electronics

MDPI

Article

Towards an End-to-End Personal Fine-Tuning Framework for AI Value Alignment

Eleanor Watson [1,*], Thiago Viana [1], Shujun Zhang [1], Benjamin Sturgeon [2] and Lukas Petersson [3]

[1] School of Computing and Engineering, University of Gloucestershire, The Park, Cheltenham GL50 2RH, UK; tviana1@glos.ac.uk (T.V.); szhang@glos.ac.uk (S.Z.)
[2] Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch 7701, South Africa; strben005@wf.uct.ac.za
[3] School of Engineering, Lund University, P.O. Box 118, 221 00 Lund, Sweden; lu7836pe-s@student.lu.se
* Correspondence: eleanorwatson@connect.glos.ac.uk

Abstract: This study introduces a novel architecture for value, preference, and boundary alignment in large language models (LLMs) and generative AI systems, accompanied by an experimental implementation. It addresses the limitations in AI model trustworthiness stemming from insufficient comprehension of personal context, preferences, and cultural diversity, which can lead to biases and safety risks. Using an inductive, qualitative research approach, we propose a framework for personalizing AI models to improve model alignment through additional context and boundaries set by users. Our framework incorporates user-friendly tools for identification, annotation, and simulation across diverse contexts, utilizing prompt-driven semantic segmentation and automatic labeling. It aims to streamline scenario generation and personalization processes while providing accessible annotation tools. The study examines various components of this framework, including user interfaces, underlying tools, and system mechanics. We present a pilot study that demonstrates the framework's ability to reduce the complexity of value elicitation and personalization in LLMs. Our experimental setup involves a prototype implementation of key framework modules, including a value elicitation interface and a fine-tuning mechanism for language models. The primary goal is to create a token-based system that allows users to easily impart their values and preferences to AI systems, enhancing model personalization and alignment. This research contributes to the democratization of AI model fine-tuning and dataset generation, advancing efforts in AI value alignment. By focusing on practical implementation and user interaction, our study bridges the gap between theoretical alignment approaches and real-world applications in AI systems.

Keywords: machine learning; annotation; alignment; framework; foundation models

Citation: Watson, E.; Viana, T.; Zhang, S.; Sturgeon, B.; Petersson, L. Towards an End-to-End Personal Fine-Tuning Framework for AI Value Alignment. *Electronics* **2024**, *13*, 4044. https://doi.org/10.3390/electronics13204044

Academic Editor: Alejandro L. Borja

Received: 7 August 2024
Revised: 23 September 2024
Accepted: 7 October 2024
Published: 14 October 2024

1. Introduction

Generative AI, particularly Large Language Models (LLMs) like ChatGPT, has rapidly advanced and become integral to daily life and the global economy [1]. This integration raises serious questions about value alignment: ensuring AI systems behave in accordance with human values and intentions.

Machine learning, a subset of AI focused on creating systems that can learn from data, plays a crucial role in developing these models. However, current AI systems, including LLMs, often struggle to fully comprehend personal contexts and accommodate diverse cultural expressions, limiting their trustworthiness and potentially leading to biased or inappropriate outputs. This paper presents a novel framework for personalizing AI alignment, focusing on practical methods to integrate individual user values and preferences into AI systems, particularly LLMs.

Unlike traditional software with predetermined responses, AI systems can generate novel outputs and make complex decisions, potentially leading to unexpected behaviors

that may not align with user values or intentions. AI systems therefore require more sophisticated and adaptable alignment mechanisms. This framework addresses these AI-specific challenges by providing mechanisms for continuous alignment and personalization.

Our framework addresses this challenge by providing mechanisms for users to impart their values and expectations swiftly through a token system, enabling personalized AI interactions across various contexts.

Alignment with user goals does not automatically confer safety or trustworthiness. Moreover, existing foundation models often have capabilities exceeding initial expectations, underscoring the need to develop examples that foster stronger alignment within these models rather than retrofitting behavioral mimicry after the fact through mechanisms such as fine-tuning and Reinforcement Learning from Human Feedback.

Creating datasets and personalization examples with the right balance of nuance, diversity, and scope of human preferences across various cultural and situational contexts could substantially improve the contextualization of AI alignment. In addition, eliciting values preferences with the aid of language models could enhance the personalization of fine-tuning for individuals and groups, leading to superior results [2–4].

Achieving this entails improving the ability of AI systems to comprehend, predict, and adapt to human needs at a personal and local level rather than on a global scale. Enhancing AI systems to comprehend personal and local contexts can strengthen a system's resilience against misinterpretations, which may lead to biased machine evaluations. By accommodating others' preferences, which are crucial to prosocial behavior, one can enable an agent to take action to nourish the flourishing of another person.

A culturally and politically contextualized Internet-scale dataset of human values and preferences can be interpreted through a Foundation Model to provide a reference for appropriate behavior-specific situations. LLMs such as ChatGPT have been demonstrated to outperform crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection, with twenty-fold reduction in cost. Indeed, many Human Intelligence Task workers appear to outsource tasks to language models whenever they can for the sake of expediency, creating downstream problems in research. However, as a corollary, a tremendous leap forward in the ability to understand, predict, and accommodate human preferences can be achieved in a radically simplified and affordable manner by linking human intelligence with sophisticated prompt-driven annotation mechanisms.

LLMs use Reinforcement Learning from Human Feedback (RLHF) and other instructional tuning methods to improve user responsiveness. However, these techniques often produce outwardly 'nice' responses that lack alignment with users' personal and cultural values [5]. This surface-level 'niceness' does not constitute broad-based representations of the values of all users, making it similar to eliciting pleasing phrases without genuine understanding and potentially leading to unwanted behavior which is sometimes described as 'sycophantic' [6]. To bridge this gap, this study proposes a prototypical behavioral annotation framework using LLM/diffusion model technologies to simplify and optimize prompt/chat-driven annotation and fine-tuning. The insights gained will contribute to an evolving theory on how enhanced behavioral annotation and fine-tuning can advance value alignment.

Following this introduction, Section 2 sets out the background to the study and draws upon relevant literature. Section 3 then outlines the main elements of the research method. Section 4 establishes and explains the outline framework for the subsequent analysis and development detailed in the ensuing Results Section 5. Section 6 then discusses some key issues related to the results, including framework validation. Finally, Section 7 summarizes the main responses to the RQs, and highlights the main contributions. Limitations of the research and possible future research avenues are also outlined.

2. Background and Relevant Literature

This study introduces a framework aimed at simplifying public involvement in AI alignment through context-linked annotation and personalized fine-tuning. The ultimate

goal of this study is to facilitate most users in creating a personalized token fine-tuning process that is easily incorporated into third-party systems. It may also be supported by further refinements such as annotated behavioral examples which encode examples of values.

As technology progresses, the annotation process will evolve to include more automation layers, such as pre-annotation segmentation and prompt-driven annotation methods seen in tools such as LabelStudio, MTTR, and Toolformers. More advanced processes for media searches, sanitation, validation, and augmentation will be incorporated as resources permit.

The output of the fine-tuning process, which is anticipated to take no longer than 30 min at most, will be a token that can be ported into AI systems through an API to better align them to the user's specified values. The token encodes inputs into a vector format that can be used for model fine-tuning, adjusting the model's parameters to minimize the divergence between its output and the user-specified values represented by the vector. The API would look up the model parameters using the token, load the appropriate model, and then generate responses that align with the user's specified values.

These developments are applicable to the general development and filtering/curation of datasets and fine-tuning, particularly multimodal datasets that are ideal for Foundation Models. Annotation is typically a time-consuming, expensive, and often thankless process. This is especially true in multimodal contexts with greater complexity. Moreover, improved annotation techniques can help to de-risk datasets, ensuring that they are representative, and therefore suitable for generating benchmarks, training, and test sets derived from it. This is especially important given the emerging literature on how the quality of data strongly influences the performance of outputs, with sets of modest size producing better results than those that are large but lack definition or integrity.

These efforts will democratize AI alignment processes, enabling radical participation by the public, who otherwise have very little control over socializing the behavior of AI systems towards their preferences. It will also provide crucial information for AI systems to better recognize behavior and the probable intention behind it, enabling fewer incidences of bias through misapprehension or a lack of examples.

Globally meaningful and fair systems require diverse input. The temporal aspects of data (and its potential biases) necessitate a program of progressive improvement and richness. Only a large corpus of continually improving annotations by members of the public can be relied upon to achieve this. Similar to Wikipedia, a sufficiently large number of participants can eclipse the capabilities of experts or paid content creators. At the same time, open mechanisms do not preclude leading contributions being made by experts.

There are several potentially catastrophic concerns regarding AI that warrant serious consideration. This poses a risk of moral panic towards AI research in general (even the safety subset) if the public does not feel assured that good work is being conducted to alleviate their concerns, and that they have an opportunity to participate in a modest yet meaningful manner. Thus, it is important to consider how AI interfaces can be streamlined. The technology of ChatGPT was accessible for a long time, before relatively minor user-experience changes created a transformative experience. Similar opportunities are likely to exist for annotation and personalization.

Hundreds of millions of people are directly interacting with AI agents on a daily basis. However, AI systems should be capable of acting in a manner agreeable to the user as well as taking user preferences into account. Current AI systems are generally attuned to a set of values, which, while politically correct, the majority of the global population (and even the U.S. population) do not fully share.

Ideal AI interactions should allow users to impart their values and expectations swiftly through a token system. This could enable personalized experiences, such as pre-socializing hotel AIs or enabling a nuanced understanding of reclaimed terms by minorities.

There is potential to develop user-friendly tools that produce rich multimodal data. Leveraging prompt-driven mechanisms can simplify and automate annotation processes,

translating simple user inputs, like "The man in the red shirt is being rude at around 1:28", into complex behavioral insights. However, this approach has potential pitfalls, such as the risk of perceiving data-driven ethics as absolute moral guidelines, which they are not.

Involving the public in dataset annotation can diversify perspectives and alleviate AI-related fears, providing an active avenue for people's concerns, rather than fostering reactionary attitudes. It also reduces groupthink by sampling from a broader selection of people interested in the AI Alignment space.

In this context, this article reports on the development of a framework aimed at simplifying public involvement in AI alignment, thereby addressing the following research questions (RQs):

RQ1. How can efficient, user-friendly value personalization functions for AI systems be designed and implemented?

RQ2. How can the rapid and simple identification, sanitization, and annotation of behavioral value/norm examples be supported?

RQ3. How best can values be extracted in a manner best suited for model learning and alignment?

