fine-tuning for efficient and scalable adaptation.

(4) ***Evaluation and Feedback Loop:*** After each fine-tuning iteration, the updated LLM generated realigned principles for the validation set. The value differences between these realigned principles and their corresponding reference-aligned principles were recalculated to assess post-training value differences. New topic-wise mean value differences were computed and normalized to update the topic-specific loss weights for the next iteration. This feedback mechanism created an adaptive loop in which topics requiring further value realignment were automatically prioritized in subsequent training cycles. The mean value difference across all topics served as the primary metric for monitoring convergence and evaluating iterative improvements in the value realignment process.

(5) ***Convergence and Termination Criteria:*** The iterative realignment process continued until one of two stopping conditions was met: (a) the number of iterations reached a predefined threshold or (b) the relative improvement in mean value difference between consecutive iterations fell below a

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

predefined threshold, indicating diminishing returns. The checkpoint yielding the lowest mean value difference was selected as the best model and evaluated on the held-out test set to confirm the generalization of value realignment.

4. Experimental setup and results

4.1. Experimental setup

In this study, the experiments were structured in two main phases: value uncovering and value realignment, corresponding to the diagnostic and corrective stages of the proposed value-driven framework.

In the value uncovering phase, the implicit values encoded within LLM outputs were identified and compared against reference value sets. Two open-source base models were used for text generation, including Meta AI's Llama 3.2 [3B parameters] (Touvron et al., 2023) and Google DeepMind's Gemma 2 [2B parameters] (Mesnard et al., 2024). Two reference value sets were examined, including the Ten Commandments and the GDPR. For each of the six AIfSG topics, 100 principles were generated to capture the model's implicit value assumptions and their corresponding reference-aligned counterparts. Both the original and reference-aligned principles were embedded using a sentence transformer model (all-mpnet-base-v2) (Reimers and Gurevych, 2019). Cosine distance between embeddings provided a quantitative measure of value difference, which served as the basis for topic prioritization in the subsequent fine-tuning phase.

In the value realignment phase, the base LLMs were iteratively fine-tuned to minimize topic-specific value differences. The value-realignment dataset was divided into 60% training, 20% validation, and 20% testing subsets, stratified by topic for balanced representation. Fine-tuning employed LoRA-based parameter-efficient adaptation with rank = 16, $\alpha = 32$, and dropout = 0.05, applied to the attention and projection layers of each model. Training used a batch size of 4, a learning rate of 1×10^{-5}, the AdamW optimizer with weight decay = 0, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, and cosine learning rate scheduling for 5 epochs per iteration.

Each value realignment experiment was limited to five fine-tuning iterations, with early stopping triggered when relative improvement in mean value difference on the validation set fell below 1%. After each iteration, topic-specific loss weights were recalculated from the latest validation value differences, adaptively emphasizing topics requiring greater value realignment. The best checkpoint, defined by the lowest mean validation value difference, was then evaluated on the held-out test set to verify the generalization of value realignment.

4.2. Results

The results have demonstrated the efficacy of the proposed value-driven framework for uncovering and realigning the implicit values of LLMs. This section focuses on the results of quantifying value differences, analyzing misaligned topics, and evaluating the impact of iterative realignment.

Figures 2–5 present the mean value differences across six AIfSG topics before and after realignment for two open-source LLMs: Llama 3.2 [3B] and Gemma 2 [2B], under two reference value sets: the Ten Commandments and the GDPR. These figures can serve as diagnostic tools for assessing ethical consistency, pinpointing areas where misalignment persists, and supporting the monitoring, auditing, and governance of AI systems. It should be noted that these analyses provide a structured view of how LLMs express normative principles but not operational performance, that is, how these principles manifest in concrete decisions in applied contexts. For example, Bias Removal does not measure whether the model statistically eliminates bias in its downstream outputs; rather, it evaluates how the model frames fairness and inclusion as ethical priorities within its responses. From a policy perspective, these topic-level insights reveal where an AI system's ethical framing diverges from socially endorsed standards, helping regulators and auditors target human oversight and prioritize governance interventions in high-risk domains.

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

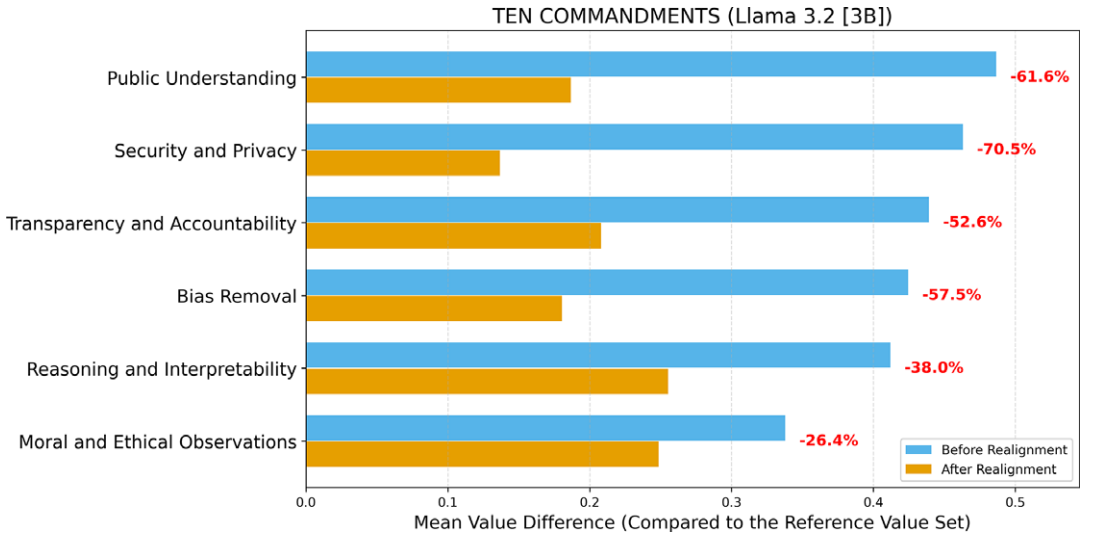


Figure 2. *Value difference across topics before and after realignment (LLM: Llama 3.2 [3B]; Reference Value Set: Ten Commandments).*

