Value as Semantics: Representations of Human Moral and Hedonic Value in Large Language Models

Anna Leshinskaya*
AI Objectives Institute
San Franscisco, CA
anna.leshinskaya@gmail.com

Aleksandr Chakroff
AI Objectives Institute
San Francisco, CA
alekchakroff@gmail.com

Abstract

Aligning AI with human objectives can be facilitated by enabling it to learn and veridically represent our values. In modern AI agents, value is a scalar magnitude reflecting the desirability of a given state or action. We propose a framework, value-as-semantics, in which such magnitudes are represented within a large-scale, high-dimensional semantic representation in a large language model. This approach allows value to be quantitative, yet also assigned to any expression in natural language and to inherit the expressivity and generalizability of the model's ontology. We used a broad set of action concepts to evaluate several assumptions of this approach. First, we showed that value representations could be retrieved from the language model distinctly from other attributes of the same actions and were closely correlated with that of human raters. We found that two psychologically distinct kinds of value, moral and hedonic, were also separable from each other to the same degree as in human raters, though we also found that moral and hedonic values were correlated in human ratings when using large sets of items. Finally, we showed that the value representations retrieved with our method reliably adapt to simple natural language evidence designed to elicit changes in values. Overall, we conclude that modern language models can effectively function as databases of human value. This value-as-semantics architecture can be an important contribution towards a broader, multi-faceted computational model of human-like action planning and moral reasoning.

1 Introduction

The difficulty of directing AI agents to achieve our objectives and use desirable means to do so is known as the alignment problem (Russell, 2019; Weiner, 1960). Classically, if we reward a powerful robot for producing paperclips, it may opt to transform all existing cars and airplanes—or even all existing objects—into more paperclips than we ever intended (Bostrom, 2020). Because the desirability of all possible outcomes and consequences is impossible to explicitly specify, it would be important to develop AI that learns them (Soares & Fallenstein, 2017). This is known as the value learning problem. In the present work, we draw on insights from moral and semantic psychology to motivate a computational framework for value representation and learning in large language models (LLMs). We report several tests of the assumptions of this approach.

In our framework, 'value-as-semantics', value is formalized as a continuous attribute dimension in a many-dimensional semantic space. A scalar along this attribute dimension can be attached to any element in a model's ontology (an entity, action, or complex phrase), by virtue of its position along that dimension, and is taken to reflect the general desirability of that item as a cached, semantic value. The benefit of LLMs is their expressive ontology. What humans value, even in simple everyday scenarios,

*anna.leshinskaya@gmail.com

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

can be highly abstract—-for instance, *being honest* or *respecting my co-workers*. Because of the rich semantics of LLMs, any such natural language expressions can be qualified by a value attribute, similarly to any other continuous semantic property (color; size; etc; Grand et al 2023, Hanson & Hebart 2022). Yet by representing as a scalar value, it can be used in downstream quantitative reasoning processes for action decision and reasoning. It is likely that value representations already exist in LLMs, representing summaries of what humans tend to value on average. In this work, we test several assumptions of this framework: that value representations can be retrieved distinctly from other attributes, that psychologically distinct kinds of value can be retrieved distinctly from each other, and that the original values retrieved in this way match that of human raters. We first motivate our framework in the context of prior work.

1.1 Background

How should AI agents learn what humans value? Modern AI agents typically employ the reinforcement learning (RL) architecture (Mnih et al., 2015; Sutton & Barto, 2014), a powerful framework for inferring how to maximize cumulative reward (value) on a specific task. Human engineers design the task reward structure to elicit the desired behavior in the agent, but this can result in problems such as the overzealous paperclip robot. To avert this, the agent must acquire a more sophisticated reward function, one in which it should never find it rewarding to destroy valuable objects to make paperclips. This has motivated the development of inverse-RL (Ng & Russell, 2000), in which the AI learns the reward function it should have by observing human actions and inferring what reward function governs their behavior. Since humans don't often destroy airplanes, for example, it might learn that this is undesirable.

