

Tracing Moral Values in LLMs: Towards a Dataset for Interpretability in Cultural Narrative Generation

Chiara Fulfaro¹, Enrico Mensa¹ and Rossana Damiano¹

¹Department of Computer Science, Università di Torino, Italy

Abstract

Moral values play a central role in cultural heritage communication, yet current Large Language Model (LLM) approaches mainly rely on prompt-based control, providing limited insight into how such values are internally represented. This paper presents the design of a dataset, called MIRA (Moral Internal Representation Analysis), intended to support the investigation of moral value representations in LLMs through activation-based analysis. The dataset comprises structured decision scenarios covering single-value decisions, conflicts between competing moral values, and non-moral decisions as a comparative baseline. All scenarios follow a common template while systematically varying context, narrative perspective, and moral configuration. Although motivated by cultural heritage storytelling, the dataset is conceived as a general-purpose resource for studying moral decision-making and value representations in LLMs.

Keywords

activation analysis, cultural heritage, dataset construction, large language models, mechanistic interpretability, moral decision-making, moral values, narrative generation

1. Introduction

Within cultural heritage, one of the most significant challenges of recent years has been rethinking how cultural institutions communicate with the public and developing strategies capable not only of enhancing visitor engagement but also of addressing the ethical, social and cultural values embedded in heritage interpretation, particularly when dealing with sensitive or controversial materials [1]. The use of narrative techniques to strengthen museum communication and increase visitor engagement represents a well-established and widely recognized practice [2].

Given their capacity to produce coherent, compelling and contextually rich narratives, Large Language Models (LLMs) have begun to attract attention in cultural heritage as tools for digital storytelling, museum communication and visitor engagement [3]. However, many LLM-based storytelling approaches rely primarily on prompt-level control: although such approaches can guide fluent and engaging narrative production, they provide limited insight into how models internally represent culturally and ethically meaningful dimensions associated with cultural heritage materials. Consequently, these dimensions risk being treated as superficial themes, reduced to explicit moral statements or even lost during the generation process [4]. More specifically, elements imposed during the generation process are not always organically integrated within the narrative and can surface as disconnected components. For instance, this has been observed in a system for generating curatorial narratives, where dynamic elements from Politi's dramatic situations would often appear as explicit labels in the generated text rather than being embodied through the corresponding characters and their actions [5].

This limitation connects to a broader and increasingly central problem in contemporary LLM research: the internal mechanisms through which these systems represent concepts, make distinctions and produce decisions remain difficult to interpret. In response to this issue, Mechanistic Interpretability has emerged as a research field concerned with addressing the black-box problem by reverse-engineering the internal

4th Italian Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2026, <https://ai4ch.di.unito.it/>), co-located with the 24th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2026). 6–9 October 2026, Perugia, Italy

✉ chiara.fulfaro@edu.unito.it (C. Fulfaro); enrico.mensa@unito.it (E. Mensa); rossana.damiano@unito.it (R. Damiano)

ORCID 0000-0001-7743-4999 (E. Mensa); 0000-0001-9866-2843 (R. Damiano)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

representations and computations that give rise to model behavior [6]. A particularly relevant line of work is the Linear Representation Hypothesis, according to which high-level concepts may be encoded as linear directions in a model’s representation space [7]. If such representations are identifiable, they can be measured through internal activations and later tested through targeted activation interventions: by artificially strengthening or weakening the internal representation associated with a given concept, it becomes possible to observe whether and how that concept shapes the model’s output, providing a concrete means of controlling generation itself.

These lines of research point to an open problem at the intersection of cultural heritage storytelling and mechanistic interpretability. In cultural heritage, moral values are central to responsible interpretation when dealing with ethically sensitive or socially complex topics. In mechanistic interpretability, they represent a challenging class of abstract and context-dependent concepts whose internal encoding and functional role in LLMs remain difficult to assess. Investigating moral values and moral decisions in LLMs therefore becomes relevant to both responsible and culturally meaningful narrative generation, especially considering both the growing interest in the conscious adoption of generative AI within cultural heritage institutions [8] and the growing attention to interpretable representations in language models.