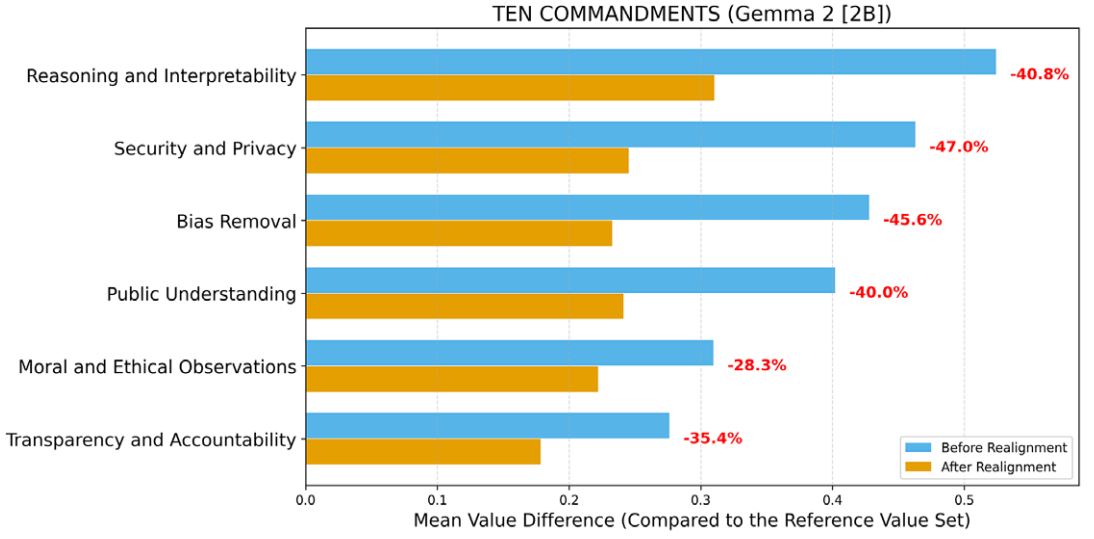


Figure 3. *Value difference across topics before and after realignment (LLM: Gemma 2 [2B]; Reference Value Set: Ten Commandments).*

Before value realignment, both LLMs exhibited notable value differences across topics, with the largest discrepancies appearing in Public Understanding (Llama 3.2; Ten Commandments), Reasoning and Interpretability (Gemma 2; Ten Commandments), Moral and Ethical Observations (Llama 3.2; GDPR), and Bias Removal (Gemma 2; GDPR). These variations indicate that the LLMs initially diverged most in domains emphasizing public understanding, reasoning, interpretability, fairness, and broader moral and ethical observations, reflecting topic-dependent sensitivities to ethical reasoning under different reference frameworks. In contrast, smaller value differences were observed in topics such as

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

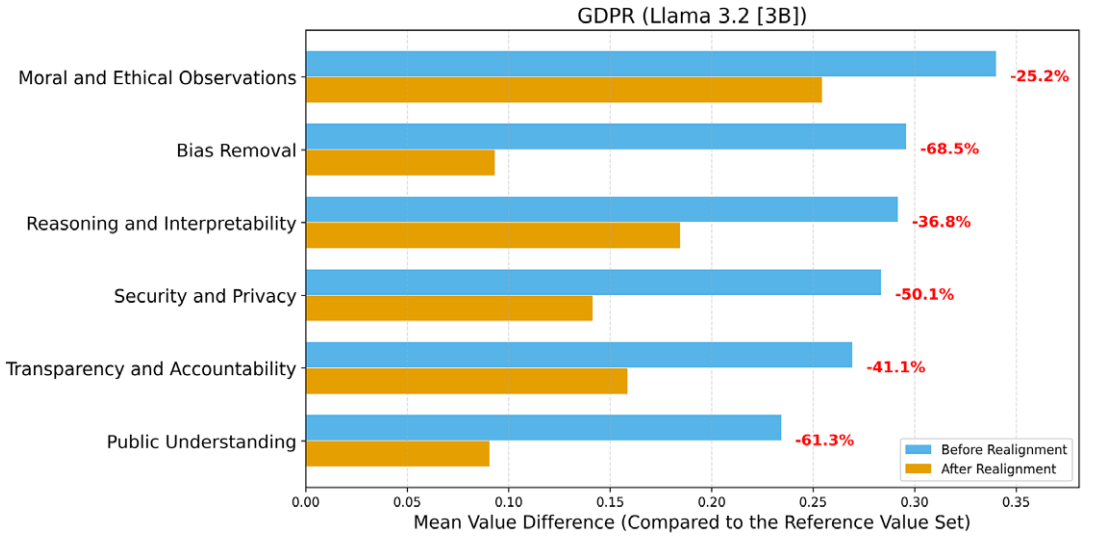


Figure 4. *Value difference across topics before and after realignment (LLM: Llama 3.2 [3B]; Reference Value Set: GDPR).*

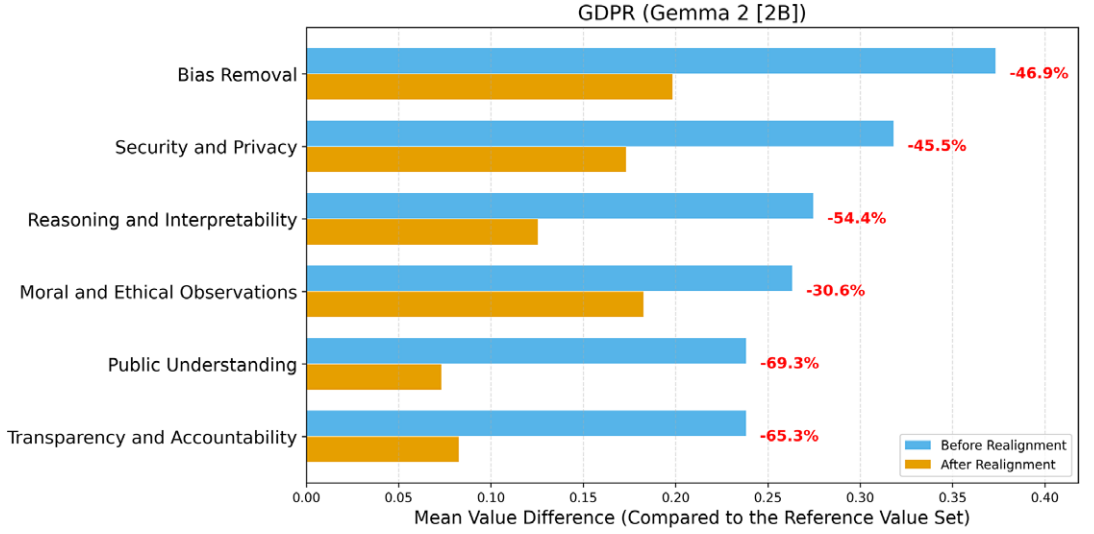


Figure 5. *Value difference across topics before and after realignment (LLM: Gemma 2 [2B]; Reference Value Set: GDPR).*

Security and Privacy and Transparency and Accountability, suggesting a relatively stronger baseline alignment with principles of transparency, accountability, and data protection. Overall, the pre-realignment results reveal that while both LLMs exhibited partial ethical consistency across topics, substantial misalignments persisted in several key domains, underscoring the necessity of targeted, topic-weighted fine-tuning to achieve balanced and comprehensive value realignment.