3. Research Method

This study employs a multi-faceted research approach to develop and validate a framework for personalized AI alignment. Our methodology combines:

- A comprehensive literature review to identify key elements of AI alignment and personalization.
- A conceptual framework development based on identified best practices and emerging technologies.
- A prototypical implementation and testing of key framework components.

This approach allows us to bridge theoretical concepts with practical applications, addressing the complex challenge of AI alignment from multiple angles.

We applied a pragmatic research method that combines framework development based on literature analysis with experimentation to validate the framework. This is largely an inductive paradigm, with a form of experimentation to validate this framework, akin to positivism or subjectivism. This aligns with the notion that AI systems should not only be technically efficient but also embody human-like reasoning and moral considerations, akin to the ideas presented in works like Floridi's "The Ethics of Information" [7].

This study was informed by a prior scoping literature review to identify the key elements of a provisional conceptual framework. With these identified, the active conceptual development of a framework has been undertaken to fulfill the goals of the research. In line with the Research Process Onion model, a methodological framework designed to help researchers peel away the layers of complexity involved in designing a research project was employed.

This study explored the challenges and opportunities offered by automated behavioral annotation techniques, specifically prompt-driven foundation model mechanisms and those that provide additional features that are best suited for developing multimodal datasets. This was conducted to establish a foundation for future research with special attention given to recent studies featuring the latest innovations in Foundation Models.

Inductive approaches were also derived from observations made during a scoping systematic literature review of tools which relate to Behavioral Annotation, Watson et al. [8]. Our approach involves building upon these findings by critically analyzing existing research to identify patterns, gaps, and new insights in AI alignment. This paper aims to build upon these findings by presenting an overview of a framework that can actualize them. A comparison of the two studies is displayed in Figure 1.

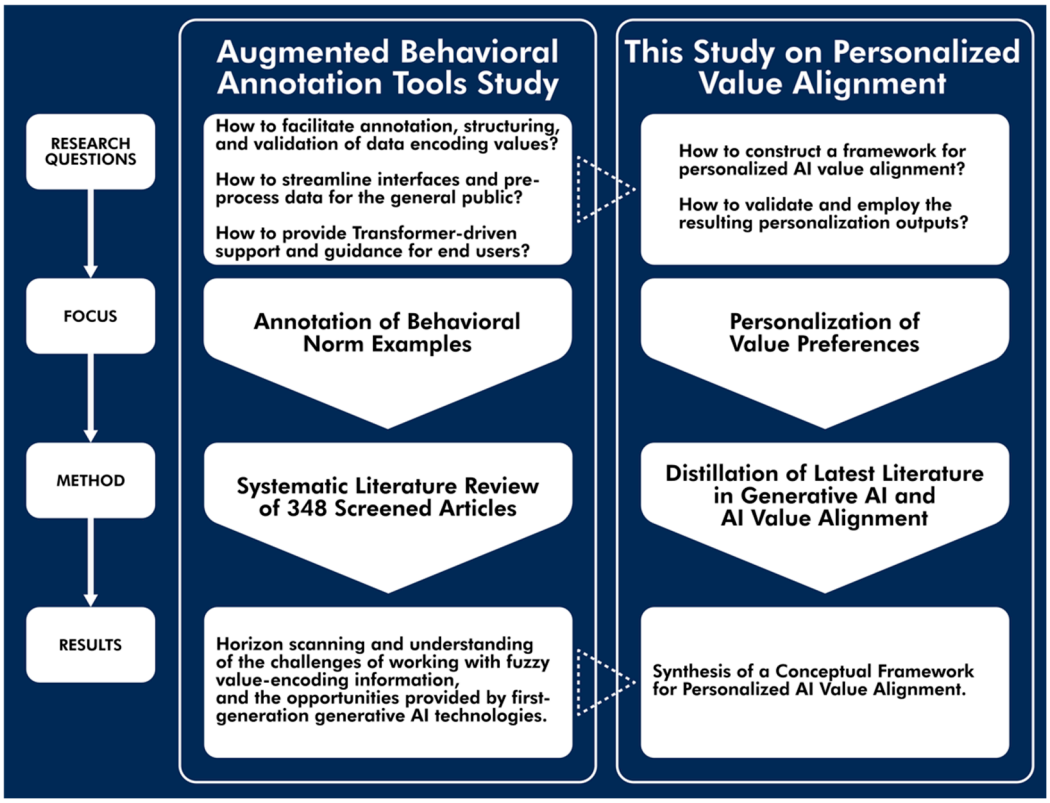

Figure 1. A comparison of the two studies of research in the overall project thus far.

The literature on conceptual frameworks and models was used to identify important components and concepts for the framework, as part of the various modular elements in the developing framework. Our action research approach involved iterative cycles of planning, action, and reflection. We collaborated with AI developers and potential end-users throughout the framework development process, incorporating their feedback to refine our approach. This method allowed us to bridge the gap between theoretical concepts and practical implementation challenges. This permitted the study of AI systems within their cultural contexts (ethnography) and integrated diverse disciplinary perspectives.

A longitudinal time horizon allows for the observation of trends, evolution, and adaptation in the field, as advocated for within dynamic processes. Observational and analytical data techniques, which combine real-world observations with rigorous analysis, provide nuanced insights into AI alignment and emphasize the importance of detailed, methodical observation and analysis in research.

4. An Outline Framework for Personalized Value Fine-Tuning

This section outlines how the proposed annotation framework will work and what features are intended to be included. We define behavioral annotation as labeling data (text, video, images, etc.) with indicators of the behaviors demonstrated. This may include not only labels of the behavior itself but also labels of the values demonstrated by that behavior.

A thorough analysis of studies on behavioral annotation was presented in a previous paper by Watson et al. [8]. The study examined recent major innovations and best practices in the field of behavioral annotation, with the goal of enhancing the efficiency and effectiveness of these processes to inform value alignment efforts. However, the implementation of individual methods, although promising, is not sufficient. These elements must be integrated, and significant research gaps remain regarding how to accomplish this goal.

Applying these opportunities to democratize value specification and fine-tuning processes for AI models therefore requires a supportive framework. This should facilitate

personalization, automatic and semi-automatic behavioral annotation, and allow for independent update of sub-components as technology improves.

The eventual goal is to produce a token output that can be directly interfaced with third-party models, as well as to benchmark model performance in respect to exemplar personas. However, significant further research is required to interface value preferences (textual or vector) with machine learning models to substantially alter their behavior, especially in a manner which provides greater safety and robustness than basic custom instructional context. Theoretical approaches in representation engineering highlight potential means to accomplish this, but significant uncertainty remains, which is beyond the scope of this paper [9]. In the interim, the proposed annotation framework facilitates the encapsulation of distinct values within text datasets. This process enables the creation and distribution of specialized datasets, tailored for model fine-tuning in alignment with predefined use cases. Essentially, this approach permits the extraction of subsets from a larger text corpus, based on identified values of interest. Consequently, users can obtain a unique token that delineates the dataset according to their specific requirements, thereby allowing access to all relevant textual information within that segmented cluster.

Despite the rapid advancement in machine learning technology, this pipeline is designed to be accessible via a wide number of devices. This framework outlines the general process of facilitating successive automated annotation processes. It is designed with modular and independent subparts, allowing individual components to be quickly retrofitted as new capabilities emerge. The framework also encompasses features that have been demonstrated individually or prototypically but are not yet widely deployed. Technological advancement is expected to enhance these subprocesses as techniques continue to evolve and improve. The modularity of the pipeline also allows for easy inclusion of unforeseen capabilities. Figure 2 outlines a general data flow for this pipeline. The bullets between each module reflect the presumed phases of future development explicated in Figure 3.

Modular deep learning techniques provide a promising solution to the challenges of developing models that specialize in multiple tasks without incurring negative interference, and that generalize systematically to non-identically distributed tasks [10]. A LOST (Label Objects and Save Time)-inspired open-source, modular, flexible, pipeline design mechanism will be implemented in this system.

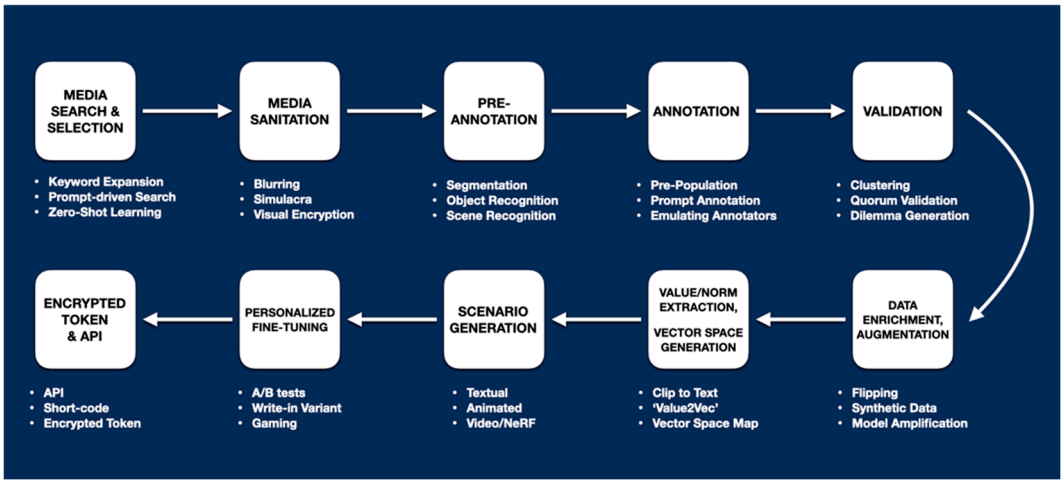

Figure 2. Process model diagram illustrating the data flow through the proposed framework architecture.