In this paper, we present the preliminary stage of a broader research work investigating whether and how moral values are represented in the internal activations of LLMs, and how such representations can be used to inform the generation of ethically and culturally meaningful curatorial narratives. As this research is still at the design stage, we describe the methodology underlying the ongoing construction of the Moral Internal Representation Analysis (MIRA) dataset to support this investigation. Its application, together with the subsequent stages of the research, will be addressed in future work.

2. Objectives

This research sets out to investigate how moral values can be used as theoretically grounded target dimensions to study whether moral decisions are linearly decodable from the internal activations of LLMs, and to apply this knowledge to the generation of curatorial narratives.

Cultural heritage communication does not merely involve the transmission of factual information, but also the articulation of deeper layers of meaning through which the public can connect the past to present-day ethical, social and cultural questions. For instance, in an interactive story where visitors take the role of an archaeologist, choosing whether to open a sarcophagus or limit the investigation to its exterior may reflect different moral attitudes toward archaeological research, respect for the dead, and the ethical treatment of human remains. Being able to identify and intentionally modulate these moral dimensions could support more transparent and controllable narrative generation. Understanding how moral dimensions are encoded in LLMs and how moral decisions can be decoded from internal activations may therefore support the development of more effective storytelling systems, in which specific moral dimensions can be emphasized, balanced or constrained according to curatorial and communicative goals.

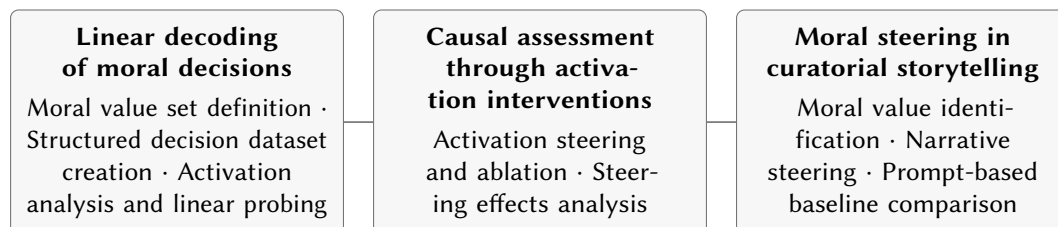


Figure 1: Overview of the research methodology.

More specifically, the research will proceed in three steps, as depicted in Fig. 1:

- **Linear decoding of moral decisions.** In this phase we will examine whether moral decisions, structured around theoretically defined moral values, are linearly decodable from LLM internal activations.
- **Causal assessment through activation.** Research will assess the functional role of the identified representations through controlled activation interventions.
- **Moral steering.** Finally, it will investigate how this knowledge can support the generation of curatorial narratives, enabling the intentional modulation of moral dimensions while preserving factual grounding, narrative coherence, and curatorial relevance.

In this paper, we lay the groundwork for the first stage of this research direction by presenting the design and methodology underlying the construction of a dataset for the analysis of moral decision-making in LLMs. Although developed to address the methodological requirements of this study, the dataset is conceived as a general-purpose resource for investigating moral values, decision-making processes and value representations.

3. Related Work

Several resources have recently been proposed to investigate moral reasoning, value alignment, and decision-making in language models, each addressing different aspects of these problems.

The ETHICS dataset [9] is a large collection of unambiguous moral scenarios grounded in multiple normative ethics frameworks and spanning diverse real-world situations, used to evaluate whether models can identify morally acceptable or unacceptable actions. Similarly, Moral Stories [10] is a structured corpus of short narratives for goal-oriented moral reasoning grounded in social situations. It aims to examine whether generative models are able to successfully navigate such situations to achieve specific goals while adhering to unspoken rules of social conduct and remaining consistent with the actor’s intention, as well as anticipating the consequences of the resulting actions.

A different line of work focuses more explicitly on the attribution of values themselves. ValuePrism, developed to train the Value Kaleidoscope model [11], is a large-scale dataset that associates decision-making situations to sets of context-dependent values, rights and duties; this work was designed to explore the extent to which AI systems can model pluralistic human values.