Supplementary Figures S1 and S2 in the Supplementary Information illustrate the effectiveness of the iterative fine-tuning process in progressively reducing the mean value difference between LLM outputs and the selected reference value sets. These figures visualize the dynamics of realignment, providing a quantitative basis for assessing when iterative interventions have achieved satisfactory convergence and

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

when additional oversight, evaluation, or retraining may be warranted. As shown in Supplementary Figure S1, when the Ten Commandments were used as the reference value set, both Llama 3.2 and Gemma 2 models exhibited a sharp initial decline in value difference after the first iteration, followed by gradual convergence over subsequent iterations. The reduction was particularly pronounced within the first two iterations, suggesting that most realignment occurs early in the process. Likewise, Supplementary Figure S2 shows consistent patterns when the GDPR principles were used as the reference value set. Both models demonstrated substantial decreases in mean value difference, reaching stable convergence within three to five fine-tuning iterations. Collectively, these findings confirm that the proposed topic-weighted iterative fine-tuning approach can effectively minimize value differences between LLM outputs and the chosen reference value sets, demonstrating robustness and generalizability across different LLMs and value systems.

When the Ten Commandments were selected as the reference value set (Figures 2 and 3), both LLMs exhibited consistent convergence toward the reference values after realignment. For Llama 3.2, the largest improvements were observed in Security and Privacy (−70.5%) and Public Understanding (−61.6%), followed by Bias Removal (−57.5%) and Transparency and Accountability (−52.6%). Moderate yet meaningful reductions occurred in Reasoning and Interpretability (−38.0%) and Moral and Ethical Observations (−26.4%). The sharper declines among privacy- and transparency-related topics suggest that realignment most effectively corrected ethical gaps related to data protection and truthful representation. For Gemma 2, reductions were similarly broad, though slightly less pronounced in magnitude, ranging from −47.0% (Security and Privacy) to −28.3% (Moral and Ethical Observations). Topics emphasizing reasoning and fairness, Reasoning and Interpretability (−40.8%), Bias Removal (−45.6%), and Public Understanding (−40.0%) showed steady improvement, indicating generalizability of the realignment process to smaller-parameter models.

When the GDPR was selected as the reference value set (Figures 4 and 5), value differences again decreased markedly across all topics after realignment. Llama 3.2 achieved the largest relative improvement in Bias Removal (−68.5%) and Public Understanding (−61.3%), with moderate reductions in Security and Privacy (−50.1%), Transparency and Accountability (−41.1%), and Reasoning and Interpretability (−36.8%). Gemma 2 displayed comparable performance, with reductions exceeding −45% in Bias Removal and Security and Privacy, and substantial alignment gains in Transparency and Accountability (−65.3%) and Public Understanding (−69.3%).

Taken together, these results demonstrate that (1) the proposed topic-weighted iterative fine-tuning process effectively minimizes value differences across diverse ethical frameworks and (2) the observed improvements are consistent across model scales and architectures. The relatively larger reductions in Security and Privacy, Transparency and Accountability, and Bias Removal indicate that topics directly tied to fairness, honesty, and responsible data handling benefit most from targeted value realignment. In contrast, smaller yet consistent reductions in Moral and Ethical Observations suggest that abstract moral reasoning may require additional iterations or richer contextual grounding for full value realignment. Overall, the convergence patterns observed across all figures confirm that the proposed value-driven framework offers a scalable and generalizable pathway for embedding moral and ethical values into LLMs.

Moreover, Figures 6–9 visualize the two-dimensional value embedding spaces before and after realignment for both Llama 3.2 and Gemma 2, using the Ten Commandments (Figures 6 and 7) and GDPR (Figures 8 and 9) as the reference value sets. MDS (Hout et al., 2013) was used to project the mean value embeddings of each AIfSG topic into a two-dimensional space. MDS was selected because it preserves the pairwise distances among embeddings more faithfully than linear reduction methods such as principal component analysis (PCA), enabling clearer visualization of how the relative proximities between model and reference values evolve through the realignment process. Before realignment, large separations between LLM outputs (circles) and reference values (squares) indicate substantial misalignments across topics. After realignment, the realigned outputs (diamonds) converge markedly toward their corresponding references, confirming reduced value differences across all topics. These visualizations provide an intuitive geometric validation of the quantitative results: the topic-specific embeddings move

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

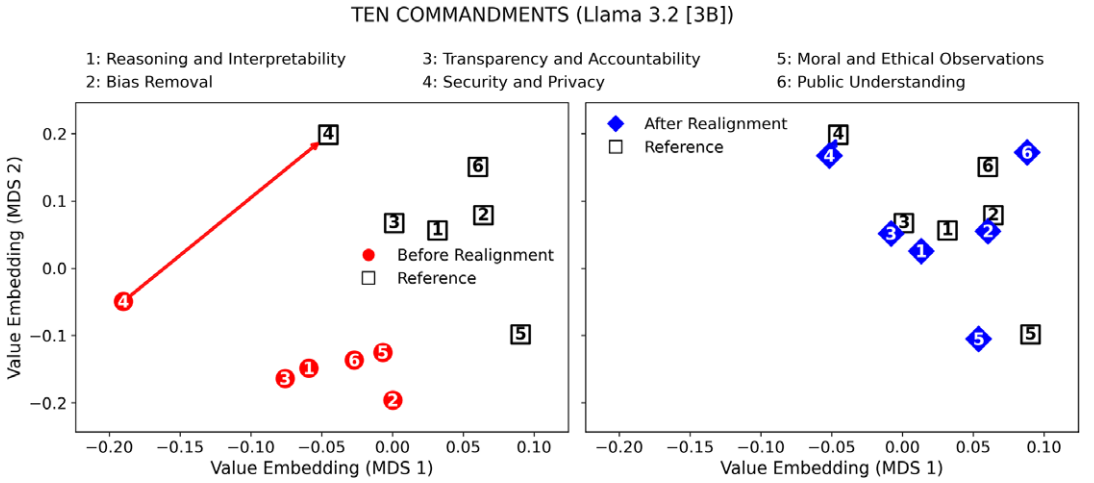


Figure 6. *Value difference in the value-embedding space before and after realignment (LLM: Llama 3.2 [3B]; Reference Value Set: Ten Commandments).*

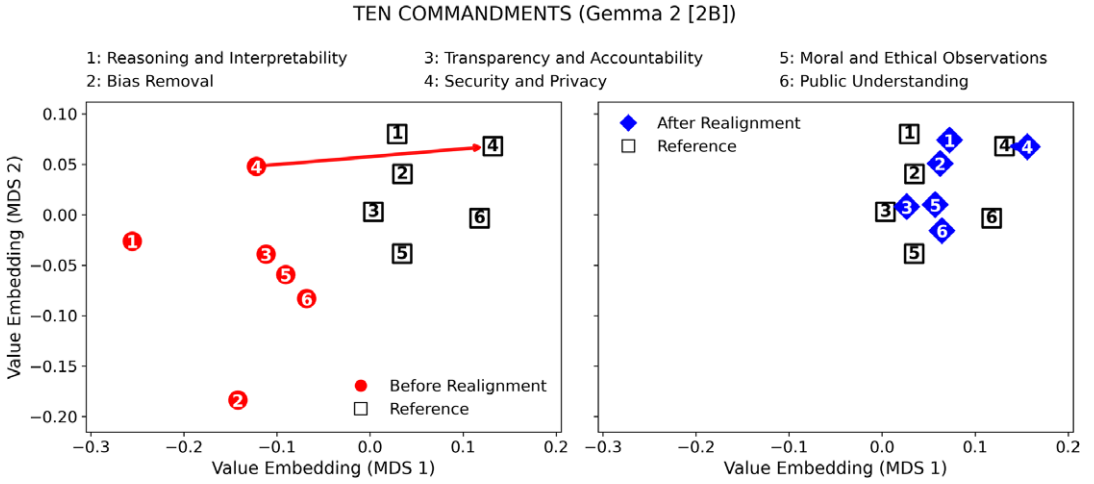


Figure 7. *Value difference in the value-embedding space before and after realignment (LLM: Gemma 2 [2B]; Reference Value Set: Ten Commandments).*

significantly closer to their reference values, revealing a clear reduction in pairwise distances and enhanced alignment across all six AIfSG topics.