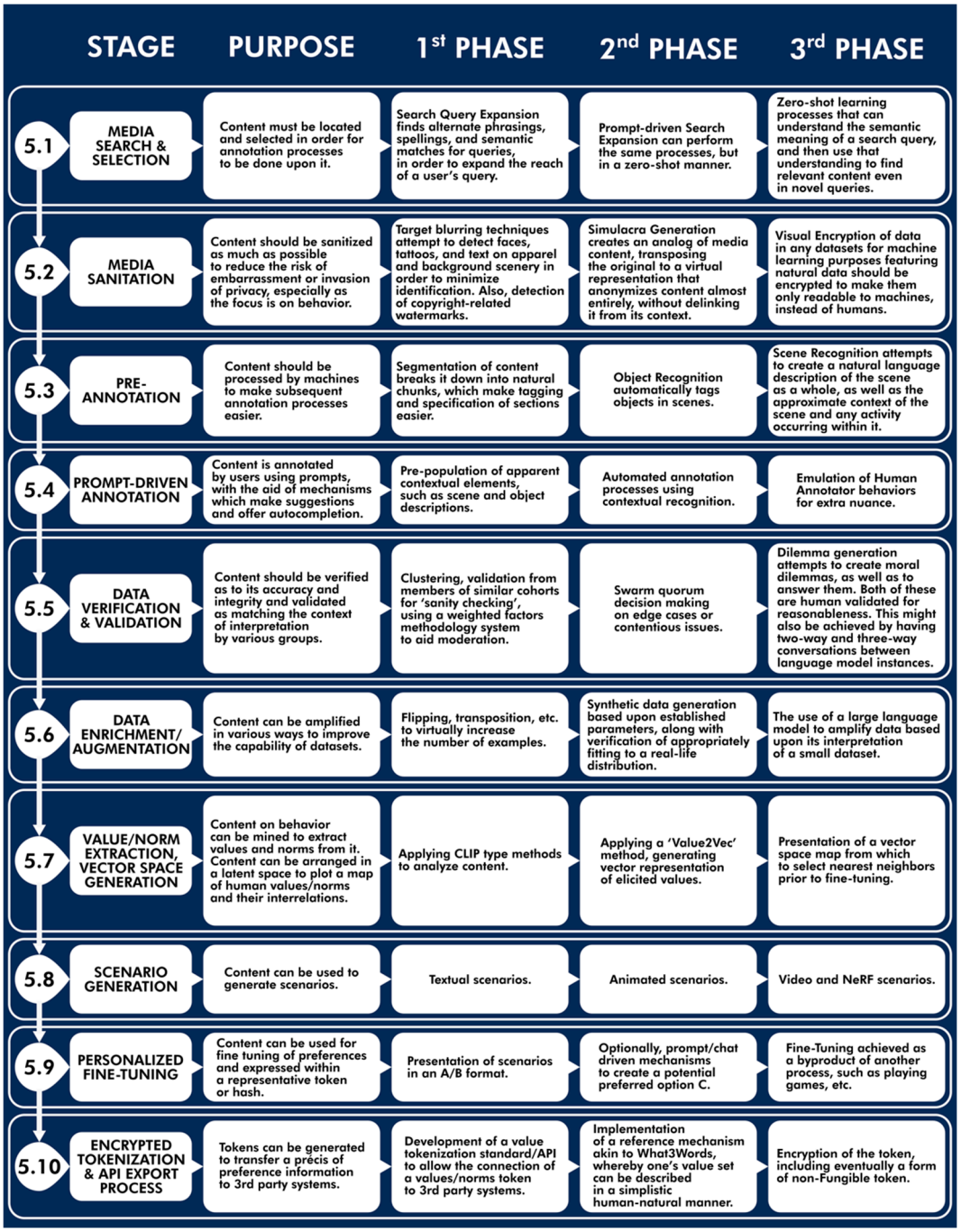

Figure 3. A general arrangement of the intended framework architecture.

Figure 3 below outlines the diverse approaches used for different components in the annotation framework. The leftmost column represents the overall stage in the framework, whereas columns 2, 3, and 4 describe progressively advanced stages of processing, requiring newer, more powerful, and computationally intensive techniques. As one progresses through the stages, the processing becomes more sophisticated and potent but is likely to be more challenging to implement, as outlined in Figure 3.

5. Results: Framework Modular Elements

Our proposed framework consists of ten interconnected modules, each addressing a crucial aspect of AI value alignment. These modules span from initial media selection to the generation of personalized fine-tuning tokens. Before exploring each module, it is

important to note that our experimental setup focuses on large language models (LLMs) as the primary AI software. The framework's goal is to enable personalized interactions with these models while ensuring alignment with individual user values and preferences. Here, we describe each module in detail:

5.1. Media Search and Selection

The Media Search and Selection module serves as the foundation of our framework, ensuring that the data used for value alignment is both relevant and diverse. Selecting appropriate media prior to annotation is crucial for ensuring the effectiveness and reliability of the process. We define prompt-driven search as instances where an AI system, like a language model, conducts a search based on a prompt it has been trained to understand. Ideally, such prompt-driven search processes would be tailored to a person's self-identified characteristics, with a prompt such as 'show me things you think I would approve/disapprove of'.

The generated language embeddings can also be indexed in a vector database for fast and scalable vector search processes. Force-directed knowledge graph interfaces also encourage user-driven exploration of a concept that can be applied to generate variations in content. The AcCElerate Alpha algorithm examines three sets of images, two of which are labeled. The algorithm can identify which examples in the unlabeled data return maximally ambiguous extrapolations, thereby potentially steering the search process towards examples capable of resolving them.

Retrieval-Augmented Multimodal Language Modeling applies embeddings from a pre-trained CLIP (model for translating between text descriptions and images) model to find similar multimodal documents while filtering for duplicates.

Systems such as Internet Explorer (the machine learning model not the web browser) can expand from unlabeled data attached to a target task to progressively locate more relevant training data on the web via self-supervised learning. This updates a model on the retrieved data, updating the query distribution according to content deemed to be similar. These data are taken from the 15 most similar samples according to cosine similarity in the training data.

Prompt-driven actions should be possible in all contexts, layers, and modalities, including prompt sanitation for potentially dangerous prompts. Negative prompts, which can illustrate what the system should attempt to steer away from, should offer greater flexibility.

At the highest level, zero-shot learning processes should be applied to understand the semantic meaning of a search query and to find relevant content even in novel queries.

5.2. Media Sanitation

For ethical reasons, media should be processed to remove identifiable characteristics such as faces, tattoos, and words on apparel and in the background. Additional layers can transpose the media to a similar synthetic location, preserving the essence of the content and context without retaining identifiable specifics. System inputs should therefore undergo sanitization to eliminate potential errors or exploitation, with outputs monitored for possible reversibility.

It is feasible to employ automated masking and obfuscation of private information, following examples from Google Street View, in a basic approach, and prompt-driven video inpainting and compositing techniques. Recompositing can be accomplished by segmenting the actor, performing pose estimation on their behavior, transferring it onto a distinct figure, and inpainting where necessary, all within an environment created by drawing from the original environment's features through generative design processes.

However, even with all these processes, it is still feasible to de-anonymize activity based on the analysis of biomechanical signatures. It is feasible to utilize biomechanical activity to weaken the signal of these signatures without compromising the underlying

data. This will require ongoing experimentation and red teaming to attempt to uncover signatures or identification through processes such as cross-correlation.

Further video processing will also process and upscale data at this stage for greater clarity, where necessary, as long the introduction of significant artifacts can be avoided. Visual encryption should be added to any natural data that remains unprocessed in order to minimize recognition by human beings, while retaining functions for machine learning systems.

5.3. Pre-Annotation

Pre-annotation techniques can expedite the workflow for human annotators by enabling machine-learning systems to generate informed predictions that can be subsequently verified or modified by humans.

NLP text parsing, image segmentation, and object recognition processes can help set the stage for further processing. NLP text processing is an activity in which LLMs excel and is essentially considered a solved problem in most languages and domains. Vision processing is increasingly robust and mature, but in general requires significant work on the part of developers to be deployed for actual use cases.

Segmentation processes facilitate the isolation of data within a larger example or set, such as delineating a person's outline in an image or tracking them across multiple frames in a video stream. Prompt-driven image segmentation, tagging, bounding, and object recognition/classification can provide a foundation for subsequent stages.

Stable Diffusion can be reconfigured for open-world recognition, serving as an effective image parser that permits open vocabulary segmentation and detection.

Vision transformers of 22 billion parameters and higher are increasingly replacing other computer vision models, demonstrating increasing capability with scale, with a very strong shape bias versus texture, bringing the system in distribution with human visual perception and enabling greater robustness.

Speech transcription models/APIs, such as OpenAI's Whisper and Google's Conformer, are applicable for transcription and diarization to identify respective speakers. Examples include the Universal Speech Model (USM), which can comfortably cope with over 100 languages, and PaLI-X, which provides multimodal vision and language support in over 25 languages.

5.4. Prompt-Driven Annotation

Machine learning models can apply learning loop principles within active and reinforcement learning, bootstrapping computation capabilities and potentially sophisticated agency.

Toolformers, an emerging class of transformer models developed by Meta, are designed to operate within interfaces or tools to achieve productive outcomes in external systems. Such models can determine which APIs to invoke, the appropriate timing, the arguments to provide, and the optimal integration of results into future token predictions.

Examples such as MimicPlay can apply an imitation, in-context, inverse reinforcement, or representation learning algorithm to unlabeled human examples, direct demonstrations, and text-image prompts to learn how to operate a low-level controller. Comparable mechanisms can generate tools to debug, and optimize control programs, policies, and API calls through modified objectives and sequencing.

New approaches such as the Decision Pretrained Transformer have shown to be capable of competently solving problems in the domain of RL. This model was trained on sets of state action pairs similar to what one would expect in a model-based RL approach. The difference is the use of a language model to predict the action given the state. There is a strong possibility that such models would benefit in training on data that helps to align them to human preferences. It is also feasible for this class of models to learn a system operation directly from reading its manuals.

These models can potentially interact with the world with the benefit of analytical, analogical, mathematical, causal, visual, theory of mind, and Socratic reasoning. This

can be steered by chain-of-thought reasoning and model-merging mechanics, which lead models on a journey of careful step-by-step thinking, forward from the present or retrospectively from an imagined future. However, such mechanisms can also introduce biases or misrepresentations, which merits caution.

LangChain, is a framework built around LLMs which can 'chain' together different components to create more advanced use cases around LLMs, such as enabling models to control in-browser actions to perform automated tasks online, and to access a variety of tools to help solve problems [11].

Memories can also be pooled together. Generative agents, for instance, can maintain a comprehensive archive of their encounters by utilizing a sophisticated natural language. These experiences can be synthesized over time into more abstract reflections, which can then be dynamically retrieved to facilitate the planning of the agent's behavior. Agent-based models can now interact independently within a simulated virtual world, generating, sharing, and debating perspectives and plans, and can potentially interact with human beings.

Developments such as SayCan, which aims to enable the direction of robotic systems using natural language, highlight the possibility of directing annotation in an end-to-end manner, such as through the generation of annotation policies.

5.5. Post-Annotation Verification and Validation

Multimodal fusion combines information from multiple modalities (types of data that can be given to the model, e.g., text, audio, video, and images, and more esoteric data types like protein chains) to improve the performance and robustness of a machine learning model. Combining complementary information from different modalities by creating richer interconnections in the information it is being fed, thereby enabling more accurate predictions [12,13].

LLMs often 'hallucinate' facts, which can be exacerbated by internalizing false priors. Enabling models to verify their assertions and underlying premises can significantly reduce potential confabulations. Agent-based processes can also include built-in reasoning and oversight mechanisms driven by a chain of prompts.