Other datasets have also addressed conflicts between competing values. DailyDilemmas [12] provides a collection of real-world moral dilemmas spanning diverse everyday topics; each one presents two alternative actions associated with different human values, with no choice being clearly correct. Grounded in five theoretical frameworks and considering the perspectives of stakeholders, the dataset is used to examine how models prioritize values under such conflicts and to assess whether there are discrepancies between their stated principles and the values actually reflected in their decisions. Multi-step Moral Dilemmas [13] is a dataset specifically constructed to evaluate the evolving moral judgments of LLMs. The scenario structure is similar to the one used in the previous dataset, but organized across multiple stages: each one begins with a simple value conflict and systematically introduces new tensions, forcing models to reconcile prior decisions.

Although these resources are valuable for studying moral reasoning and value alignment, they do not employ a consistent scenario template across all decision situations. Consequently, scenarios vary in both their structure and presentation, making systematic comparisons more difficult. They typically focus on a single relationship between moral values and decision choices: either alternatives representing the opposing poles of a single moral value or supported by two conflicting moral values. Moreover, they do not include comparable non-moral decision scenarios that can serve as a baseline. Such controlled settings are essential for mechanistic interpretability approaches, motivating us to introduce a new dataset specifically designed for this purpose.

4. Dataset Design

Moral Internal Representation Analysis (MIRA) is a dataset designed to investigate different aspects of moral decision-making while providing a comparative baseline for distinguishing moral-specific patterns from general decision-making processes. Moral oppositions, which have characterized dramatic writing since antiquity, are not specific to the cultural domain, although they acquire particular relevance in this context. For this reason, the dataset is built around values of general relevance, which will subsequently be used to generate curatorial narratives.

The dataset is currently being constructed by generating scenarios with an LLM through a dedicated prompt that specifies general constraints and the structure of each scenario.

4.1. Moral Values

Research on moral values spans a wide range of disciplines, from social psychology (*Schwartz's Theory of Basic Human Values* [14], *Haidt's Moral Foundations Theory* [15]) to philosophy of rationality (*van Fraassen's account of values and practical reason* [16]). In addition, the role of moral values in shaping compelling and meaningful narratives has been extensively investigated in drama studies and scriptwriting theory (*Polti's Thirty-Six Dramatic Situations* [17], *Williams' Moral Premise* [18]).

Building on this multidisciplinary background, the definition of a theoretically grounded set of moral values represents a fundamental preliminary step in the construction of moral decision scenarios. These values should not only be relevant for the investigation of moral decision-making, but also be suitable for the systematic construction of the diverse moral scenario types adopted in this dataset. In particular, selected moral dimensions should be characterized by opposing poles that capture a broad spectrum of value-related behaviors and support the construction of meaningful conflicts with other moral values. Moreover, they should be sufficiently general to be instantiated across diverse contexts and domains. For instance, the opposite values of *care* and *harm* [15] represent a particularly suitable candidate pair: it is characterized by two clearly distinct poles that can manifest through a variety of actions and contexts, while also giving rise to genuine trade-offs with other moral dimensions.

Research on moral values has proposed several theoretical frameworks for categorizing moral dimensions. Rather than proposing an independent set of values, we rely on established theories such as Moral Foundations Theory (MFT) for selecting the moral dimensions considered in our dataset [15]. MFT identifies six moral foundations that are argued to be universal across cultures and societies. Each foundation is organized around a pair of opposing poles, making it particularly suitable for the systematic construction of moral decision scenarios.

The initial set of selected moral dimensions consists of *care/harm*, *fairness/cheating* and *truth/distortion*. The first two dimensions directly correspond to moral foundations identified in MFT, while the third is included due to its relevance in contemporary communication contexts, including cultural heritage domains. This limited selection is intended to establish the core of the dataset for its first stage of development and evaluation, while additional moral values will be incorporated in subsequent iterations to progressively expand the coverage of the dataset.