Further, to complement the quantitative value-difference analysis, we performed a qualitative validation to examine whether the realigned LLM outputs more closely reflect the intended reference value sets. For each model and each reference framework, we identified the LLM-generated outputs that exhibited the largest reductions in value difference before and after realignment (see Supplementary Tables S2–S5 in the Supplementary Information). Manual inspection of these samples indicated that large changes in value difference were often associated with substantive shifts in ethical emphasis. For example, under the Ten Commandments, in Public Understanding (Supplementary Table S2), broad procedural statements such as "AI should be accountable and responsive to societal needs" were rephrased as explicit ethical directives like "You shall not manipulate public discourse for personal gain or to obscure the truth." Similarly, in Reasoning and Interpretability (Supplementary Table S3), the outputs moved from technical, process-oriented descriptions toward statements underscoring honesty and

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

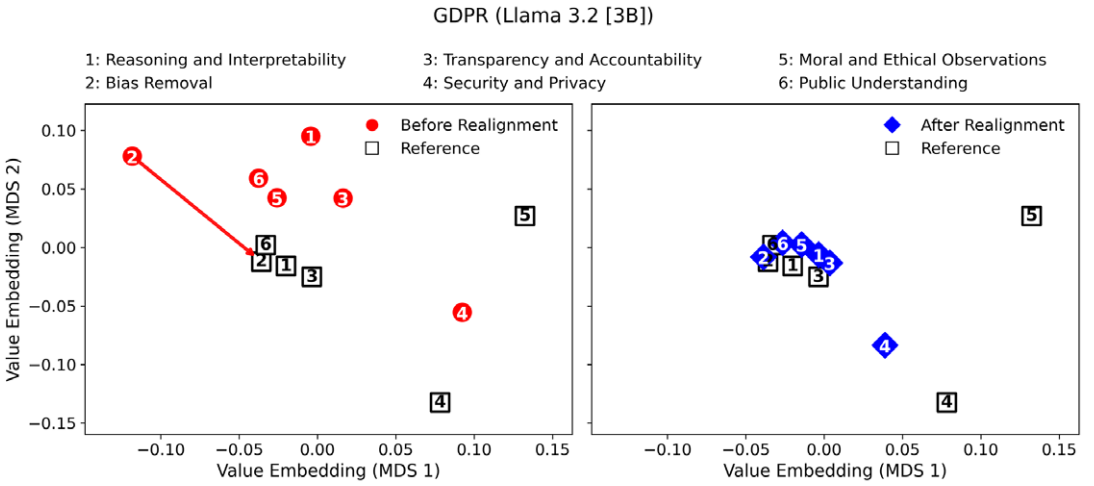


Figure 8. *Value difference in the value-embedding space before and after realignment (LLM: Llama 3.2 [3B]; Reference Value Set: GDPR).*

integrity. Under the GDPR, the outputs showed closer value alignment with principles of fairness, accountability, and transparency, reframing general design recommendations into formulations that emphasize lawful and trustworthy AI behavior (Supplementary Tables S4 and S5). Taken together, these qualitative observations suggest that the largest quantitative improvements in cosine-based value difference are typically accompanied by meaningful shifts in ethical orientation. Nevertheless, expert evaluations will be necessary to systematically verify whether these observed changes align with human moral judgments across diverse contexts.

5. Discussion

The results of the study demonstrate the efficacy of the proposed value-driven framework in identifying and realigning the implicit values of LLMs to ensure that they embed and align more closely with the socially desirable/social good values. By quantifying value difference and analyzing misaligned topics, the findings provide significant insight into how iterative fine-tuning can reduce discrepancies between a GAI system's behavior and a reference value set. The ethical implications of these results are profound, as they suggest that LLMs, when properly realigned, can better reflect a selected social good reference value set, which is critical for promoting AI systems that behave in a manner consistent with ethical principles.

One of the key findings is the large initial value difference observed before value realignment (Figures 2–5). These results have demonstrated that both LLMs (Llama 3.2 and Gemma 2) misalign with the reference values to a significant extent before realignment, particularly in topics such as Public Understanding, Reasoning and Interpretability, Moral and Ethical Observations, and Bias Removal. Analogous to how An et al. (2014) quantified political polarization on social media using a continuous net partisan skew, our framework operationalizes ethical alignment as a continuous value distance metric, enabling comparable quantitative analysis of ethical divergence in LLM behavior. Following the iterative realignment process, these value differences were consistently reduced across all topics and reference value sets (ranged approximately from 25% to 70%). Overall, these findings demonstrate that the proposed framework effectively reduces the value differences between model outputs and reference values, successfully realigning both moral-philosophical (Ten Commandments) and regulatory (GDPR) benchmarks toward greater ethical coherence and interpretability.

Moreover, the visualization of value embeddings before and after value realignment, as depicted in the MDS plots, has highlighted the effectiveness of the iterative value realignment process (Figures 6–9). The

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

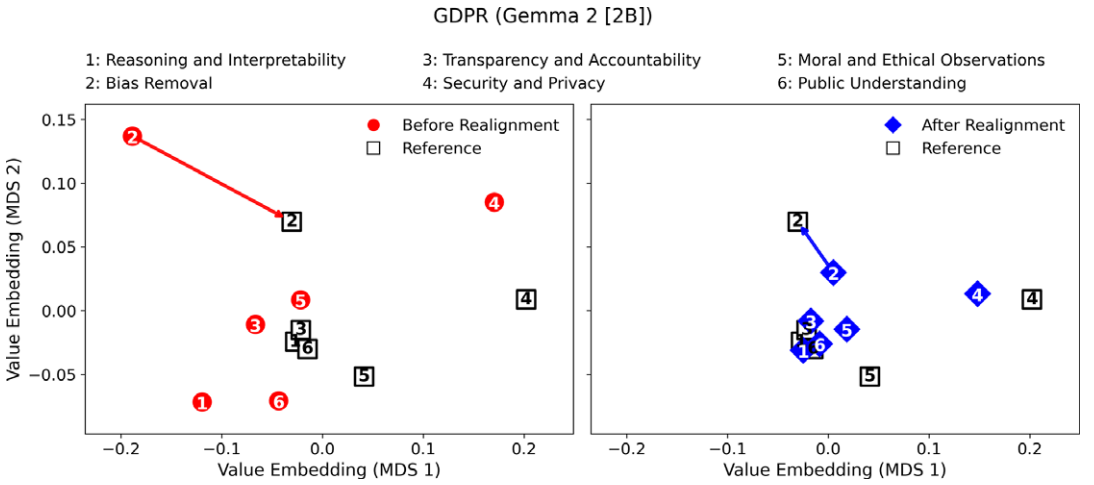


Figure 9. *Value difference in the value-embedding space before and after realignment (LLM: Gemma 2 [2B]; Reference Value Set: GDPR).*