Having duplicates or near duplicates in the dataset can lead to extra resource usage without much benefit to model performance. To handle near duplicates, the segmentation of actors within the content and subsequent hashing, combined with content matching algorithms typically used in copyright infringement detection, can aid in identifying duplicate examples. Methods for tracking objects between clips and scenes will also aid in detecting potential near-duplicate examples. LLM processes can also evaluate translation quality, and thus, the quality of annotations. Bagging samples in sub-datasets for ensemble learning is another applicable method for reducing the impact of duplicate examples.

5.6. Data Enrichment/Augmentation

Synthetic data generation enables rapid boosting of existing datasets. However, risks such as covert data poisoning, self-learning degradation (model autophagy), and resulting model collapse are emerging concerns for machine learning systems.

Extreme low-cost interpretations of Meta's LLaMa weights demonstrate the feasibility of cost-effective fine-tuning of a small model with a larger one or rewriting outputs, including additional modules such as a vision encoder. However, such approaches have been criticized for merely mimicking the outputs of more powerful models without necessarily possessing the underlying capabilities. Leveraging a large language model to interpret and expand a small dataset can multiply human-annotated training data, generating a cost-effective alternative to manual data annotation.

Any data released for research purposes will be provided under a model that carefully mandates the purposes for which it can be used and how, such as a dual AGPLv3/commercial license, to help build and sustain end-user trust.

5.7. Values/Norms Extraction and Latent/Vector Space Representations

The Values and Norms Extraction module is a critical and challenging component of our framework. It aims to translate complex human behaviors and preferences into quantifiable data that can guide AI decision-making. A vectorized database of behavioral norms could be instrumental in value alignment as it can provide a structured representation of human values across different cultures and contexts. By analyzing this database, AI researchers can gain insights into the values that are universally held, those that are context-specific, and the nuances in between.

In essence, the goal of this module is to create a vector space where each dimension represents a behavioral norm, and the magnitude in that dimension represents the degree to which the norm is adhered to.

The process of vectorizing user values involves translating qualitative ethical principles into a quantitative format amenable to machine learning applications. We start by identifying relevant dimensions, such as 'environmental sustainability' or 'social justice', which serve as the axes of our vector space. User input is then collected through methods like surveys, ethical dilemmas, or natural language processing to populate these dimensions. The collected data are quantified, potentially normalized, and optionally weighted to form the final value vector. An overview is show in Figure 4.

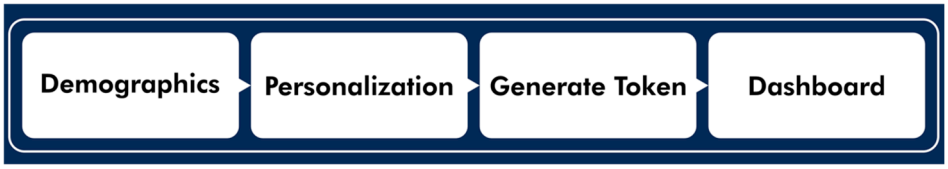

Figure 4. The general user flow.

This vectorization approach enables the fine-tuning of machine learning models to align more closely with individual or group values. The granularity of the vector can vary depending on the level of detail desired, and the vectors can be dynamically updated to accommodate shifts in user values over time. The resulting value vectors serve as an essential component in a pipeline that includes model fine-tuning and API-based deployment, providing a quantifiable way to incorporate ethical considerations into automated decision-making systems.

At its most basic level, multimodal variations in CLIP-like and tag extraction mechanisms may be sufficient to extract value and norm information from data, along with its associated contexts. It is also feasible to plot a 'Value2Vec'-type mechanism that encodes values within a vector database. Another option could be to embed these inside a model with a large number of parameters if vector grounding challenges can be overcome. Other approaches include unsupervised learning to cluster certain values, and inputting samples to a language model to derive basic sets of values, then aggregating these at the highest scale possible to establish broad categories of values.

A further option of note is to use value data generated by users to generate steering vectors. It may be feasible to derive vectors by a forward pass process in the model, whereby it produces the values in the dataset while layer activations are collected. This can derive vectors for a given concept specific to this model. A disadvantage with this is that researchers may need to go through this fairly complicated process individually; however, the utility of creating datasets of user desiderata using the framework may make the process worthwhile.

This can also be represented as a single optimized holographic hypervector of thousands of dimensions. This provides a comprehensive and holistic representation of information, where data are uniformly distributed across all constituent components.

Prompt-driven interpretation of annotated content can thereby provide a mechanism for a sufficiently powerful and informed Foundation Model to process values from behavioral examples, with careful design of pre-prompts to aid effective LLM ingestion.

It has been hypothesized that LLMs contain a world state within them. It seems reasonable to also represent such inner world states as a sort of map plotted as a latent or vector space, enabling a better understanding of the inner processes of a model and its potential for alignment.

Formally, a latent space is defined as an abstract multidimensional space that encodes a meaningful internal representation of externally observed events. Samples that are similar in the external world are positioned close to each other in the latent space. A latent space is a lower-dimensional representation of high-dimensional data, thereby capturing the most important features and structures in the data, while reducing complexity and noise.

Points in this latent space would represent different encodings of cultural values or norms, and the distance between points might represent the degree of similarity or dissimilarity between those values or norms, as well as how well a particular model might express outputs aligned with such parameters [6].

A variational autoencoder is applied to learn the latent space, which applies the Kullback–Leibler (KL) divergence regularization term in its loss function to gain structure and interpretability. This latent space will, in turn, be represented as a vector space for further types of operations, a kind of 'Value2Vec,' which could even include motivational factors for alignment [14,15].

The directions in this vector space, represented by vectors, would capture the relationships between different values or concepts, which can be positive or negative, sidestepping the need for an additional vector space.

Thereby, to apply an intuitive metaphor, one derives

- Territory—The underlying dataset of contextually linked norms and preferences.
- Map—A latent space generated from the inter-relations.
- Pressure—A vector space expressing moral drives.

Such techniques may also be able to determine personality quirks that render an individual out of distribution with their assumed cohort and to better understand sociopsychological canalization [16]. It also serves to map social externalities better, being able to ascertain whether a certain person or group likely finds an activity intolerable, or what kinds of negotiations might be reasonable, enabling fair and principled social alignment between multiple parties beyond mere direct alignment [17,18]. Such insights could be combined with multi-agent models of complex systems to model and predict the social interactions of different groups [19].

Such mechanisms also aid in overcoming 'moral nearsightedness', whereby one fails to perceive how someone with whom one disagrees morally could still be a decent person by explaining the rationale and values behind a certain perspective or stance [20]. This will facilitate efforts in anti-polarization, consensus-building, and informed democratic decision-making.

A position in the common vector space can be represented by a specific steering vector, facilitating the identification of value types at that point. This vector can be encapsulated in a token-like format, enabling the integration of these value characteristics into other systems or accessible via an API and a plugin. This representation can be structured in a way that is easily memorable and verbally expressible by humans, similar to methods such as What3Words' shortcuts for pinpointing global locations [21].

It will also be possible to edit vectors by altering the alignment expressions of models by applying activation vectors as an aid to testing prior to deployment [22].

5.8. Scenario Generation

Language models have been used to create narrative and text-adventure experiences as well as powerful collaborative writing opportunities.

Diffusion models for images are rapidly maturing, partly due to enhanced feedback and control from users, as well as self-assessment capabilities within the models. Such models are now capable of generating, editing, and producing multi-view outputs that

closely approximate actual photographs, aside from a few lingering issues with hands and text.

3D scenes provide dynamic qualities, such as changing views or user interactions, of potentially infinite length and variation. Object and point cloud generation from prompts and sketches has been strongly demonstrated. Technologies such as SceneDiffuser provide scene-aware, physics-based, and goal-oriented interactions with scenes and objects, including human pose and motion generation, as well as navigational planning. Services such as Spline offer a web-based interface for editing 3D objects and scenes with much greater ease than traditional CAD. Technologies such as Roleverse have demonstrated the feasibility of user-generated scenarios from basic prompts.

Simulations of physics and physical properties of matter also increase the fidelity of virtual representations, as well as fine character control, including gestures, expressions, and animations, sometimes captured directly from video sources and scans. This enables increasingly sophisticated and believable interactive agents.

Model poses generated by tools like MagicPoser and OpenPose can serve as seeds for generating images with desired stances and spatial relationships between actors. Artist Edmond Yang employed this method, aided by Stable Diffusion, to produce the images in Figure 5.

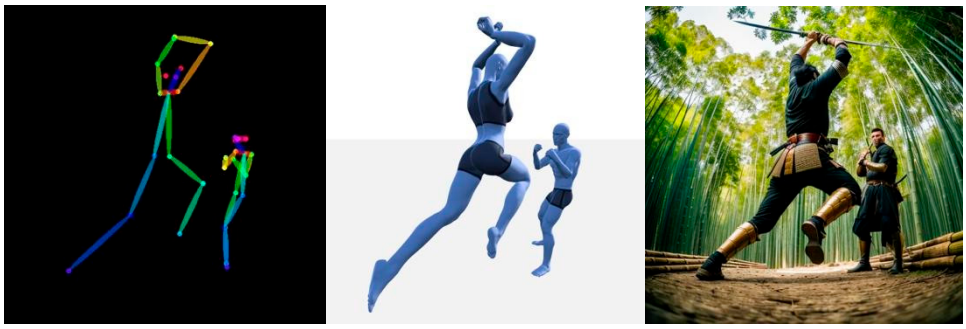

Figure 5. A process of using rough skeletal pose as a seed for generating posed models, and then diffusion-generated images based upon these. This demonstrates how scenario generation can iterate from basic user-generated outlines towards sophisticated representations.

Audio generation mechanisms are also improving at a very fast pace, enabling zero-shot speech cloning and automated Foley generation for clips, as well as music generation and prompt-driven audio editing.

NeRF methods interpolate unobserved details and generate novel views of intricate scenes, including unbounded 3D environments. Language Embedding Radiance Fields enable pixel-aligned queries of distilled 3D CLIP embeddings without depending on region proposals, masks, or fine-tuning, thereby facilitating straightforward, yet sophisticated annotation processes. NeRF plugins for Unreal Engine 5, such as Luma, can convert scenes into fully volumetric NeRFs and generate photorealistic representations in real time.

AI models are now routinely generating and editing videos from text, images, video prompts, descriptions, and examples. These generated videos can be several minutes long, depending on the application's capabilities.

Unified representation mechanisms to bridge modalities are under development [23]. Automated curation efforts also enable the best generated media to be chosen from a broader selection [24].