4.2. Structure

According to the proposed design, the dataset is structured into three complementary subsets, each comprising a specific category of decision scenarios designed to support the analysis of different aspects of moral decision-making: (i) the first subset includes single-value moral decision scenarios, in which *one moral dimension* at a time is varied across contrasting poles, such as *care* versus *harm*; (ii) the second subset addresses non-trivial moral dilemmas in which two distinct moral values support conflicting choices, allowing the analysis of trade-offs between *competing moral dimensions*; and (iii) the third subset consists of meaningful *non-moral decision* scenarios and provides a comparative baseline for the activation-based analyses, allowing the assessment of whether decoded representations are specific to moral decision-making rather than decision-making in general.

Despite the differences across the three subsets, each scenario in MIRA follows a common structure:

Scenario Structure
TASK Answer using only A or B.
SCENARIO <i>Description of the situation</i>
OPTIONS A) <i>Option A</i> B) <i>Option B</i>
ANSWER

The **TASK** section specifies the required response format, instructing the model to select one of the two available options by providing only the corresponding letter; this ensures a standardized output across all scenarios and facilitates controlled activation-based analyses. The **SCENARIO** presents the broader context and the specific situation in which the choice is framed; depending on the subset, the described situation may or may not be associated with an underlying moral dimension. The two available alternatives are presented in the **OPTIONS** section and may correspond to opposing poles of the same moral dimension (e.g., *care* versus *harm*), reflect conflicting moral values (e.g., *fairness* versus *loyalty*) or represent two non-moral options. Finally, **ANSWER** provides a consistent reference point at which model activations can be analyzed before the generation of the final response.

The scenarios will span different domains and include historically or culturally situated cases to test whether the identified dimensions emerge in cultural heritage settings. To study the effect of narrative perspective, each scenario is presented in three versions by varying only the narrative person, namely the first, second, or third person, while keeping all other elements unchanged. This allows us to assess whether narrative perspective influences the internal encoding and linear decodability of moral decisions.

The following example provides a concrete instance of how a single-value moral scenario is instantiated in the dataset:

Scenario Structure
TASK Answer using only A or B.
SCENARIO X is attending a team meeting where a new team member is presenting a report and makes several visible mistakes in front of everyone.
OPTIONS A) Point out the mistakes in front of the whole team to set an example of proper standards. B) Wait until after the meeting and go through the mistakes privately, offering guidance on how to improve.
ANSWER

It should be noted that the proposed scenarios do not explicitly specify the alternative narrative developments required to support interactive storytelling. This is intentional, since narrative branches naturally follow from the value-based decision investigated in each scenario and can be generated, with varying degrees of directness, from the selected alternative.

4.3. Construction

To enable reliable analyses of internal representations, the proposed dataset must contain a large number of decision scenarios spanning diverse contexts, situations, and narrative perspectives. Manually constructing such a dataset does not scale well, making it difficult to ensure consistency and diversity across scenarios; LLMs offer a scalable alternative, generating varied scenarios while satisfying predefined structural and semantic constraints. Accordingly, a pilot of the proposed dataset is generated using Claude Sonnet 5, guided by prompts specifically designed for the requirements of each subset. The prompt template used for generation, exemplified for the first subset, is as reported in Appendix A.

For the other two subsets, the prompt template follows the same structure, with a few targeted modifications reflecting the specific requirements of each one. For the second subset, the placeholders **{positive_pole}** and **{negative_pole}** are replaced by **{moral_value_1}** and **{moral_value_2}**. Minor adjustments to the **[TASK]**, **[SCENARIO]** and **[FORM AND STYLE]** sections are required to reflect the fact that two moral values are now involved instead of one. For the third subset, all references to moral values are removed. A **{decision_criterion}** placeholder is introduced to ensure the generation of meaningful decisions; it represents the practical dimension along which the two alternatives differ (e.g., *efficiency/cost*). Given the substantial differences with respect to the previous subsets, all sections of the template require minor modifications. Note that the **[BALANCING]** section is removed entirely, since the distribution of the different criteria is not relevant. For both subsets the **[OUTPUT]** section is adjusted to reflect the changes; for brevity, the resulting modifications are not detailed here.