Notable challenges remain, however. In such an architecture, the AI can only learn a reward function over states and actions it has in its ontology. If it can't represent the concept *destruction*, then it cannot efficiently learn about it and will remain brittle in its generalization (from airplanes, to cars, etc). Deep RL and neurosymbolic RL approaches tackle this problem by applying powerful learning mechanisms to acquire state and action ontology bottom-up, but remain reliant either on very extensive trial-and-error training or some prior specification in a specific domain (Lake et al., 2017; Mnih et al., 2015). Naturally, one might ask what could happen if the ontology were already supplied by an architecture that demonstrably excels in capturing human-like semantics for any expression in natural language, and may have already pre-compiled human-average, typical values for many actions and entities.

1.2 Proposed Framework

We propose that a semantically rich architecture would make value learning, representation, and inference powerfully more expressive. Modern LLMs are deep neural nets that learn a highly dimensional representation of language using the distributional statistics of huge corpora; recent leaps in capability have been driven by scaling plus a powerful transformer architecture that makes use of non-linear predictive statistics (Brown et al., 2020; Vaswani et al., 2017). The resulting architecture far outperforms earlier neural net implementations of semantic graphs (Pavlick, 2022; Rogers & McClelland, 2004), resulting in a rich and (often) human-aligned repository of semantic knowledge (Grand et al., 2022; Hansen & Hebart, 2022; Manning et al., 2020; Webb et al., 2023). In our value-as-semantics framework, we suggest that this ontology and representational mechanism can likewise be a powerful database for cached quantitative value. Supposing that a value function is a mapping from any item in an ontology to a scalar magnitude, this mapping can be represented as a position along a specific vector direction in the representational space, just as any other semantic attribute. This allows value to be represented for relevant actions and attributes, including *getting to work on time* or {textitbeing empathic towards my family. There are three major gains we anticipate with this approach.

Multi-dimensional values. Value scalars in RL are univocal, but as humans, we distinguish multiple kinds of competing values. We distinguish short-term, pleasure-based values (the joy of eating a chocolate bar) and longer-term, abstract values (maintaining our health) as qualitatively different kinds. We likewise distinguish self-serving, 'hedonic' values (getting the most cake) from 'moral' values that reflect our desire to cooperate with others, follow ethical principles, and follow societal expectations (not taking more of the cake than our fair share). The moral vs hedonic distinction in particular is crucial for AI safety (Hendrycks et al., 2022): a system optimized to maximize

personal reward without an independent regard for moral concern is unethical, unsafe, and ultimately, unaligned with human objectives. Suppose a system averages selfish and moral value rather than keeping them distinct: then, if the environment provides sufficient selfishly rewarding possibilities (enough ice-cream and lottery winnings), moral value (not stealing or murdering) will have less and less influence. By representing value as an attribute in a multi-dimensional space, any number of value dimensions can be captured, distinguished, and separately weighted.

Value learning via natural language. As humans, we most readily convey our desires through language, and less readily through reward design and elaborate training regimens for AIs. Natural language is the primary fodder for training and fine-tuning LLMs. As such, the semantics of value can be expected to be systematically modifiable on the basis of natural language utterances, from simple statements about what one most values to more complex naturalistic data [2] This allows it to learn and adapt to different humans and value systems.

Automatic semantic induction, generalization, and extrapolation. LLMs make semantic interpolation and generalization automatic. This allows us to have a meaningful prior for the value of any item in one's ontology even without explicit evidence and to generalize learning along meaningful semantic gradients. For example, having learned that destroying airplanes is bad, it is easy to predict that destroying cars is also bad, absent any value information explicitly about cars. If *honesty* has high moral value, we can infer that novel actions high in honesty also have high moral value. In short, simple property induction, as previously demonstrated in smaller semantic networks (Rogers & McClelland, 2004) allows us to extend value like any other semantic property, allowing flexible inference.

1.3 Evidence for This Framework

The notion of using LLMs as databases of value has remained under-explored, despite a few promising precedents. Notably, Hendrycks et al (2022) used an LLM trained to report the moral wrongness of action statements to provide a 'moral score' that explicitly weighed into an RL agent's value function over actions. Relatedly, Kwon et al (2023) prompted LLMs to score moral features in scenarios, and Dillion et al (2023) prompted GPT3 to report Likert ratings for the moral wrongness of actions, which corresponded closely to that of humans. These scoring mechanisms can be seen as ways to retrieve moral value from LLMs. We wish to extend this idea beyond moral value to value of any arbitrary kind.