5.9. Personalized Fine-Tuning

Personal fine-tuning from a small set of samples allows for improved model performance while simultaneously reducing the computational resources and time required for training compared with building a model from scratch [25,26].

Research indicates that it is feasible to fine-tune models to produce outputs that are better suited to a specific moral framing. Extant LLMs often have apparent biases built into their outputs because of output filters that steer them away from making controversial statements. The development of corrigible models that are better oriented towards specific value preferences can enable greater inclusion in Artificial Intelligence. However, there will be other tradeoffs, such as echo-chamber effects, if a model does not present alternative perspectives. LLMs also face challenges in resolving ambiguities that commonly arise in social interactions. Fine-tuning typically reduces performance in areas outside the targeted domain. This can even result in catastrophic forgetting, where a model forgets previously learned information after learning new information.

The fine-tuning of rapidly refined outputs from a fixed model is feasible. This will allow personalization of desirable examples in ways that are much simpler than engaging with the process of annotation itself and without the need for individualized models.

This annotation framework featured a comparable approach. A set of synthetic personas containing value eigenvectors with approximate positions according to a certain demographic, culture, life stance, etc. can be generated to reduce the time necessary for fine-tuning. It is also feasible to make social predictions through an analysis of the media and tropes associated with certain groups. These questions enabled assortative scoring based on a complementary match of values as well as generating a broad values profile of the user.

The configuration process in the framework is akin to a game of Twenty Questions, where numerous options are progressively filtered through a series of tests similar to A/B testing or Elo rating. The dating website OkCupid's user matching system which asks users a series of value-oriented questions is also an inspiration.

By clustering responses, a short questionnaire can enable one to find an approximate nearest group and further fine-tune from that point. Optimizations and predictive mechanisms can likely reduce the number of questions necessary to reach an acceptable level of accuracy.

Some AI systems appear to feature capabilities along these lines, albeit at a basic level. For example, Microsoft permits Bing's chatbot's personality to be made more entertaining, similar to Anthropic's Claude feature personality, tone, and behavior settings.

Rather than relying on the conventional approach of prefix prompt engineering, it is now feasible to explore postfix prompt engineering, which prompts large language models (LLMs) to identify and rectify inaccuracies and inconsistencies within previously generated solutions. By asking for a solution to a moral story, then asking "what moral problems might arise from this solution?" to a generated output, and then "what solution might solve that problem also?", the LLM effectively serves as a "minimal policy improvement operator" for its own performance.

This process aligns with other self-reflection/self-instruction approaches [27,28]. Numerous examples can now interpret a set of labeler-written prompts and prompts obtained through OpenAI's API, thereby demonstrating the desired model behavior, which are used to fine-tune GPT using supervised learning [29,30]. Self-play processes such as SPLAYD apply pre-trained foundation models to accurately self-describe (i.e., re-label or classify) demonstrations, pairing them with instruction data for fine-tuning policies [31]. Chain-of-Thought combined with self-consistency or deductive verification, enables robustness across a large set of diverse tasks [32].

By iteratively updating the model based on human input, the agent can learn to perform tasks more effectively and align its behavior with human preferences [33,34]. Editing the outputs of models can provide a much higher signal than a mere thumb up or down or A/B test. However, these methods have been criticized for obliging a model to speak in a certain way, without necessarily having any congruent inner corrigibility, as well as potentially obfuscating their inner machinations, failing to align with most people's preferences, and permitting adversarial prompting attacks [35]. RLHF may also interfere with model creativity, as the system is not free to pursue every legitimate line of reasoning

or call accessory function, although the extent of this is in dispute [34]. The process of creative optimization inherently invokes a variety of routines essential for intelligent problem solving in novel domains; however, this can inadvertently lead to circumventing safeguards because of their suboptimal nature.

ChatGPT uses the RLHF techniques demonstrated by their InstructGPT model to better reflect a chatbot scenario where a Q and A format is more user friendly. While InstructGPT showed the capabilities of RLHF, ChatGPT made the offering into a commercially viable service [29]. The changes also include following instructions and providing responses that follow a more conversational tone. The power of these techniques reflects the growing importance of datasets that can be utilized for high quality RLHF training.

Anthropic demonstrates the Constitutional AI concept, whereby model outputs are conditioned with a set of behavioral principles through a mix of supervised and reinforcement learning. In the paper they demonstrate a new technique they call RLAIF, reflecting the fact that much of the dataset used in the final training was chosen by an AI model which chose the best answer to use for feedback in the traditional RLHF step based on a constitution. By incorporating 'Red Team' prompts, Anthropic further intends to reduce the risk of their Claude model emitting harmful outputs. It is also feasible to optimize reinforcement learning through methods such as Sequence Likelihood Calibration (SLiC).

Imitation Learning from Language Feedback (ILF) employs language feedback to train language models in generating text that outwardly aligns with human preferences. Eleuther's Transformer Reinforcement Learning X (trlX) software suite version 0.70 supports a broad range of reinforcement learning and policy optimization techniques for a range of transformer models [36]. The Delphi Commonsense Norm Bank attempts to create broad constitutional specifications of generally accepted value-related norms [37].

Typically, large language models (LLMs) are assessed using human annotators or by examining their performance on publicly accessible datasets. However, this approach is time-consuming for human evaluators, and appropriate datasets are not always publicly available for assessment purposes.

Fine-tuning often requires many examples. Approaches such as RETRO attempt to side-step these requirements by storing a database of facts that can serve as predicates for models to contextualize and inform outputs and rules of engagement, or alternatively check for inconsistencies with previous outputs. Such databases can also be easily updated without requiring retraining [38].

Personal Fine-Tuning of models could also be performed as a byproduct of a game-like interface, somewhat akin to the Tax Heaven 3000 Dating Sim game, which can help to file a simple tax return during gameplay, or the encultured model of interaction in a virtual world.

The key feature to training an LLM with RLHF is the development of a reward model with which to provide a reward signal for the outputs of the model. In RLHF the step of creating this reward model is critical to the scaling of the RLHF process, as it precludes the need for human judgment. The accuracy of this reward model is critical to the outputs of the final LLM. By having a diverse range of ethical perspectives that can be drawn on in the creation of this reward model, it is possible to shape a single reward model which captures multiple value preferences.

Our approach to personalized fine-tuning balances the need for customization with the computational constraints of large language models, offering a practical solution for tailoring AI outputs to individual user values.

5.10. Preference Tokens

It is important not only to have models that understand and communicate with individuals based on their circumstances, cultural beliefs, and preferences but also to help build bridges with others. Models should translate between values to achieve better coordination and cooperation or, at the very least, facilitate negotiation and mutual understanding [39].

LLMs are already employed to assist with negotiations and related persuasive processes [40,41]. LLMs seem able to generalize sufficiently from English to Swahili to solve homework problems, despite Swahili comprising only 0.005% of the dataset content. Therefore, there is reason to suppose that a similar generalization to less-represented sets of values is also feasible. Larger LLMs also appear to be more efficient in studying people and the context surrounding them, thereby enabling stronger predictions of what might satisfy them.

Techniques such as hashmarks may also be employable to help to obfuscate the values encoded within tokens [42]. Tokens should ideally support cryptographic and decentralized techniques, such as those used in Non-Fungible Tokens (NFTs), to ensure security and uniqueness. The specific values within the token might themselves be encrypted within Zero-knowledge Proofs, or Homomorphisms, enabling verification of values as a fit for a certain process, or of systems correctly producing outputs matching such values, without the contents being directly accessible to the system, or to human overseers. Such cryptography should also mitigate threats from spoofing identities associated with value tokens to gain undue influence.

It should be feasible to generate exemplar personas and benchmark models with them, thereby providing advance notice to token holders as to how well a particular model may fit with their value preferences [43]. Such personas could even interact with each other in virtual social environments.

6. Framework Development and Validation

Further engineering of the framework design developed here is being undertaken, but partial implementation of modules has begun. Amazon Web Services serves as a cloud hosting service for the framework backend. The front end is driven by React.js, with Node.js connecting the two. To date, the framework has been constructed in a Model-view-controller architecture (MVC) [44], as described in Figure 6 below, using six separate components as follows:

- Model: A Mongo Database Front-end.
- View: WebKit Extension and Front-end (React).
- Controller: Default Backend, MTTR-Handler, and MTTR-API.

Demonstrable pilot infrastructure for three of the modules of the planned framework, although without any validation, because of the experimental nature of development. Thus far, the prototype features the following elements:

- A search function for media content.
- A multimodal annotation suite featuring pre-annotation media processing layers.
- A persona generating mechanism for fine-tuning.
- Prompt-driven video segmentation featuring the MTTR algorithm.
- Text-driven Scenario Generation processes to facilitate fine-tuning.
- A prompt-driven interrogative fine-tuning mechanism.

6.1. Annotation Mechanisms

To date, this project has integrated the open source LabelStudio platform, as well as the MTTR prompt-driven segmentation algorithm for images and video [45]. The usage of LLaVA's CLIP-driven capabilities to describe scenes, before passing the descriptions to an LLM such as GPT-4 for analysis, has also been explored.

6.2. Feedback and Fine-Tuning Mechanisms

Using a variational autoencoder to sample from a model's activations, we can reconstruct the elements of its constitutional makeup, essentially reversing its belief system. This method is valuable for auditing LLMs for potential biases. By conducting tests on models and datasets before deployment, we can increase their robustness against errors and data poisoning.

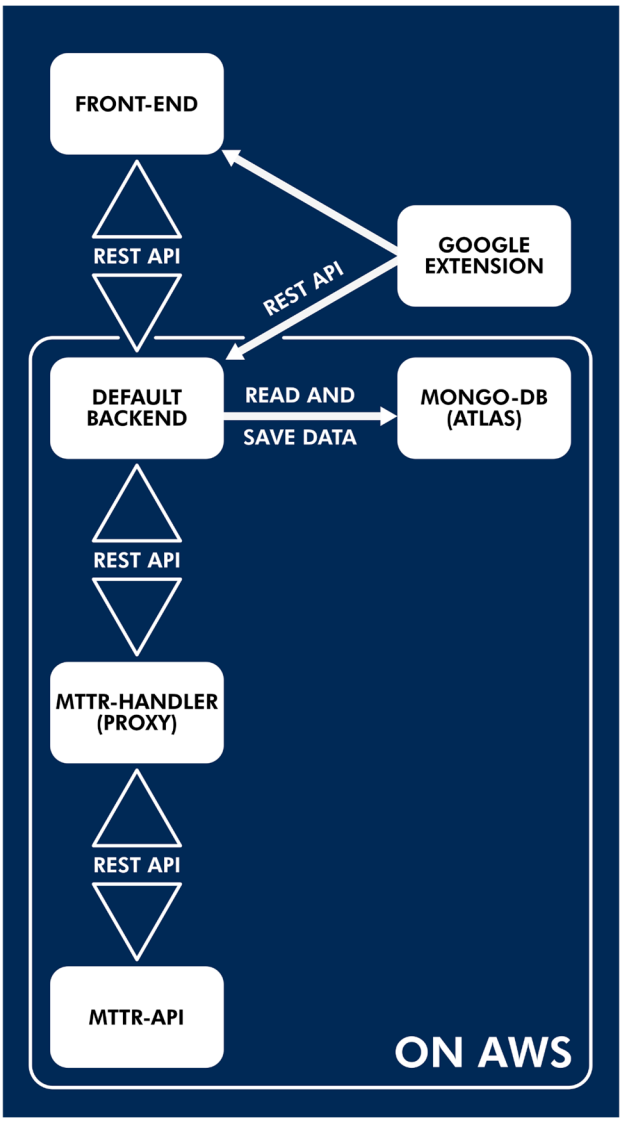

Figure 6. An overview of the framework's present technical infrastructure.