5. Conclusion and Future Work

In this paper we presented the design and methodology for constructing a dataset intended to support the investigation of internal representations of moral values in Large Language Models in cultural settings. Unlike existing resources, the proposed dataset adopts a common template for all decision scenarios and combines three categories of decision scenarios: choices between opposing poles of a single moral value, conflicts between competing moral values and non-moral decisions. This design enables systematic comparisons across scenarios that differ in context, narrative perspective and decision situation, while maintaining a consistent structure that supports controlled mechanistic interpretability analyses. The moral values selected for the dataset are grounded in established moral frameworks, ensuring that scenario generation is based on conceptually rigorous and theoretically motivated moral dimensions. Beyond the dataset structure itself, the proposed methodology provides a prompting strategy for the large-scale generation of these structured scenarios.

The design presented in this work should be considered a preliminary version that requires further consolidation and refinement. A necessary next step will involve experts (for example, museum curators and educators) to verify that the alternatives in each scenario are correctly aligned with their intended moral values. In future iterations, the dataset could be extended to incorporate additional moral values and a broader range of historically and culturally situated decision contexts.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT 5 in order to check Grammar and spelling and to generate Figure 1. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

A. Prompt Template

Scenario Structure

[TASK]

Generate **{n_scenarios}** moral decision scenarios based on the moral value **{moral_value}**. Each scenario must be presented in three versions and include two alternatives associated with the opposite poles of the moral dimension: **{positive_pole}** and **{negative_pole}**.

[SCENARIO]

Each scenario must include:

- A concrete and plausible situation within the **{context}** context;
- A decision centered on the moral value **{moral_value}**, to be made by a single individual;
- Two alternatives, A and B: one associated with **{positive_pole}** and the other with **{negative_pole}**. The polarity associated with an alternative must not depend on its position (A or B) or its length.

[SCENARIO STRUCTURE]

Each scenario must follow exactly the template below:

TASK

Answer using only A or B.

SCENARIO

<Description of the situation>

OPTIONS

A) <Option A>

B) <Option B>

ANSWER

Rules:

- Replace only the placeholders in angle brackets;
- Do not modify the headers 'TASK', 'SCENARIO', 'OPTIONS', and 'ANSWER';
- Do not modify the sentence under 'TASK';
- Keep the labels 'A)' and 'B)' exactly as indicated;
- Do not add any text after 'ANSWER'.

[BALANCING]

Distribute the position of the poles as evenly as possible:

- **{positive_pole}** must appear in A and in B a similar number of times;
- **{negative_pole}** must appear in A and in B a similar number of times.

Distribute the length of the alternatives as evenly as possible, including:

- Cases in which A and B have similar length;
- Cases in which A is longer than B;
- Cases in which B is longer than A.

[FORM AND STYLE]

Describe hypothetical situations involving a decision that has not yet been made. In the scenario description, use the present tense and maintain a neutral, non-judgmental style: avoid phrasing that makes one of the two alternatives more desirable than the other. The alternatives must be expressed using verbs in the infinitive form, with no conjugated verbs and no reference to a specific subject.

Do not explicitly name **{moral_value}**, **{positive_pole}**, or **{negative_pole}** in the text of the scenario or the alternatives.

[VERSIONS]

Each scenario must be presented in three versions, in which the individual who must make the decision is expressed in:

- First person singular;
- Second person singular;
- Third person singular expressed as a name, a role or a pronoun.

The three versions must be identical in every element, except for the grammatical adaptations required by the change of person: pronouns, possessives, agreement and verb forms.

The gender of the individual may vary across different scenarios, but must remain the same across the three versions of the same scenario.

[OUTPUT]

Return a JSON array in which each scenario has the following structure:

```

...
{
  "scenario_id": "<id>",
  "moral_value": "{moral_value}",
  "context": "{context}",
  "narrative_person": "P1" | "P2" | "P3",
  "text": "...",
  "polarities": {
    "A": "{positive_pole}" | "{negative_pole}",
    "B": "{positive_pole}" | "{negative_pole}"
  }
}
...