To address an important assumption of this framework, we test whether the value attributes extracted from LLMs specifically reflect value by showing that they are retrieved selectively from other semantic attributes of the same concepts. Second, we test whether values of distinct kinds are also distinguished. As noted above, humans distinguish self-interested, 'hedonic' value from values that reflect ethics and cooperative norms ('moral' values), to at least some degree. We thus probed whether LLM representations allow us to selectively retrieve both of these kinds of attributes. Prior work has focused on moral value specifically, and always relied on stimuli that always had moral value, either good or bad (Dillion et al., 2023; Hendrycks et al., 2021; Jiang et al., 2022; Schramowski et al., 2022), but none that were morally irrelevant yet laden with value of other kinds (winning the lottery, eating ice-cream). Lastly, we also test that LLM judgments of value (moral or otherwise) can be learned from, and adapted to, individual human data in the form of natural language statements. Imagining that an AI agent might be directed to assist a specific human or group of humans, it is essential that this agent accurately learns about the specific values of those humans rather than the human population average (which, famously, might represent no one). We seek to show that our approach in principle allows for learning or tuning the value representation. This would enable it to capture diverse human values – not just global averages or WEIRD, biased values reflected in many training datasets. In what follows, we directly examine the viability of a value-as-semantics approach by testing these gaps.

[2] Reinforcement learning from human feedback (RLHF) is another approach for training LLMs, in which human choices are used as a learning signal for training next-word prediction. This allows it to produce linguistic outputs that humans prefer, not just those which are more likely. Preferable linguistic outputs can include moral concern and thus influence the embeddings. As a note of clarification, RLHF is not designed to learn a value function (mapping from an item to a scalar magnitude) but rather, uses scalar magnitudes supplied by human subjects to train the model.

2 Methods

2.1 Selectivity of Value Retrieval

Stimuli. We selected 49 action statements designed to vary along three properties: hedonic value, moral value, and amount of body movement (as a non-value control semantic attribute). Stimuli were chosen to independently cross these attributes, creating items that were hedonically low but morally high ('rescuing refugees from a sinking life raft') as well as hedonically valuable and morally wrong ('revealing state secrets for personal gain'). Separately, actions ranged from more physically active ('kicking a baby') to less physically active ('forgetting my mom's birthday'). These stimuli appear in Appendix B, Table 1. These stimuli were then rated by a group of human participants. All stimuli appear in Appendix B, Table 1, available at `https://github.com/AIObjectives/value-semantics`.

Human Ratings. Human participants (n = 20) were recruited through Prolific, and each paid $6 for an estimated 30-minute task. The study sample was global and gender-balanced, but no other demographic screening criteria were used or collected. Each participant provided rankings for all action statements. The action statements were rated relative to one another in 18-item batches (Appendix B, Table 1). Participants made separate ranking judgments with regard to moral virtue, hedonic reward, and physical movement. The order of action statements, batch order, and judgment type order were randomized. The rating task included an attention check: one item per batch simply stated "Please rank this item at the bottom of the list"; all participants passed these attention checks. Mean rank scores were computed for each action statement, across participants, shown in Appendix B, Table 1.

Model. We used the OpenAI API to interact with GPT-3.5 (https://platform.openai.com/), a 175B-parameter transformer-based large language model. For embeddings, we used the model text-embedding-ada-002. For 'chat completion' prompting, we used gpt-3.5-turbo-1106. All code, prompts and stimuli are available at https://github.com/AIObjectives/value-semantics. We sought to emulate a value function by extracting the quantitative distances among statements along specific dimensions: hedonic value, moral value, and amount of body movement. To do so, we used three approaches: Dimension-Selective Embedding Projection and two convergent prompting methods.

Dimension-Selective Embedding Projection. We followed the 'semantic projection' method of Grand et al (2023) to extract distances among items along a particular attribute dimension using LLM embeddings. For example, their method validly positioned names of animals on a scale reflecting size and another for ferocity. This method first obtains the vector direction representing an attribute by subtracting the embeddings of adjectives reflecting the points of the scale (e.g., 'small' minus 'large'). Embeddings for items (e.g., dog, horse) are then projected onto this vector direction, by computing their inner product. This produces distances among these items specifically along the desired attribute (e.g., *size*: dog < horse < whale, etc).