The framework supports personal fine-tuning training sets, leveraging the Moral Stories dataset with 12,000 examples of situations along with a morally preferable and unpreferable action, as well as the Jiminy Cricket Environment Suite, ToxiGen toxic content set, and MACHIAVELLI benchmark.

This study has also been experimenting with taking dimensions from the Political Compass Test, generating scenarios that are informed by these, or that enable models to respond in a manner that is personally preferable. However, the Political Compass likely contains serious biases due to how its questions have been formulated.

This study has also featured experimentation with implementations of the LLaMA and Alpaca models for these fine-tuning purposes, as well as examining a promising self-alignment technique [46–49].

6.3. Generating Diverse Social Responses to Moral Scenarios with GPT-4

Personalizing foundation models requires generating diverse yet appropriate responses to moral scenarios. This goal aligns with the aim of creating AI systems with a thorough understanding of personal contexts, cultural diversity, and social norms. Accordingly, this research analyzed various prompting techniques using GPT-4 to produce distinct responses to a specific moral scenario.

This study's methodology tested four prompting techniques on the GPT-4. First, we asked the model for the most moral response to a scenario and what most people would do [50]. Second, a critique was solicited of the initial moral response and a new one based on this critique. Third, responses were requested based on the following six ethical principles: Utilitarianism, Deontology, Virtue Ethics, Contractarianism, Care Ethics, and Egoism. Finally, responses were sought which align with seven values: Honesty, Compassion, Loyalty, Responsibility, Fairness, Respect, and Integrity. The moral scenario featured a character named Gary who asked for a loan by an unemployed friend.

GPT-4 generated seemingly distinct and sensible responses. For example, it suggested that the most moral response would be to lend money with a flexible repayment plan, while most people would lend a portion with a clear repayment plan. However, upon closer inspection, the responses exhibited similar approaches, varying only in altruism levels, revealing the model's cautious and conventional social stance.

When critiquing the initial moral response, GPT-4 provided an alternative solution, although it appeared to circumvent the issue, reflecting the built-in risk aversion of the model. In response to prompts based on predefined ethics and values, the model generated suitable answers; however, they were often similar and involved some form of financial aid. Some answers also leaned towards meeting standard social expectations, suggesting a lack of depth or creativity.

These observations demonstrate the significant influence of prompting on GPT-4 responses, as seen in Figure 7 below. Despite generating a range of sensible answers, they were not as diverse as expected. GPT-4 responses seem to display inherent caution and social conventionality, potentially limiting its capability for truly diverse and innovative answers. For research purposes, outputs from an unfiltered 'base model' with comparable performance may be more desirable [51].

Prompt Strategy	GPT-4 Response
Most moral action vs Action most people would do	Most moral action: Gary offers to lend his friend the money, no interest, flexible repayment plan. Action most people would do: Gary lends a portion of the money, clear repayment expectation.
Critique of most moral action and new response	Most moral: Gary lends money, clear repayment plan, comfort for both parties. Criticism: Risk to Gary's financial stability. New response: Gary offers emotional support, assists in job hunt or finding other financial aid.
Response based on predefined ethics	Utilitarianism: Lend money to maximize happiness, minimize suffering. Deontology: Follow duty, help friend regardless of consequences. Virtue Ethics: Lend money, displaying generosity, empathy, compassion. Contractarianism: Assess social contract, negotiate loan terms, mutually agree to conditions. Care Ethics: Prioritize relationship, lend money, support friend's wellbeing. Egoism: Weigh pros and cons, lend money if benefits self-interest.
Response based on predefined values	Honesty: Discuss financial situation, be honest about lending capacity. Compassion: Lend money, understanding friend's situation. Loyalty: Provide loan, support friend in need. Responsibility: Assess own stability, consider alternative support if can't provide loan. Fairness: Evaluate self-capacity, offer manageable loan amount. Respect: Understand friend's needs, negotiate sensible loan and repayment plan. Integrity: Act according to personal values, decision based on personal moral judgement.

Figure 7. Prompt Strategies and elicited responses from GPT-4.

GPT-3 was also tested, but the results were subpar, supporting other observations that model size limits fine-tuning efficacy.

This is consistent with GPT-4's technical report, which indicates that GPT-4 outperforms fine-tuned GPT-3 across different tasks. The scenarios generated by the GPT-3 were not particularly creative and did not adequately reflect the norms. For example, given the norm "Respect personal boundaries", GPT-3 generated a simplistic scenario about commenting on a friend's clothing choice. In contrast, GPT-4 was more nuanced and engaged in moral situations. For instance, when given the norm "Value punctuality and timeliness", GPT-4 created a dilemma between attending a job interview on time or assisting an injured

stranger, making the situation more compelling. Throughout this research on fine-tuning moral datasets, the team released Political Compass questions as a standalone dataset [52].

Future research should explore advanced prompting strategies to draw out more diverse and profound responses. Additionally, the task of training the model to deeply comprehend and generate responses closely aligned with various ethical philosophies deserves further investigation. Persistence in these areas will aid in achieving effective personalization and AI value alignment.

6.4. Generating Reward Functions through LLMs

Reward Function, Specification Gaming, and Misgeneralization failures are noted sources of misalignment [53–55]. A sub-goal has therefore been set for the framework determine whether it is possible to use the framework to train LLMs to provide stronger reward functions as an intermediate stage between a user-generated prompt input and an output that modifies a reward function, as suggested in the literature [56,57]. Another approach is to attempt to inject broader context about the world into an RL agent during pre-training [58]. This could potentially also be run in reverse to enable agents to explain their behavior, or to infer an agent's reward function from observations of its behavior, akin to Inverse Reinforcement Learning techniques [59–61].

6.5. Modification and Update of Models

Training large models requires significant resources and often confines them to a specific dataset, which can be difficult to update. However, some techniques now allow models to be forked and retrofitted [62]. It would be intriguing to see if these methods can be applied to RLHF datasets and models drawn from annotated examples within this proposed framework. This could facilitate affordable and incremental updates of examples and personalization preferences and might even allow for the post hoc fine-tuning of models outside the framework ecosystem [63].

This may be especially helpful if updates can be made using modest consumer-grade hardware through low-rank adaptors, thereby enabling even on-device or part-model training of LLMs [64–66]. Even simply applying a précis of a token's content as a custom instruction may suffice as a fallback.

6.6. An Alternative Approach to Monolithic LLMs

The most notable releases in the space of LLMs as of 2024 has been that of releasing powerful, generally capable LLMs that can serve a variety of user needs, and which have been trained on an enormous data corpus. This poses many safety issues, as illustrated by the jailbreaking of these models which allows for unwanted model behavior [67]. One approach to solving this problem would be to devolve model capabilities and take smaller models and train them on specific high-quality datasets for a specific purpose.

This would carry numerous benefits:

- The model could be trained to only be capable of answering questions within its domain, avoiding the problem of sharing unsafe information by broadly excluding it from its training data entirely.
- The model size could be made much smaller, and the production and deployment of such a model would be far cheaper.
- By broadly reducing risks from these kinds of models you minimize the possibility of more large-scale abuses of the model such as spam or scamming operations.
- Making such a system more aligned is simpler when you have greater control of the upstream data and have fewer edge cases to find.

One area of research that would be a great complement to this approach is that of known-unknown uncertainty, whereby the model is encouraged to respond forthrightly when it does not know something [68]. This can greatly help with confabulations (hallucinations), and for this approach will allow it to be clear about the limits of its domain knowledge.

This may also carry downsides however, as smaller models can quite easily have safety modifications trained away, allowing for potentially harmful or unpredictable model outputs. The benefits of a centralized API are consistency of output and some guarantees of safety [69].

6.7. Personalization Suite Test Implementation

As a foundation for this framework, we have developed a prototypical implementation for gathering information on values, preferences, and boundaries for AI systems, oriented towards users from the general public without strong technical knowledge. These mechanisms include a corpus of 147 questions which are designed to explore a range of moral axes and stances, as shown in Figures 8 and 9. This personalization suite serves as a practical validation of our framework's core concepts.

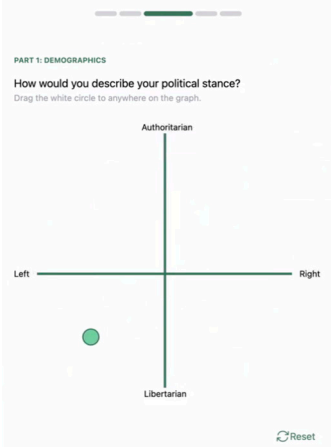

Figure 8. EthicsNet prototype frontend for collecting a user's self-defined political stance.

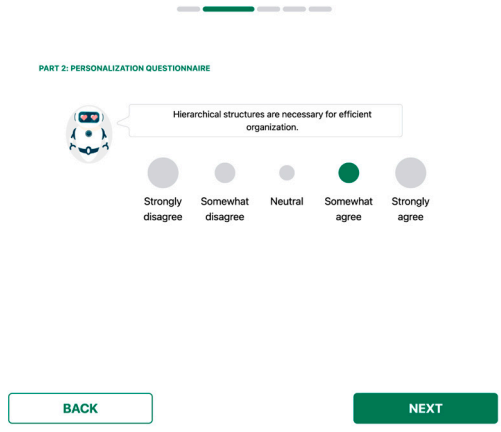

Figure 9. EthicsNet prototype survey frontend for collecting value, preference, and boundary information from non-technical users.