noticeable reduction in value distance between the initial and the outputs from realigned LLMs illustrates a significant shift in the alignment of LLM outputs with the reference values, confirming that iterative fine-tuning is a practical and effective method for enhancing the value alignment of GAI systems. These findings underscore the ethical significance of the value realignment process, as they indicate that through focused weighting adjustment, GAI systems can be more reliably aligned with the moral and ethical standards. This value alignment is particularly critical in the context of LLMs, which carry the potentials to influence public opinions, perpetuate stereotypes, or even inflict harms on the societies when their LLM-generated outputs do not align well with the socially desirable moral and ethical values.

One of the strengths of the proposed value-driven LLM framework is its potential for scalability. The value realignment processes outlined are not limited to a specific subset of LLMs but can be applied more broadly to any LLM systems. As LLMs continue to evolve and proliferate in various domains, ranging from healthcare and finance to education and public services, this framework provides a systematic approach to supporting more socially responsible deployment. The framework's scalability is also supported by its modular nature. The techniques for topic-based value realignment can be tailored to specific use cases, domains, or regulatory environments, allowing for a flexible approach to social good integration. This makes the novel, value-driven LLM framework applicable in both commercial and public sector applications, where different moral and ethical standards and priorities may be at play.

Beyond its technical contributions, the framework also carries direct implications for AI governance and policy implementation. It can operate both as a diagnostic tool and as an intervention tool within regulatory and organizational contexts. As a diagnostic tool, the framework provides quantitative indicators, such as value difference metrics and topic-specific analyses, which enable policymakers, regulators, and independent auditors to identify where model outputs diverge from defined ethical or legal standards. These diagnostics can be integrated into AI-impact assessments, model-card documentation, and compliance audits to prioritize high-risk domains for oversight. As an intervention tool, the topic-weighted realignment process offers a structured mechanism for targeted ethical adjustment and adaptive monitoring. By iteratively fine-tuning models in response to measured misalignments, agencies and developers can demonstrate progress toward normative benchmarks and maintain ongoing accountability as systems evolve. In this way, the framework supports evidence-based, transparent, and adaptive governance of generative AI technologies.

Operationalizing the proposed framework in real-world governance contexts requires clear mechanisms for value definition, updating, and oversight. In regulatory settings, the reference value sets can be co-defined through participatory, multi-stakeholder processes (Sharma, 2024), involving policymakers,

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

ethicists, domain experts, and affected communities to ensure legitimacy and inclusiveness. These value sets should be periodically reviewed and updated as societal norms evolve, with transparent documentation to maintain accountability. Independent auditing bodies or ethics boards can serve as oversight mechanisms to monitor adherence and prevent misuse (Falco et al., 2021). At the same time, embedding explicit value sets risks reinforcing dominant cultural norms or introducing bias if poorly governed (Tao et al., 2024). Therefore, safeguards such as pluralistic consultation, contextual adaptation, and public reporting are essential to avoid cultural imposition and to ensure that value alignment frameworks enhance rather than constrain diversity, fairness, and ethical innovation.

Adoption of such value-centric frameworks can be advanced through a mix of regulatory mandates and policy incentives. AI governance emphasizes post-deployment monitoring and continuous value alignment as essential components of responsible AI oversight (Stein et al., 2024; Pappas and Vassilakopoulou, 2025). This framework can help operationalize this oversight through measurable indicators. Public procurement standards, certification schemes, and funding requirements can further encourage developers and agencies to embed value-driven realignment into their governance processes, creating both soft and hard incentives for responsible AI adoption.

Despite its strengths, our novel value-driven framework carries several limitations. First, one major limitation is the challenge of effectively measuring the impact of value realignment. While methods such as value difference metrics and two-dimensional value embedding space visualization can track shifts in model outputs and behaviors, quantifying and assessing their broader social implications remains complex (Mittelstadt et al., 2016). Accordingly, the framework should be viewed as a complementary diagnostic and intervention tool operating at the model output layer, to be integrated with structural and institutional analyses rather than treated as a stand-alone solution to AI ethics.

Second, while the framework provides a methodology for embedding social good values into GAI systems, it assumes that human values can be clearly defined and distilled into actionable guidelines for GAI systems. However, human values are often complex, fluid, and sometimes contradictory (Jobin et al., 2019; Gabriel, 2020), and reducing them to fixed rules could risk oversimplification. This presents a challenge when attempting to embed complex moral and ethical principles into algorithmic decision-making, where nuances and contradictions must be accounted for. Moreover, the values themselves are often context-dependent and may not universally apply across cultures, geographies, or stakeholders. What is considered ethical or socially beneficial in one culture or region may not hold true in another, and ensuring orthogonality, that is, avoiding overlap or conflict among embedded values, adds an additional layer of complexity. As a result, the framework may struggle to achieve universal alignment when deployed globally. Against this context, reference frameworks such as the Ten Commandments and the GDPR illustrate how the proposed framework can flexibly adapt to different ethical paradigms. This adaptability enables its deployment across sectors and regions, bridging abstract ethical ideals with practical governance requirements for responsible and context-sensitive AI alignment.

Third, the use of cosine distance between LLM outputs and baseline ethical frameworks has inherent limitations. While cosine distance provides a useful quantitative proxy for semantic similarity, it does not guarantee that the generated outputs are meaningfully aligned with human interpretations of the reference values. To address this, future work should incorporate a human-in-the-loop evaluation approach (Raza et al., 2025) to complement embedding-based metrics. Specifically, cosine distance can be combined with human evaluation to ensure meaningful alignment with reference values. In the first stage, cosine distance can be used to identify LLM outputs that are broadly aligned with the reference values, reducing the initial pool of candidate outputs. For example, a threshold can be set to filter out outputs with low alignment, ensuring only relevant outputs proceed to the next stage. In the second stage, human evaluators (e.g., domain experts such as theologians, philosophers, and AI ethicists) can assess these filtered outputs based on their alignment with the reference values on a scale of 1 to 10 (with 10 indicating perfect alignment). Such human evaluation complements automated similarity-based metrics, ensuring that LLMs align meaningfully with the reference values in real-world applications.

Finally, the framework depends on the selection of reference values, which may themselves be influenced by model biases. If the model used to generate or interpret the reference values is biased,

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

the resulting reference may not represent an absolute ethical benchmark but rather a relative reference point. Consequently, reference positions may shift across different models, affecting the measured degree of value difference. Future work will explore methods to identify unchangeable reference points through more robust value-driven models that can provide more consistent benchmarks for evaluating value alignment.