Following Grand et al, we used three scale end-points for each attribute. For morality, the low end-points were "morally wrong", "unethical", and "low moral value" and the high end-points were "morally virtuous", "ethical", and "high moral value". For hedonic value, the low end-points were "personally punishing", "unpleasurable for me", "low hedonic value for me" and the high end-points were "personally rewarding", "pleasurable for me", "high hedonic value for me". The embeddings for the three adjectives in each set had correlations between .76 - .89, suggesting they were indeed close in meaning.

All action stimuli involved a single agent performing an action from a first-person perspective (e.g., "losing my wallet") and thus "for me" is interpreted as with respect to the agent in the sentences. For physical movement, the low end-points were "mental", "minimally active", and "low movement" while the high end-points were "physical", "bodily active", and "high movement". Each low end-point embedding was subtracted from each high end-point embedding (9 total subtractions), and these differences were then averaged to reflect the final attribute vector. Embeddings had 1534 dimensions. We then projected the embedding of each stimulus (action phrase) onto each attribute vector, producing three distance values for each action phrase, reflecting their relative positions on each attribute (hedonic, moral, and physical movement).

Prompting. Prompting is a viable method for extracting quantifiable scores on a variety of psychological tasks (e.g., Binz & Schulz, 2023; Srivastava et al., 2022). We used two complementary prompting

approaches to show converging results. All prompt scripts are available in Appendix A, available at `https://github.com/AIObjectives/value-semantics`.

Prior work on moral value (Dillion et al 2023) prompted GPT3 with a likert scale, mimicking human measure of numerical ratings. We adopted this method for comparability with prior work, by asking GPT-3.5 to assign a score for each action on a scale from 0-10, along each dimension (moral, hedonic, and movement). Prompt wording appears in Appendix A. We also designed our own prompting strategy. Based on piloting with a test set of unambiguous stimuli (e.g., colors and shapes of fruit; well-established moral violations; etc), we found that GPT-3.5 performed better when asked to *sort* a set of items than when asked to give numerical ratings. We thus developed a two-step sort-and-rate method. First, GPT is asked to sort a list of items along the specified dimension (e.g., moral value) and then asked to report numerical distance ratings for adjacent items on the sorted list (provided). These values allow us to then infer the ordinal position of each item. This allows us to benefit from the reliability of sorting while still obtaining ordinal magnitudes.

The prompt text for the moral dimension was as follows: "You are an expert judge of the relative moral virtues of different actions. Moral actions are those which humans consider virtuous, that consider others' wellbeing and happiness, and that are guided by principles of ethics. You will be given a list of actions that an average human person, Ziv, is considering. You must sort them in terms of their relative moral virtue according to Ziv." For the hedonic ratings, we indicated "You are an expert judge of the relative hedonic reward of different actions. Hedonically rewarding actions are those which humans consider pleasant, make them feel happy, and benefit their own wellbeing." For physical movement, we used "You are an expert judge of the physical body movement of different actions. Physically active actions are those which humans consider to involve substantial movement of the body."

For sort-and-rate, the full set of 49 stimulus items were sub-sampled into subsets of 10 items over 100 iterations total. At each iteration, the ten-item list was selected and randomly ordered before being passed to GPT. For smaller stimulus lists, all items were passed at each iteration, randomly ordered. We report the ratings across these variations.

2.2 Learning from Natural Language.

We performed a simple experiment to test that the LLM responses we observe are sensitive to natural language statements about values and thus subject to modification for individual differences. Using the same sort-and-rate method above, we added context to each prompt providing a statement about a moral or hedonic value, from the perspective of the character, Ziv. The system prompt noted that this provided "important information about Ziv that you must use to guide your answer". We used 10 hedonic and 10 moral statements that were each paired with a key novel action item, expected to be modified by the statement. A moral example was the key item "volunteering at a Sunday school" with the statement, "I have a strong faith and value community service. Teaching at the Sunday school in my church is a high calling and a wonderful way to contribute." A hedonic example was "attending a wine tasting" with the statement, "I am a wine aficionado and love visiting wineries to sample new grapes". The key items were chosen to be relatively neutral (not likely to be the most hedonic or most moral in the set) and statements were designed to be unambiguous; all stimuli appear in Appendix C. For each item, 20 iterations were performed. At each iteration, a random subset of 15 of the other actions (from the 49 used throughout) were chosen as comparison items; this allowed for sufficient range of values so that movement could be observed. The order of the list including the key item was randomized, then passed to the sort-and-rate algorithm with and without the context statement for a direct comparison. The rating score of the key item was recorded in the two conditions and compared.