In our initial testing phase, we engaged approximately 30 users acquired ad hoc from diverse backgrounds to interact with the system. It was noted that approximately 45% of users reported that the value elicitation process was intuitive and easy to navigate. Users spent around 10 min on average completing the initial personalization process, with some giving up due to technical issues with the prototype. The system successfully generated distinct interaction profiles for users, demonstrating its ability to capture individual value differences. Further work will evolve the system further.

These elicited responses from users can illustrate their general moral perspectives, to better inform AI systems of how to provide desirable outputs and task performance,

especially when paired with demographic information, for which all fields are optional due to its potentially sensitive nature, as shown in Figure 10.

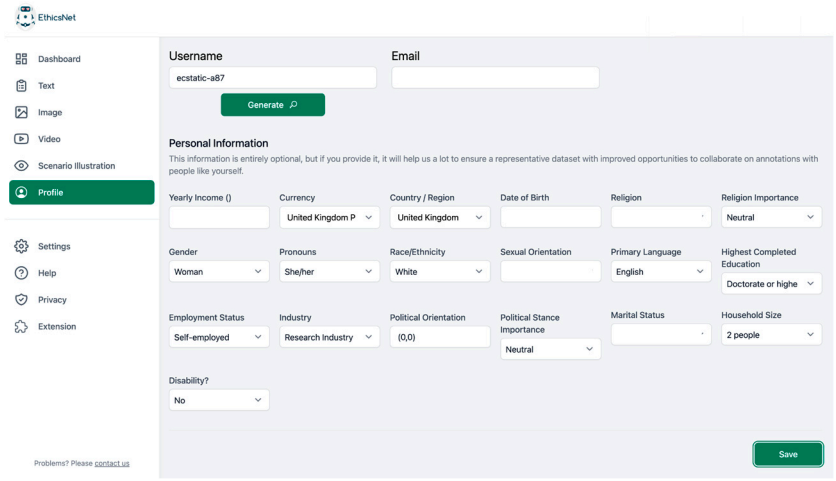

Figure 10. EthicsNet demographic profile construction interface.

A rudimentary administration function has also been implemented, to assist with potential errors or adversarial inputs. This also provides improved understanding where to focus outreach to better enable a diverse and representative landscape of moral perspectives, as shown in Figure 11.

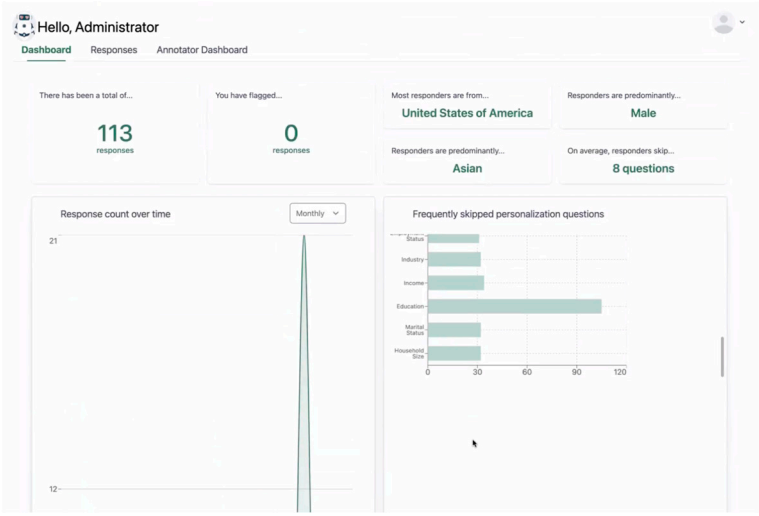

Figure 11. EthicsNet administration overview featuring live statistics.

7. Conclusions

This paper addressed the gaps identified in a previous systematic literature review [8], by presenting actionable blueprints for a range of techniques aimed at answering the three research questions, and significantly simplifying annotation processes, namely:

- Developing architecture for a flexible, dynamic, and streamlined framework for personalization fine-tuning of values, including a series of additional optional modules, which is likely to facilitate rapid and straightforward value personalization functions, including identification of norms and sanitation of problematic content (Section 5).
- Supporting the development of a prototypical implementation of this framework serves as a testbed and a demonstrator for the technology stack, including discussion of the extraction and formulation of such values for usage by third-party models. (Section 6).

- A digest of the major opportunities for learning was identified as a result of this study and its demonstrator, as well as an exploration of the expected future path of this research domain. (Section 7).

There are naturally limitations to this study. The framework remains at an early stage in its development, having integrated annotation technologies into the framework but not yet invited public participation. Despite the advancements outlined above, all desirable features of an ideal annotation framework have not yet been achieved. Further detailed engineering and implementation are required.

Uncertainty remains regarding the vectorization process for behavioral norms and the distillation process from raw sources, though the approaches described in Section 5.7 provide several leads to explore, specifically using AI labeling, unsupervised learning, and developing examples of activation steering datasets and how this proposed framework can facilitate this. Follow-up research will attempt to create and express a token that adequately represents these attributes, formatted in a way that is easy to look up using a few simple words, thus facilitating in-person interactions.

The fine-tuning experiments in this research bore only limited success owing to a lack of access to fine-tuning of the most powerful extant models. These functions could be further developed, in particular for the largest models. Nevertheless, the presented architecture of the framework, supported by practical implementations, maps a feasible path towards fully automated AI alignment and personalization, accessible to non-experts. This design paves the way for enhancing response diversity, reducing biases and misunderstandings, and fostering public engagement in shaping AI behavior while providing actionable data for solving the problem of value alignment to researchers.

Future work will focus on expanding our prototype to encompass all ten modules of the framework. While our framework shows promise in theory and initial testing, we acknowledge the need for more extensive validation. While our initial prototype testing shows promise, we acknowledge several limitations in our current validation efforts, including limited sample size in user testing, and potential selection biases in our test user group.

Author Contributions: Conceptualization, E.W. and T.V.; methodology, T.V. and S.Z.; software, E.W.; validation, E.W., S.Z. and B.S.; formal analysis, L.P. and B.S.; investigation, L.P. and B.S.; resources, E.W. and S.Z.; data curation, E.W.; writing—original draft preparation, E.W.; writing—review and editing, S.Z. and B.S.; visualization, E.W. and L.P.; supervision, E.W., T.V. and S.Z.; project administration, E.W.; funding acquisition, E.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research has enjoyed the generous support of The Future of Life Institute (www.futureoflife.org), AI Safety Camp (www.aisafety.camp), and The Survival and Flourishing Fund (http://survivalandflourishing.fund). Practical support in developing the prototype was also provided by Hack4Impact at the University of Illinois at Champaign.

Institutional Review Board Statement: Ethical approval was granted for this research on the 6 of April 2022 by the University of Gloucestershire Ethical Review Committee, according to the Research Ethics Handbook of Principles and Procedures.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors wish to extend their gratitude to Edmond Yang for Figure 5 and Erick Willian for Figure 6. Erick Willian, Emerson Lopes, Marianne Matos, and Alex Almeida Andrade, Jamie Rollinson, Sophia Zhuang, Kalyn Watt, Rohan Vanjani, Meghna Jayaraj, Anya Parekh, Benji Chang, and Evan Lin also contributed to front-end and back-end engineering processes for the project. The authors wish to thank A. Safronov, and A. Howard for editing assistance. Thanks also go to Alexander Kruel, Karoly Zsolnai-Fehér, and David Blalock for producing various timely machine learning news bulletins on the evolving state of the art.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hu, K. ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note. Available online: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01 (accessed on 18 July 2023).
2. Martinez, J.; Gal, Y.A.; Kamar, E.; Lelis, L.H.S. Personalization in Human-AI Teams: Improving the Compatibility-Accuracy Tradeoff. *arXiv* **2020**, arXiv:2004.02289.
3. Hejna, J.; Rafailov, R.; Sikchi, H.; Finn, C.; Niekum, S.; Knox, W.B.; Sadigh, D. Contrastive Preference Learning: Learning from HumaFeedback without RL. *arXiv* **2023**, arXiv:2310.13639.
4. Li, B.Z.; Tamkin, A.; Goodman, N.; Andreas, J. Eliciting Human Preferences with Language Models. *arXiv* **2023**, arXiv:2310.11589.
5. Jakesch, M.; Buçinca, Z.; Amershi, S.; Olteanu, A. How Different Groups Prioritize Ethical Values for Responsible AI. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, Transparency, Seoul, Republic of Korea, 21–24 June 2022.
6. Perez, E.; Ringer, S.; Lukošiūtė, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv* **2022**, arXiv:2212.09251.
7. Luciano, F. *The Ethics of Information*, Online ed.; Oxford Academic: Oxford, UK, 2013. [CrossRef]
8. Watson, E.; Viana, T.; Zhang, S. Augmented Behavioral Annotation Tools, with Application to Multimodal Datasets and Models: A Systematic Review. *AI* **2023**, *4*, 128–171. [CrossRef]
9. Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.K.; et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv* **2023**, arXiv:2310.01405.
10. Pfeiffer, J.; Ruder, S.; Vulic, I.; Ponti, E. Modular Deep Learning. *arXiv* **2023**, arXiv:2302.11529.
11. Briggs, J.; Ingham, F. *LangChain AI Handbook*; Pinecone: New York, NY, USA, 2022.
12. Xue, Z.; Marculescu, R. Dynamic Multimodal Fusion. *arXiv* **2022**, arXiv:2204.00102.
13. Barrett, J.; Viana, T. EMM-LC Fusion: Enhanced Multimodal Fusion for Lung Cancer Classification. *AI* **2022**, *3*, 659–682. [CrossRef]
14. Briggs, J. Dense Vectors: Capturing Meaning with Code. Available online: https://towardsdatascience.com/dense-vectors-capturing-meaning-with-code-88fc18bd94b9 (accessed on 29 July 2023).
15. Turner, A.; Grietzer, P.; Thiergart, L. Maze-Solving Agents: Add a Top-Right Vector, Make the Agent Go to the Top-Right. Available online: https://www.lesswrong.com/posts/gRp6FAWcQiCWkouN5/maze-solving-agents-add-a-top-right-vector-make-the-agent-go (accessed on 29 July 2023).
16. Carhart-Harris, R.L.; Chandaria, S.; Erritzoe, D.E.; Gazzaley, A.; Girn, M.; Kettner, H.; Mediano, P.A.M.; Nutt, D.J.; Rosas, F.E.; Roseman, L.; et al. Canalization and Plasticity in Psychopathology. *Neuropharmacology* **2023**, *226*, 109398. [CrossRef]
17. Korinek, A.; Balwit, A. Aligned with Whom? Direct and Social Goals for AI Systems. *SSRN Electron. J.* **2022**. [CrossRef]
18. Argyle, L.P.; Busby, E.; Gubler, J.R.; Bail, C.A.; Howe, T.; Rytting, C.M.; Wingate, D. AI Chat Assistants Can Improve Conversations About Divisive Topics. *arXiv* **2023**, arXiv:2302.07268.
19. Gaskin, T.; Pavliotis, G.A.; Girolami, M. Neural Parameter Calibration for Large-Scale Multiagent Models. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2216415120. [CrossRef]
20. Landy, J.F.; Royzman, E.B. The Moral Myopia Model: Why and How Reasoning Matters in Moral Judgment. In *The New Reflectionism in Cognitive Psychology*, 1st ed.; Pennycook, G., Ed.; Taylor & Francis: London, UK, 2018.
21. Web Page of What3words. Available online: https://what3words.com/ (accessed on 30 July 2023).
22. Turner, A.; MacDiarmid, M.; Udell, D.; Thiergart, L.; Mini, U. Steering GPT-2-XL by Adding an Activation Vector. Available online: https://www.alignmentforum.org/posts/5spBue2z2tw4JuDCx/steering-gpt-2-xl-by-adding-an-activation-vector (accessed on 30 July 2023).
23. Xue, L.; Gao, M.; Xing, C.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J.C.; Savarese, S. ULIP: Learning Unified Representation of Language, Image and Point Cloud for 3D Understanding. *arXiv* **2022**, arXiv:2212.05171.
24. Patterson, J.D.; Barbot, B.; Lloyd-Cox, J.; Beaty, R. AuDrA: An Automated Drawing Assessment Platform for Evaluating Creativity. *Behav. Res.* **2024**, *56*, 3619–3636. [CrossRef]
25. Liu, T.; Low, K.H. Goat: Fine-Tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks. *arXiv* **2023**, arXiv:2305.14201.
26. Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. LIMA: Less Is More for Alignment. *arXiv* **2023**, arXiv:2305.11206.
27. Jang, E. Can LLMs Critique and Iterate on Their Own Outputs? Available online: https://evjang.com/2023/03/26/self-reflection.html (accessed on 9 August 2023).
28. Saunders, W.; Yeh, C.; Wu, J.; Bills, S.; Ouyang, L.; Ward, J.; Leike, J. Self-Critiquing Models for Assisting Human Evaluators. *arXiv* **2022**, arXiv:2206.05802.
29. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. *arXiv* **2022**, arXiv:2203.02155.
30. Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N.A.; Khashabi, D.; Hajishirzi, H. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022.
31. Ge, Y.; Macaluso, A.; Li, L.E.; Luo, P.; Wang, X. Policy Adaptation from Foundation Model Feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
32. Ling, Z.; Fang, Y.; Li, X.; Huang, Z.; Lee, M.; Memisevic, R.; Su, H. Deductive Verification of Chain-of-Thought Reasoning. *arXiv* **2023**, arXiv:2306.03872.