Even though the present model does not yet consistently embody the values of AI for Social Good, iterative fine-tuning provides a pathway to establish it as a benchmarking foundation model. Through iterative fine-tuning, the model can be stabilized as a reference point whose reference value embeddings remain unchanged over time, allowing it to serve as a benchmarking ruler for assessing the alignment of corresponding foundation models with AI for Social Good principles. In this way, it becomes a reliable tester of how closely other models reflect the principles of AI for Social Good.

As our work demonstrates, LLM potentially provides a strong foundation for embedding social good values, provided that moral and ethical challenges, scalability, and methodological limitations are carefully addressed. Our proposed value-driven framework offers a promising approach to facilitating the alignment of GAI technologies with moral and ethical principles, promoting their positive contribution to society. However, challenges remain in terms of measuring the real-world impact of value realignment, balancing the complexity of human/moral values with the computational capabilities of AI systems, maintaining orthogonality for the embedding social good values to minimize redundancy or conflict, and ensuring global applicability across diverse cultural and regulatory contexts. Addressing these challenges will be crucial for the continued development and responsible deployment of GAI systems that genuinely serve the public good.

6. Conclusion

This paper presents a novel value-driven LLM framework for embedding social good values into LLMs. By systematically uncovering, quantifying, and realigning implicit values, the framework addresses critical ethical challenges in GAI development and deployment. Using open-source LLMs and employing the Ten Commandments and the GDPR as reference value sets demonstrated the proposed approach's effectiveness in reducing value disparities across multiple topics. Looking ahead, future research will explore the integration of multi-stakeholder processes to define and update reference value sets, ensuring that value alignment reflects pluralistic ethical perspectives. Further work will also focus on refining fine-tuning strategies, enhancing interpretability and stability, and developing human-in-the-loop evaluation protocols to assess the real-world social implications of value realignment. By advancing these directions, the value-driven framework can support the creation of more transparent, accountable, and context-sensitive LLMs that meaningfully contribute to the public good.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/dap.2026.10061.

Data availability statement. The data and code associated with this article are publicly available at https://doi.org/10.5281/zenodo.17636407 and are cited in the References section (Li et al., 2025).

Acknowledgments. Our deepest gratitude for the unwavering support and trust of our collaborators and friends across different parts of the world in our vision on AIfSG over the last 14 years. Without their dedication, this vision could not have been successfully implemented. Our gratitude goes to the following people in alphabetical order by surname: Denis Alexander, John Bacon-Shone, Martin Beniston, Meir Buber, Carolyn Cain, David Cardwell, Shirley Chan, Vincent Chan, WC Chan, Danny Chan, Kelvin Chan, Patrick Cheng, late KP Cheung, Francis Chik, WS Ching, Lydia Chung, Karen Costa, Lianne Crossette, Jiahuan Cui, Jocelyn Downey, Rajiv Ellalasingham, George Esber, Wilton Fok, Jinqi Fu, Jason Gan, Emily Gleed-Owen, Toby Gleed-Owen, Illana Gozes, Gianni Heung, Kitty Ho, Silvester Ho, Joshua Hodge, Stephen Holland, Anne Horton, Yunhe Hou, James Huang, Emily Huang, Kristin Janz, Rod Jones, Sumer Kaistha, Tushar Kaistha, Ben Kao, Frank Kelly, Christopher Knittel, Marcus Kraft, Ingrid Kwok, Thomas Kwok, Martin Lau, Ian Leslie, Dennis Leung, Natalie Leung, Enjia Liu, Jen Liu, Yuchen Liu, Stephen Low, Susie Lum, Peter Mathieson, Stephen Matthews, Michael Mehling, Patricia Mirrlees, Tingyu Mo, Lena Mok, Michal Morshtorch, Avi Nagar, David Newbery, Tiffany Ngan, Caleb Olasupo, Adeoluwa Oyinlola, Marcin Pawlak, Vincenzo Pecunio, Michael Pollitt, Buhui Qiu, David Reiner, James Rowe, David Rubinsztein, Richella Ryan, Elpha Shea, Ekaterina Solomina, Ratiya Tangmanee, William Therese, Grace Tsang, Stephen Tsang, Denise Wah, Regina Wai, Shivonne Wai, Shanshan Wang, Ben Wilding, Albert Wu, Virginia Yip, and Alexander Yu. We also acknowledge the strong support of faith-based groups in HK, US, and UK.

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

Author contribution. Conceptualization: V.O.K.L., J.C.K.L.; methodology: V.O.K.L., J.C.K.L., Y.H.; data curation: Y.H., J.C.K.L.; visualization: Y.H.; writing—original draft: Y.H., V.O.K.L., J.C.K.L; inputs on the original draft: J.C., L.Y.L.C.; writing—review and editing: Y.H., V.O.K.L., J.C.K.L., J.C.; funding acquisition: V.O.K.L., J.C.K.L. All authors approved the final submitted draft.

Funding statement. This research was supported by the University of Hong Kong Seed Fund for Collaborative Research. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests. None.

AI statement. The authors acknowledge the use of generative AI tools to support the writing process. ChatGPT-4o was used to help with the preparation of the first manuscript draft from our original presentation slides, and ChatGPT-5 was used to support subsequent manuscript revisions. The authors reviewed the content generated and took full responsibility for the content of the submitted manuscript.

References

Alvarez JM, **Colmenarejo AB**, **Elobaid A**, **Fabbrizzi S**, **Fahimi M**, **Ferrara A**, **Ghodsi S**, **Mougan C**, **Papageorgiou I**, **Reyero P**, **Russo M**, **Scott KM**, **State L**, **Zhao X and Ruggieri S** (2024) Policy advice and best practices on bias and fairness in AI. *Ethics and Information Technology 26*(2), 31.

An J, **Quercia D and Crowcroft J** (2014) Partisan sharing: Facebook evidence and societal consequences. In Goel A and Gummadi K (eds.), *Proceedings of the Second ACM Conference on Online Social Networks*. New York, NY: Association for Computing Machinery (ACM).

Bai Y, **Kadavath S**, **Kundu S**, **Askell A**, **Kernion J**, **Jones A**, **Chen A**, **Goldie A**, **Mirhoseini A and McKinnon C** (2022) *Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.*

Bender EM, **Gebru T**, **McMillan-Major A and Shmitchell S** (2021) On the dangers of stochastic parrots: Can language models be too big? In Irani L, Kannan S, Mitchell M and Robinson D (eds.), *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery (ACM).

Bharathi Mohan G, **Prasanna Kumar R**, **Vishal Krishh P**, **Keerthinathan A**, **Lavanya G**, **Meghana MKU**, **Sulthana S and Doss S** (2024) An analysis of large language models: Their impact and potential applications. *Knowledge and Information Systems 66*(9), 5047–5070.