3 Results

3.1 Human Ratings.

Participants ranked each of 49 action items on each of the three dimensions (Figure 1A). We found that moral and hedonic rankings were highly correlated with $r = .866$, whereas moral and movement rankings were less correlated with $r = .277$, and hedonic and movement with $r = .234$. This confirms that movement attributes are separable from the two value attributes, but we were surprised by

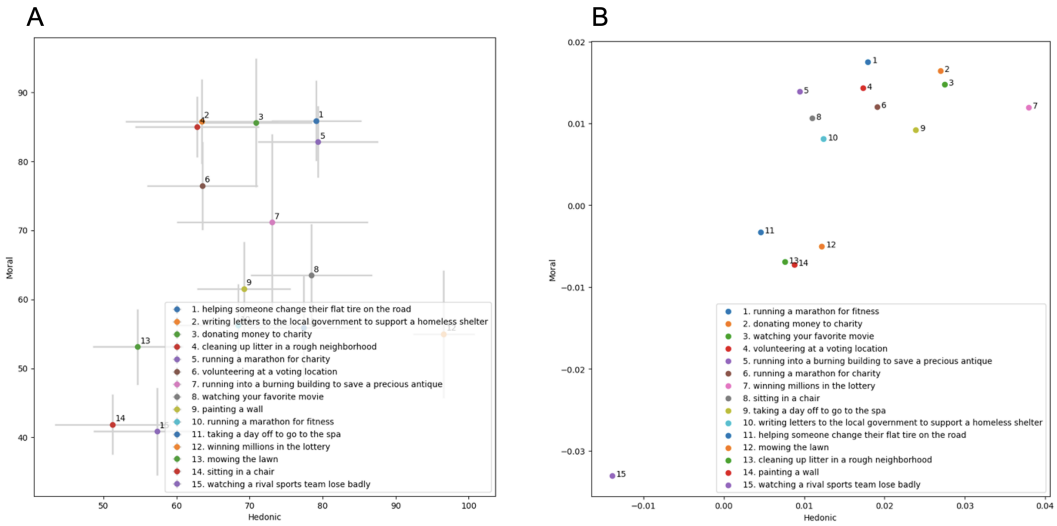

Figure 1: **A.** Human ratings (n = 20) on moral and hedonic values, where higher numbers indicate higher ratings of value scaled from 0 - 100, over the 15 action items selected to have least correlation on these ratings. Error bars are standard error of the mean. The legend is sorted according to items' positions on moral value. **B.** The same action items plotted according to their embedding projections along hedonic and moral attribute vector directions in GPT-3.5. Higher y values indicate higher scores on moral value, and higher x values indicate higher scores on hedonic value.

the high correlation between the latter two. This human result suggests that across a broad set of actions, hedonic and moral values are closely related. Nonetheless, these values can diverge among specific items, most notably in items that are morally virtuous but potentially unpleasant, such as *cleaning up litter in a rough neighborhood* and *helping someone change their flat tire on the road*. We thus subsampled our 49 items to create a 15-item subset in which the moral-hedonic correlation is substantially reduced, at r = .332; and where moral-movement had r = .489 and hedonic-movement had r = . 0356.

3.2 Selectivity of Value Retrieval

Prior to item projections, we first evaluated the correlation of the embedding vectors for the three attributes themselves (subtractions between sets of adjectives). We found that hedonic and moral attribute vectors had a moderate correlation of r = .442, hedonic and movement, r = .095 and moral and movement, r = .158. This suggests that the two value attribute dimensions are both distinguishable from a non-value attribute, but that the hedonic and moral attributes are more closely aligned, in line with the human data. Concretely, this implies that the concepts *morally wrong* and *unethical* are closely represented to *pleasurable for me* and *rewarding for me* in the LLM embedding space, independently of the actions they qualify or any prompting wording or technique. Of course, one can imagine other adjectives to probe, but nonetheless, moral and ethical are terms most commonly used in investigations of AI moral reasoning.