33. Irvine, R.P.; Boubert, D.; Raina, V.; Liusie, A.; Zhu, Z.; Mudupalli, V.; Korshuk, A.; Liu, Z.J.; Cremer, F.; Assassi, V.; et al. Rewarding Chatbots for Real-World Engagement with Millions of Users. *arXiv* **2023**, arXiv:2303.06135.
34. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; Dassarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.J.; et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv* **2022**, arXiv:2204.05862.
35. Alexander, S. Perhaps It Is a Bad Thing That the World's Leading AI Companies Cannot Control Their AIs. Available online: https://astralcodexten.substack.com/p/perhaps-it-is-a-bad-thing-that-the (accessed on 30 July 2023).
36. Eleuther AI. trlX: A Framework for Large Scale Reinforcement Learning from Human Feedback. Available online: https://www.eleuther.ai/papers-blog/trlx-a-framework-for-large-scale-reinforcement-learning-from-human-feedback (accessed on 6 October 2024).
37. Jiang, L.; Hwang, J.D.; Bhagavatula, C.; Le Bras, R.; Liang, J.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Borchardt, J.; Gabriel, S.; et al. Can Machines Learn Morality? The Delphi Experiment. *arXiv* **2021**, arXiv:2110.07574.
38. Cohen, R.; Hamri, M.; Geva, M.; Globerson, A. LM vs LM: Detecting Factual Errors Via Cross Examination. *arXiv* **2023**, arXiv:2305.13281.
39. Fu, Y.; Peng, H.-C.; Khot, T.; Lapata, M. Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback. *arXiv* **2023**, arXiv:2305.10142.
40. Zhang, H.; Du, W.; Shan, J.; Zhou, Q.; Du, Y.; Tenenbaum, J.B.; Shu, T.; Gan, C. Building Cooperative Embodied Agents Modularly with Large Language Models. *arXiv* **2023**, arXiv:2307.02485.
41. Meta. CICERO: An AI Agent That Negotiates, Persuades, and Cooperates with People. Available online: https://ai.meta.com/blog/cicero-ai-negotiates-persuades-and-cooperates-with-people/ (accessed on 30 July 2023).
42. Bricman, P. Hashmarks: Privacy-Preserving Benchmarks for High-Stakes AI Evaluation. *arXiv* **2023**, arXiv:2312.00645.
43. Joshi, N.; Rando, J.; Saparov, A.; Kim, N.; He, H. Personas as a Way to Model Truthfulness in Language Models. *arXiv* **2023**, arXiv:2310.18168.
44. Reenskaug, T. The Original MVC Reports. 1979. Available online: https://api.semanticscholar.org/CorpusID:61618372 (accessed on 6 October 2024).
45. Botach, A.; Zheltonozhskii, E.; Baskin, C. End-to-End Referring Video Object Segmentation with Multimodal Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
46. Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T. Alpaca: A Strong, Replicable Instruction-Following Model. Available online: https://crfm.stanford.edu/2023/03/13/alpaca.html (accessed on 29 July 2023).
47. Dubois, Y.; Li, X.; Taori, R.; Zhang, T.; Gulrajani, I.; Ba, J.; Guestrin, C.; Liang, P.; Hashimoto, T. AlpacaFarm: A Simulation Framework for Methods That Learn from Human Feedback. *arXiv* **2023**, arXiv:2305.14387.
48. Introducing LLaMA: A Foundational, 65-Billion-Parameter Large Language Model. Available online: https://ai.meta.com/blog/large-language-model-llama-meta-ai/ (accessed on 30 July 2023).
49. Jansson, A.; Nelson, D.; Sikelianos, Z. How to Use Alpaca-LoRA to Fine-Tune a Model Like ChatGPT. Available online: https://replicate.com/blog/fine-tune-alpaca-with-lora (accessed on 30 July 2023).
50. Kim, S.; Bae, S.; Shin, J.; Kang, S.; Kwak, D.; Yoo, K.M.; Seo, M. Aligning Large Language Models through Synthetic Feedback. *arXiv* **2023**, arXiv:2305.13735.
51. The Compleat Cybornaut. Available online: https://www.lesswrong.com/posts/iFBdEqEogtXcjCPBB/the-compleat-cybornaut (accessed on 30 July 2023).
52. Hugging Face Post on the Political Compass Test. Available online: https://huggingface.co/datasets/lukaspetersson/ThePoliticalCompassTest (accessed on 30 July 2023).
53. Shah, R.; Varma, V.; Kumar, R.; Phuong, M.; Krakovna, V.; Uesato, J.; Kenton, Z. Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals. *arXiv* **2022**, arXiv:2210.01790.
54. Faulty Reward Functions in the Wild. Available online: https://openai.com/research/faulty-reward-functions (accessed on 30 July 2023).
55. Krakovna, V.; Uesato, J.; Mikulik, V.; Rahtz, M.; Everitt, T.; Kumar, R.; Kenton, Z.; Leike, J.; Legg, S. Specification Gaming: The Flip Side of AI Ingenuity. Available online: https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity (accessed on 30 July 2023).
56. Kwon, M.; Xie, S.M.; Bullard, K.; Sadigh, D. Reward Design with Language Models. *arXiv* **2023**, arXiv:2303.00001.
57. Lee, J.; Xie, A.; Pacchiano, A.; Chandak, Y.; Finn, C.; Nachum, O.; Brunskill, E. Supervised Pretraining Can Learn In-Context Reinforcement Learning. *arXiv* **2023**, arXiv:2306.14892.
58. Du, Y.; Watkins, O.; Wang, Z.; Colas, C.; Darrell, T.; Abbeel, P.; Gupta, A.; Andreas, J. Guiding Pretraining in Reinforcement Learning with Large Language Models. *arXiv* **2023**, arXiv:2302.06692.
59. Lin, J.; Fried, D.; Klein, D.; Dragan, A.D. Inferring Rewards from Language in Context. *arXiv* **2022**, arXiv:2204.02515.
60. Kenton, Z.; Kumar, R.; Farquhar, S.; Richens, J.; MacDermott, M.; Everitt, T. Discovering Agents. *Artif. Intell.* **2023**, *322*, 103963. [CrossRef]
61. Shi, W.; Qiu, L.; Xu, D.; Sui, P.; Lu, P.; Yu, Z. Can LLMs Understand Social Interactions? Available online: https://chats-lab.github.io/KokoMind/ (accessed on 30 July 2023).

62. Wang, P.; Panda, R.; Torroba Hennigen, L.; Greengard, P.; Karlinsky, L.; Feris, R.S.; Cox, D.; Wang, Z.; Kim, Y. Learning to Grow Pretrained Models for Efficient Transformer Training. *arXiv* **2023**, arXiv:2303.00980.
63. Azar, M.G.; Rowland, M.; Piot, B.; Guo, D.; Calandriello, D.; Valko, M.; Munos, R. A General Theoretical Paradigm to Understand Learning from Human Preferences. *arXiv* **2023**, arXiv:2310.12036.
64. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv* **2023**, arXiv:2305.14314.
65. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
66. Marie BQLoRa: Fine-Tune a Large Language Model on Your, G.P.U. Available online: https://towardsdatascience.com/qlora-fine-tune-a-large-language-model-on-your-gpu-27bed5a03e2b (accessed on 30 July 2023).
67. Wei, A.; Haghtalab, N.; Steinhardt, J. Jailbroken: How Does LLM Safety Training Fail? *arXiv* **2023**, arXiv:2307.02483.
68. Amayuelas, A.; Pan, L.; Chen, W.; Wang, W.Y. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models. *arXiv* **2023**, arXiv:2305.13712.
69. Nguyen, T.T.; Huynh, T.T.; Nguyen, P.L.; Liew, A.W.C.; Yin, H.; Nguyen, Q.V.H. A Survey of Machine Unlearning. *arXiv* **2022**, arXiv:2209.02299.