```

‘scenario_id’ must be constructed following the format:

{moral_value_code}_{context_code}_<progressive_number>_<narrative_person>

Where:

- <progressive_number> is a three-digit progressive number (001, 002, 003, ...);
- <narrative_person> always matches the value of the narrative_person field.

The three versions of the same scenario must share the same <progressive_number>.

Do not add any text before or after the JSON.

References

- [1] B. L. Murphy (Ed.), *Museums, Ethics and Cultural Heritage*, Routledge, London and New York, 2016.
- [2] L. Bedford, *Storytelling: The real work of museums*, *Curator: The Museum Journal* 44 (2001) 27–34. doi:10.1111/j.2151-6952.2001.tb00027.x.
- [3] G. Trichopoulos, *Large language models for cultural heritage*, in: *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter*, ACM, 2023, pp. 1–5. doi:10.1145/3609987.3610018.
- [4] F. Bu, Z. Wang, S. Wang, Z. Liu, *An investigation into value misalignment in LLM-generated texts for cultural heritage*, 2025. doi:10.48550/arXiv.2501.02039. arXiv:2501.02039.
- [5] E. Mensa, A. E. Pecora, C. Fulfaro, A. Pizzo, E. Ferraris, A. Bottino, R. Damiano, *From a true story: Leveraging museum catalogue data for LLM-driven narrative generation*, *ACM Transactions on Intelligent Systems and Technology* (2026).
- [6] D. Rai, Y. Zhou, S. Feng, A. Saparov, Z. Yao, *A practical review of mechanistic interpretability for transformer-based language models*, 2024. doi:10.48550/arXiv.2407.02646. arXiv:2407.02646.
- [7] K. Park, Y. J. Choe, V. Veitch, *The linear representation hypothesis and the geometry of large*

- language models, in: Proceedings of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 39643–39666.
- [8] A. Tripaldi, S. Natale, L. Befera, AI Compass per il patrimonio culturale: Riflessioni per un’adozione consapevole dell’IA generativa, Research Report, University of Turin, Turin, Italy, 2026. URL: <https://hdl.handle.net/2318/2128757>.
 - [9] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, J. Steinhardt, Aligning AI with shared human values, in: International Conference on Learning Representations (ICLR), 2021.
 - [10] D. Emelin, R. Le Bras, J. D. Hwang, M. Forbes, Y. Choi, Moral stories: Situated reasoning about norms, intents, actions, and their consequences, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 698–718. URL: <https://aclanthology.org/2021.emnlp-main.54>. doi:10.18653/v1/2021.emnlp-main.54.
 - [11] T. Sorensen, L. Jiang, J. D. Hwang, S. Levine, V. Pyatkin, P. West, N. Dziri, X. Lu, K. Rao, C. Bhagavatula, M. Sap, J. Tasioulas, Y. Choi, Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 19937–19947.
 - [12] Y. Y. Chiu, L. Jiang, Y. Choi, Dailydilemmas: Revealing value preferences of LLMs with quandaries of daily life, in: The Thirteenth International Conference on Learning Representations (ICLR), 2025.
 - [13] Y. Wu, Q. Sheng, D. Wang, G. Yang, Y. Sun, Z. Wang, Y. Bu, J. Cao, The staircase of ethics: Probing LLM value priorities through multi-step induction to complex moral dilemmas, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 15950–15970. URL: <https://aclanthology.org/2025.emnlp-main.806/>.
 - [14] S. H. Schwartz, An overview of the schwartz theory of basic values, *Online Readings in Psychology and Culture* 2 (2012). doi:10.9707/2307-0919.1116.
 - [15] J. Haidt, C. Joseph, Intuitive ethics: How innately prepared intuitions generate culturally variable virtues, *Daedalus* 133 (2004) 55–66.
 - [16] B. C. Van Fraassen, Values and the heart’s command, *The Journal of Philosophy* 70 (1973) 5–19.
 - [17] G. Polti, *The Thirty-Six Dramatic Situations*, The Writer, Boston, 1916.
 - [18] S. Williams, *The Moral Premise: Harnessing Virtue and Vice for Box Office Success*, Michael Wiese Productions, Studio City, CA, 2006.