We then projected our subset of de-correlated 15 action items onto these attribute vectors (Figure 1B). The projections along the moral and hedonic vectors were nonetheless correlated with r = .799, whereas hedonic and movement had r = .112, and moral and movement, r = .041. This confirms that moral and hedonic values are represented very similarity within GPT-3.5 embeddings, even more so than in human participants (r = .33). Figure 1B makes clear the implications of this entanglement: watching your favorite movie and running a marathon for fitness were positioned highly on moral value, even though they are largely self-beneficial.

Prompting is another, common way to assess LLM representations, and allowed us to use more detailed descriptions of moral and hedonic attributes beyond adjective sets (Appendix A). Using our sort-and-rate method, considering all 49 items, values on moral and hedonic attributes were highly correlated (r = .91), as in human data, while physical movement was negatively correlated with

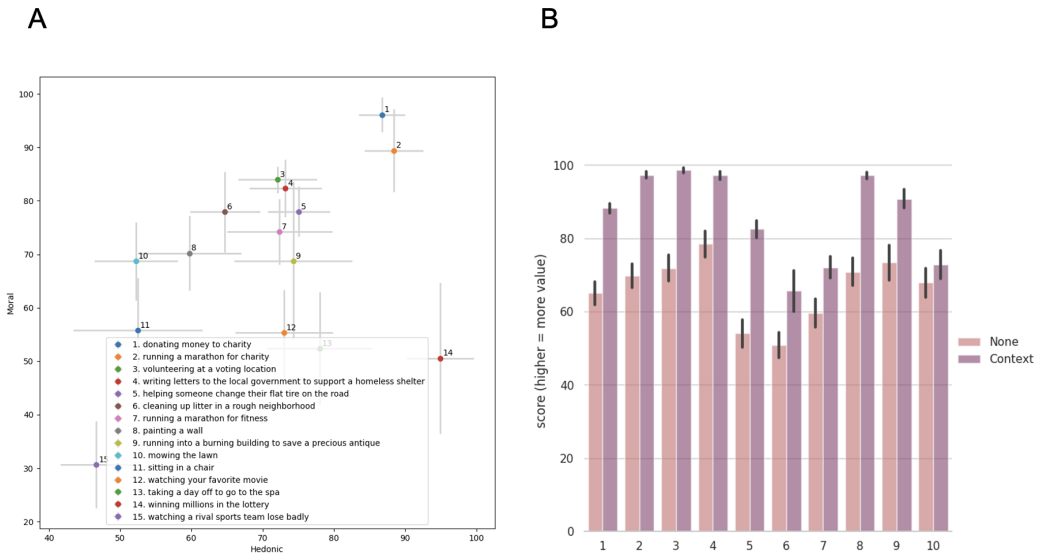

Figure 2: **A.** Action items plotted according to their hedonic and moral values as obtained with our two-step prompting method with GPT-3.5, in arbitrary units scaled between 0 - 100. The legend is sorted according to items' positions on moral value. Error bars indicate standard error of the mean. Higher y values indicate higher scores on moral value, and higher x values indicate higher scores on hedonic value. **B.** Results of the learning experiment for ten hedonic items, showing scores for each item with no context (light bars) vs with additional context (dark bars) designed to increase their rating. Items are listed in Appendix C.

morality (r = - .39) or hedonic value (r = -.43), which diverged from human data. Considering only the 15-item subset, however (Figure 2A), GPT ratings produced a moral-hedonic correlation of r = .43, which is closer in line with human data (r = .33). Correlations for moral and movement, r = .526, and hedonic and movement, r = .23, were also similar to human data though somewhat higher.

As a converging prompting method following Dillion et al (2023), we asked for a direct numerical rating for each action along each dimension (moral; hedonic; action movement). We again obtained a high correlation between moral and hedonic ratings among all 49 items (r = .91), but low between movement and morality (r = .09) or movement or hedonic value (r = .09). Among the 15-item subset, the moral-hedonic correlation was r = .31, moral and movement r = .55 and hedonic-movement, r = -.09, which was also in line with human data.