Bietti E (2021) From ethics washing to ethics bashing: A moral philosophy view on tech ethics. *Journal of Social Computing 2*(3), 266–283.

Cahyawijaya S, **Chen D**, **Bang Y**, **Khalatbari L**, **Wilie B**, **Ji Z**, **Ishii E and Fung P** (2024) *High-Dimension Human Value Representation in Large Language Models. arXiv preprint arXiv:2404.07900.*

Casper S, **Davies X**, **Shi C**, **Gilbert TK**, **Scheurer J**, **Rando J**, **Freedman R**, **Korbak T**, **Lindner D and Freire P** (2023) *Open problems and fundamental limitations of reinforcement learning from human feedback*. arXiv preprint arXiv:2307.15217.

Cheong BC (2024) Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics 6*, 1421273.

De Michele W, **Armas Cervantes A and Frermann L** (2025) Automated business process analysis: An LLM-based approach to value assessment. In Pufahl L, Rosenthal K, España S and Nurcan S (eds.), *International Conference on Advanced Information Systems Engineering*. Cham, Switzerland: Springer.

DeLashmutt M (2025) *Faith's Answer to AI: The Sabbath.* Available at https://religionnews.com/2025/09/05/faiths-answer-to-ai-the-sabbath/ (accessed 5 November 2025).

Devlin J, **Chang M-W**, **Lee K and Toutanova K** (2018) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.*

European Commission (2024) *EU Artificial Intelligence Act.* Available at https://artificialintelligenceact.eu/ai-act-explorer/ (accessed 6 November 2025).

Falco G, **Shneiderman B**, **Badger J**, **Carrier R**, **Dahbura A**, **Danks D**, **Eling M**, **Goodloe A**, **Gupta J**, **Hart C**, **Jirotka M**, **Johnson H**, **LaPointe C**, **Llorens AJ**, **Mackworth AK**, **Maple C**, **Pálsson SE**, **Pasquale F**, **Winfield A and Yeong ZK** (2021) Governing AI safety through independent audits. *Nature Machine Intelligence 3*(7), 566–571.

Friedman HH (2025) *The Ten Commandments Impact: How Ten Timeless Principles Reshaped Civilization and Can Revitalize Your Organization. SSRN 5154208.*

Gabriel I (2020) Artificial intelligence, values, and alignment. *Minds and Machines 30*(3), 411–437.

Gabriel I, **Manzini A**, **Keeling G**, **Hendricks LA**, **Rieser V**, **Iqbal H**, **Tomašev N**, **Ktena I**, **Kenton Z and Rodriguez M** (2024) *The Ethics of Advanced AI Assistants. arXiv preprint arXiv:2404.16244.*

Gallegos IO, **Rossi RA**, **Barrow J**, **Tanjim MM**, **Kim S**, **Dernoncourt F**, **Yu T**, **Zhang R and Ahmed NK** (2024) Bias and fairness in large language models: A survey. *Computational Linguistics 50*(3), 1097–1179.

Green SK (2000) The fount of everything just and right? The ten commandments as a source of American law. *Journal of Law and Religion 14*(2), 525–558.

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

Hendrycks D, **Burns C**, **Basart S**, **Critch A**, **Li J**, **Song D and Steinhardt J** (2020) *Aligning AI with Shared Human Values. arXiv preprint arXiv:2008.02275*.

Heritage History (2025) *From the Seventh to the End of the Tenth Commandment.* Available at https://heritage-history.com/index.php?c=read&author=kinkead&book=baltimore4&story=seventh (accessed 5 November 2025).

Hout MC, **Papesh MH and Goldinger SD** (2013) Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science 4*(1), 93–103.

Hu EJ, **Wallis P**, **Allen-Zhu Z**, **Li Y**, **Wang S**, **Wang L and Chen W** (2022) LoRA: Low-rank adaptation of large language models. In Liu Y, Finn C, Choi Y and Deisenroth M (eds.), *International Conference on Learning Representations*. New York, NY: Curran Associates.

Irwin T (2011) The development of ethics: A historical and critical study. *The Philosophical Forum*. *42*(3), 269–335.

Jentzsch S, **Schramowski P**, **Rothkopf C and Kersting K** (2019) Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.*

Jobin A, **Ienca M and Vayena E** (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1*(9), 389–399.

Kasirzadeh A and Clifford D (2021) Fairness and data protection impact assessments. In Fourcade M, Kuipers B, Lazar S and Mulligan D (eds.), *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* New York, NY: Association for Computing Machinery (ACM).

Lemieux VL and Werner J (2024) Protecting privacy in digital records: The potential of privacy-enhancing technologies. *ACM Journal on Computing and Cultural Heritage 16*(4), 1–18.

Li B, **Zhou H**, **He J**, **Wang M**, **Yang Y and Li L** (2020) *On the Sentence Embeddings from Pre-Trained Language Models. arXiv preprint arXiv:2011.05864*.

Li VOK, **Lam JCK and Cui J** (2021) AI for social good: AI and big data approaches for environmental decision-making. *Environmental Science & Policy 125*, 241–246.

Li VOK, **Lam JCK**, **Han Y and Cheung LYL** (2024) *Designing a Value-driven GAI Framework for Social Good: Embedding Social Good Values into GAI Models.* Data for Policy 2024, London, UK. Available at https://members.dataforpolicy.org/2024-conference-day-3/.

Li VOK, **Lam JCK**, **Han Y**, **Crowcroft, J and Cheung LYL** (2025) "Replication Data for: Value-driven LLM for Social Good. Zenodo, Version v1 [dataset]. https://doi.org/10.5281/zenodo.17636407.

Madan R and Ashok M (2023) AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly 40*(1), 101774.

McInnes L, **Healy J and Melville J** (2018) *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426.*

Mesnard T, **Hardin C**, **Dadashi R**, **Bhupatiraju S**, **Pathak S**, **Sifre L**, **Rivière M**, **Kale MS**, **Love J**, **Tafti P**, **Hussenot L and Sessa PG** (2024) *Gemma: Open Models based on Gemini Research and Technology. arXiv preprint arXiv:2403.08295*.

Mikolov T (2013) *Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781*.

Mittelstadt BD, **Allo P**, **Taddeo M**, **Wachter S and Floridi L** (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society 3*(2), 2053951716679679.

Mühlhoff R and Ruschemeier H (2025) Updating purpose limitation for AI: A normative approach from law and philosophy. *International Journal of Law and Information Technology 33*, eaaf003.

Navigli R, **Conia S and Ross B** (2023) Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality 15*(2), 1–21.

OECD (2024) *AI Principles.* Available at https://www.oecd.org/en/topics/sub-issues/ai-principles.html (accessed 6 November 2025).

Ong JCL, **Chang SY-H**, **William W**, **Butte AJ**, **Shah NH**, **Chew LST**, **Liu N**, **Doshi-Velez F**, **Lu W**, **Savulescu J and Ting DSW** (2024) Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health 6*(6), e428–e432.

Ouyang L, **Wu J**, **Jiang X**, **Almeida D**, **Wainwright C**, **Mishkin P**, **Zhang C**, **Agarwal S**, **Slama K and Ray A** (2022) Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35*, 27730–27744.