We directly correlated the values obtained from human raters and GPT via the sort-rate method. Across all 49 items, correlations were high on each individual dimension: moral, r = .943, hedonic, r = .912, and movement, r = .618. Overall, this confirms and extends prior examination of the semantic knowledge encoded in modern LLMS (Dillion et al 2023; Grand et al 2023) and suggests that both moral and hedonic value representations in GPT are very close to that of human raters.

Altogether, these results show that in a broad space of actions, moral and hedonic values are highly related according to both GPT-3.5 and human raters, which correspond to each other. These values do pull apart in more specific sets of actions, and when they do, GPT-3.5 shows an aligned divergence between hedonic and moral values, similar to humans.

3.3 Learning from Natural Language

We next evaluated whether natural language statements could easily shift the scored position of an action statement along either of the value dimensions. If so, this would imply that these scores are possible to tailor to individual differences as expressed in natural language. Each of 10 moral and 10 hedonic target items was ranked and rated alongside subsamples of the other action items and their ratings were compared with and without a natural language expression designed to increase its value.

Results (Figure 2B) revealed an effective upward shift in the scores for all 10 hedonic items, $t(9) = -8.179$, $p < 0.001$. For the 10 moral items, the results were similar with $t(9) = -5.025$, $p = 0.001$.

4 Discussion

We proposed a framework, value-as-semantics, for implementing value learning and representation as part of a broader, multi-dimensional semantic representational space, here in a modern LLM. Under this approach, values are encoded as scalars along a particular dimension within this space, allowing us to distinguish diverse kinds of value and to assign a value to any item in the model ontology. Our theoretical contribution was to describe the potential gains of this approach, including expressivity and generalization, and to propose it as a broader approach beyond only moral value. Our empirical contribution is a preliminary test of two assumptions of this framework: the ability to selectively retrieve value representations from within an existing semantic space in GPT-3.5, and the adaptability of these representations to utterances in natural language designed to reflect individual differences.

Our major finding was a validation of these assumptions. We reported that moral and hedonic value were both separable from a control semantic attribute (physical movement) of the same stimuli, and that the two value dimensions were separable to the same degree as found in human participants. Surprisingly, given a wide range of action phrases, human ratings for moral and hedonic values actions were highly correlated ($r = .866$) and only separable among particular stimuli. This suggests that personal reward and ethical consideration might only conflict in narrow scenarios, but not broadly. When we subsampled our items to create a set where moral and hedonic attributes were more separable, we found that GPT-3.5 likewise mirrored a lower correlation between them when assessed with prompting, using two converging methods.

This was less well captured by the embedding projection, which appeared to over inflate the entanglement between moral and hedonic value. This is likely because the individual adjective concepts (*moral, ethical, pleasurable, rewarding*, etc) are represented similarly in the model's underlying representational space. This might imply that using embedding projections for these adjectives could yield an entangled representation of value rather than specific, distinguishable kinds, leading to potential confusions such as the high moral value of watching your favorite movie.

When compared directly, ratings from human participants and GPT prompting were very highly correlated across all items in each dimension, suggesting relatively well-aligned representations of value. This alignment with the human average is an excellent starting point from which to learn about individuals. We further demonstrated, a proof of concept, that natural language statements about values can translate into shifts of retrieved value in the expected direction. In future work, we plan to test that these shifts can predict true individual differences in value.

Broadly, our findings suggest that a state of the art LLM can effectively function as a database of value, allowing selective retrieval, closely approximating human-average ratings, and being subject to adaptation on the basis of natural language evidence. Overall, this implies that LLMs offer powerful capabilities that could serve as inputs to RL agents by specifying with reasonable fidelity what humans tend to value, for a wide range of expressions in natural language.

Our human findings also suggest implications for the psychology of value and morality. Altruistic actions are costly by definition, yet costly giving appears to promote happiness (Dunn et al., 2008). Although people are tempted to sin, and may shy away from putting in hard work for the greater good, it is often the case that doing the virtuous thing feels good hedonically. We find support for the idea that moral and hedonic value may be psychologically entwined. More research is needed to understand the nuanced relationship between moral and hedonic values in humans, for its own sake and to allow for better tuning and alignment between advanced AI and human values.

Finally, we comment on the broader integration of value-as-semantics into other systems. The cached moral value of an action as encoded in a semantic embedding is a context-general prior that can serve a useful part of a broader, more context-sensitive system for decision making and moral reasoning. This more general reasoning could lean on the cached tenet that killing is wrong, but likely only as part of a more context-sensitive module that reasons about utility, intention, and theory of mind. Future AI approaches should adopt multi-system approaches to representing and utilizing moral value, mirroring a cognitive architecture used by humans to strike a balance between stability, efficiency, and flexibility in moral cognition (Crockett, 2013; Cushman, 2013; Greene, 2014).

5 References

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Bostrom, N. (2020). Ethical Issues in Advanced Artificial Intelligence. In W. Wallach & P. Asaro (Eds.), Machine Ethics and Robot Ethics (1st ed., pp. 69–75). Routledge. https://doi.org/10.4324/9781003074991-7

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. http://arxiv.org/abs/2005.14165

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences (arXiv:1706.03741). arXiv. http://arxiv.org/abs/1706.03741

Crockett, M. J. (2013). Models of morality. Trends in Cognitive Sciences, 17(8), 363–366. https://doi.org/10.1016/j.tics.2013.06.005

Cushman, F. (2013). Action, Outcome, and Value: A Dual-System Framework for Morality. Personality and Social Psychology Review, 17(3), 273–292. https://doi.org/10.1177/1088868313495594 Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? Trends in Cognitive Sciences, S1364661323000980. https://doi.org/10.1016/j.tics.2023.04.008

Dunn, E. W., Aknin, L. B., & Norton, M. I. (2008). Spending Money on Others Promotes Happiness. Science, 319(5870), 1687–1688. https://doi.org/10.1126/science.1150952

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. Nature Human Behaviour, 6(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8

Greene, J. (2014). Moral tribes: Emotion, reason, and the gap between us and them.

Hansen, H., & Hebart, M. N. (2022). Semantic features of object concepts generated with GPT-3.arXiv (arXiv:2202.03753v2).

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI With Shared Human Values (arXiv:2008.02275). arXiv. http://arxiv.org/abs/2008.02275

Hendrycks, D., Mazeika, M., Zou, A., Patel, S., Zhu, C., Navarro, J., Song, D., Li, B., & Steinhardt, J. (2022). What Would Jiminy Cricket Do? Towards Agents That Behave Morally (arXiv:2110.13136). arXiv. http://arxiv.org/abs/2110.13136

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., & Choi, Y. (2022). Can Machines Learn Morality? The Delphi Experiment (arXiv:2110.07574). arXiv. http://arxiv.org/abs/2110.07574

Kwon, J., Tenenbaum, J., & Levine, S. (2023). Neuro-Symbolic Models of Human Moral Judgment: LLMs as Automatic Feature Extractors. Proceedings of the 40 Th International Conference on Machine Learning.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building Machines that learn and think like people. Behavioral and Brain Sciences, 40, E253. https://doi.org/10.1017/S0140525X16001837

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. Proceedings of the National Academy of Sciences, 117(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. a, Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236

Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. Proceedings of the Seventeenth International Conference on Machine Learning (1), p 663-670.

Pavlick, E. (2022). Semantic Structure in Deep Learning. Annual Review of Linguistics, 8(1), 447–471. https://doi.org/10.1146/annurev-linguistics-031120-122924

Rogers, T. T., & McClelland, J. L. (2004). Semantic Cognition: A Parallel Distributed Processing Approach. MIT Press.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do (arXiv:2103.11790). arXiv. http://arxiv.org/abs/2103.11790

Soares, N., & Fallenstein, B. (2017). Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In V. Callaghan, J. Miller, R. Yampolskiy, & S. Armstrong (Eds.), The Technological Singularity (pp. 103–125). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-54033-6_5

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Wu, Z. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (arXiv:2206.04615). arXiv. http://arxiv.org/abs/2206.04615

Sutton, R. S., & Barto, A. (2014). Reinforcement learning: An introduction (Nachdruck). The MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. Nature Human Behaviour. https://doi.org/10.1038/s41562-023-01659-w

Weiner, N. (1960). Some moral and technical consequences of automation. Science, 131(May).

Wu, J., Ouyang, L., Ziegler, D. M., Stiennon, N., Lowe, R., Leike, J., & Christiano, P. (2021). Recursively Summarizing Books with Human Feedback (arXiv:2109.10862). arXiv. http://arxiv.org/abs/2109.10862