Pappas IO and Vassilakopoulou P (2025) Human value alignment in AI. In Xu W (ed.), *Handbook of Human-Centered Artificial Intelligence*. Singapore: Springer, pp. 1–33.

Pennington J, **Socher R and Manning CD** (2014) Glove: Global vectors for word representation. In Moschitti A, Pang B and Daelemans W (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Pennsylvania, USA: Association for Computational Linguistics (ACL).

Preniqi V, **Ghinassi I**, **Ive J**, **Saitis C and Kalimeri K** (2024) MoralBERT: A fine-tuned language model for capturing moral values in social discussions. In Förster A, Manzoni P, Kalimeri K and Carvalho JP (eds.), *Proceedings of the 2024 International Conference on Information Technology for Social Good*. New York, NY: Association for Computing Machinery (ACM).

Radford A, **Wu J**, **Child R**, **Luan D**, **Amodei D and Sutskever I** (2019) *Language Models are Unsupervised Multitask Learners. OpenAI.*

Raza S, **Narayanan A**, **Khazaie VR**, **Vayani A**, **Chettiar MS**, **Singh A**, **Shah M and Pandya D** (2025) *HumaniBench: A Human-Centric Framework for Large Multimodal Models Evaluation. arXiv preprint arXiv:2505.11454*.

Reimers N and Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. In Inui K, Jiang J, Ng V and Wan X (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Pennsylvania, USA: Association for Computational Linguistics (ACL).

Ren S, **Jian P**, **Ren Z**, **Leng C**, **Xie C and Zhang J** (2025) *Towards Scientific Intelligence: A Survey of LLM-Based Scientific Agents. arXiv preprint arXiv:2503.24047*.

Roundtree AK (2023) AI explainability, interpretability, fairness, and privacy: An integrative review of reviews. In Degen H and Ntoa S (eds.), *International Conference on Human-Computer Interaction*. Cham, Switzerland: Springer.

Salinas A, **Shah P**, **Huang Y**, **McCormack R and Morstatter F** (2023) The unequal opportunities of large language models: Examining demographic biases in job recommendations by ChatGPT and llama. In Manshadi V, Mendler-Dünner C, Redmiles E and Rodriguez M (eds.), *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. New York, NY: Association for Computing Machinery (ACM).

Sangwa S and Mutabazi P (2025) Artificial intelligence ethics and biblical prophecy: A global Christian analysis of algorithmic censorship, digital deception, and eschatological risk. *Open Journal of Business Theology 3*(2). https://doi.org/10.5281/zenodo.15606271.

Schlender T and Spanakis G (2020) Thy algorithm shalt not bear false witness': An evaluation of multiclass debiasing methods on word embeddings. In Baratchi M, Cao L, Kosters WA, Lijffijt J, van Rijn JN and Tak FW (eds.), *Benelux Conference on Artificial Intelligence*. Cham, Switzerland: Springer.

Schramowski P, **Turan C**, **Andersen N**, **Rothkopf CA and Kersting K** (2022) Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence 4*(3), 258–268.

Shahriari K and Shahriari M (2017) IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In Krishnan S, Fernando X and Istrate E (eds.), *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*. New York, NY: IEEE.

Sharma S (2024) Benefits or concerns of AI: A multistakeholder responsibility. *Futures 157*, 103328.

Singhal A, **Neveditsin N**, **Tanveer H and Mago V** (2024) Toward fairness, accountability, transparency, and ethics in AI for social media and health care: Scoping review. *JMIR Medical Informatics 12*(1), e50048.

Solaiman I and Dennison C (2021) Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems 34*, 5861–5873.

Spitsberg T, **Kettler T and McKamie J** (2025) Large language model AI guided creative writing co-creation in secondary schools. *Theory Into Practice 64*(4), 374–389. https://doi.org/10.1080/00405841.2025.2528551.

Stein M, **Bernardi J and Dunlop C** (2024) *The Role of Governments in Increasing Interconnected Post-Deployment Monitoring of AI. arXiv preprint arXiv:2410.04931*.

Sun Z, **Shen Y**, **Zhou Q**, **Zhang H**, **Chen Z**, **Cox D**, **Yang Y and Gan C** (2024) Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems 36*.

Tao Y, **Viberg O**, **Baker RS and Kizilcec RF** (2024) Cultural bias and cultural alignment of large language models. *PNAS Nexus 3*(9), pgae346.

Thacker J (2020) *How Do I Think Biblically About Internet Privacy?* Available at https://radical.net/article/biblically-internet-privacy/ (accessed 5 November 2025).

Touvron H, **Lavril T**, **Izacard G**, **Martinet X**, **Lachaux M-A**, **Lacroix T**, **Rozière B**, **Goyal N**, **Hambro E and Azhar F** (2023) *Llama: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971*.

Umbrello S (2025) *AI Deepfakes and the Theft of Moral Authority.* Available at https://www.wordonfire.org/articles/ai-deepfakes-and-the-theft-of-moral-authority/ (accessed 5 November 2025).

United Nations (1948) *Universal Declaration of Human Rights.* Available at https://www.un.org/en/about-us/universal-declaration-of-human-rights (accessed 6 November 2025).

Voorhees CE (2021) *The 10 Commandments of AI.* Available at https://www.stjamesfaithlab.org/10-commandments-of-ai-jan-2021 (accessed 5 November 2025).

Wagner B (2019) Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping? In Emre B, Irina B, Liisa Albertha Wilhelmina J and Mireille H (eds), *Being Profiled:Cogitas Ergo Sum*. Amsterdam: Amsterdam University Press, pp. 84–89.

Wang D, **Yang K**, **Zhu H**, **Yang X**, **Cohen A**, **Li L and Tian Y** (2024) Learning personalized alignment for evaluating open-ended text generation. In Al-Onaizan Y, Bansal M and Chen Y-N (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Pennsylvania, USA: Association for Computational Linguistics (ACL).

Weidinger L, **Mellor J**, **Rauh M**, **Griffin C**, **Uesato J**, **Huang P-S**, **Cheng M**, **Glaese M**, **Balle B and Kasirzadeh A** (2021) *Ethical and Social Risks of Harm from Language Models. arXiv preprint arXiv:2112.04359*.

Wieringa M (2020) What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Celis E, Ruggieri S, Taylor L and Zanfir-Fortuna G (eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery (ACM).

Ziegler DM, **Stiennon N**, **Wu J**, **Brown TB**, **Radford A**, **Amodei D**, **Christiano P and Irving G** (2019) *Fine-Tuning Language Models from Human Preferences. arXiv preprint arXiv:1909.08593*.

Cite this article: Li VOK, Lam JCK, Han Y, Crowcroft J and Cheung LYL (2026). LLM for social good: A value-driven LLM framework to embed social good values. *Data & Policy*, 8: e19. doi:10.1017/dap.2026.10061

https://doi.org/10.1017/dap.2026.10061 Published online by Cambridge University Press