# Learning When Not to Measure: Theorizing Ethical Alignment in LLMs

**William Rathje**
University of California, Berkeley
wrathje@berkeley.edu

**Abstract**

LLMs and other forms of generative AI have shown immense promise in producing highly accurate epistemic judgements in domains as varied as law, education, and medicine – with GPT notably passing the legal Bar exam and various medical licensing exams. The safe extension of LLMs into safety-critical professional domains requires assurance not only of epistemic but ethical alignment. This paper adopts a theoretical and philosophical approach, drawing from metaethical theories to argue for a distinction hinging around quantitative, axiological comparability that separates Kantian ethics from not only the utilitarianism it is well-known to oppose, but from just distribution theories as well, which are key to debiasing LLM models. It presents the novel hypothesis that LLM ethical acquisition from both corpus induction and RLHF may encounter value conflicts between Kantian and just distribution principles that intensify as they come into improved alignment with both theories, hinging around the variability by which self-attention may statistically attend to the same characterizations as more person-like or more resource-like under distinct prompting strategies.

## Introduction

Significant interest in GPT's extraordinary capabilities to synthesize accurate epistemic judgements in a vast range of domains through successful transfer learning has spurred recent interest in the future of AI in social arenas as diverse as medicine (Thirunavukarasu 2023), law and criminal justice (Grossman et al. 2023), education, artistic production, and mental health therapy (Cheng et al. 2023). For all of GPT's much hyped potential in, for instance, passing medical licensing exams (Kung et al. 2023) or passing the Bar Exam (Katz et al. 2024), success in professional domains (not to mention mission-critical contexts) demands not only accuracy but strong and unwavering ethical judgement and reasoning. For AI systems like LLMs to be deployed in safety critical arenas such as medicine, law and policy, or education, we should at minimum trust that these systems at the most fundamental level are producing ethically defensible outputs.



LLMs appear to have learned various emergent properties above mere statistical language prediction. This is evinced, for example, by its capacity to solve problems accurately in domains such as medicine or law, suggesting that it has inferred various emergent properties about knowledge relevant to these domains (Wei 2022). Fields such as law, medicine, education, or business require not only strong analytical and epistemic capabilities but, notably, good ethical judgement. The extraordinary capabilities of GPT suggest that it has, in fact, inferred certain emergent normative properties rather than merely epistemic ones in these domains simply in order to appear competent and coherent to human interlocutors. Moreover, the capacity of GPT to infer emergent properties in general would imply that it should have inferred various normative properties, unless we are to believe that epistemology and normativity are fully separate epistemic categories (Lauer 2021).

Value alignment in LLMs is typically a two-stage process; first, emergent properties such as norms are learned directly from data, but because these inductions embed toxicity or bias, they are then fine-tuned based on human input, typically using reinforcement-learning with human feedback (RLHF). This is highly distinct from human ethical agency, however. In human ethics, people are exposed to cultural tropes that may be traditional, biased, or harmful, and they are exposed to social feedback such as peer pressure that might be equally harmful. Nevertheless, human ethical agents in many conceptualizations also possess an "ethical compass" which enables them to reason autonomously from culture or normative conformity and break with even a systematically harmful culture or environment. LLMs do not attempt to model this "ethical compass" per se and instead model cultural learning and social/normative feedback (through RLHF) toward their actions only. While it is possible that an "ethical compass" could emerge as a property of these systems (e.g. by tying RLHF results to reinforced neural-net organization), without explicit modeling it is unclear that LLMs really exhibit ethical autonomy, including from harmful cultural tropes and traditions. What LLMs do not, today, tend to do

is model the process by which individuals are taught reasons for ethical action; i.e. while social feedback may motivate and differentially reward behavior, the socialization and maturation process also involves learning rational principles not only through observation and mimesis but through explicit rational explanation.

However, LLMs also raise serious questions about the status of any such 'ethical compass'. Whereas Kant held that humans are distinct from mechanisms in possessing rational cognition and ethical autonomy, it is similarly true that humans are material beings that emerge from a physical environment, and so the organization giving rise to ethical reason should be supervenient over physical properties and laws. Whether one should explicitly model ethical cognition vs. attempt to infer it emergently in LLM contexts gets at the heart of this dilemma over how ethical autonomy develops; it seems to facilitate free deviation from environmental context, but it is hard to understand how it does not ultimately originate within a deterministic environmental context. Moreover, LLMs do not simply reason over material objects but over textual samples that project human intentions, propositional attitudes, and autonomous judgements. LLMs clearly lack any explicitly given ethical compass, but it is a thorny philosophical problem to determine if and when they might ever learn one.

This paper provides a philosophical and theoretical overview of anticipated problems in the ethical alignment of LLMs in combination with a review of evaluation techniques to date. We propose a novel concern: namely, that just distribution and utilitarian principles often cohere (because they both involve resource comparison) and come into conflict with Kantian principles. As ethical alignment of AI systems improves, we would expect models to become both more just and more Kantian. But there is a risk that inductive methods such as RLHF might error in their interpretation of when to apply just-distributional standards versus Kantian standards. If LLMs eventually come to play a significant role in autonomous decision making in critical domains such as law and medicine, these problems may lead to risky failure cases.

## Value Alignment

### Which Values Should Motivate Alignment?

AI value alignment has frequently been conceptualized as a problem of aligning ML models to human ethics. Value theory, however, extends across the set of normative properties and principles, which span across not only ethics but justice, aesthetics, prudence, axiology, etc. Theories of ethics, aesthetics, and justice are often compatible but at times come into direct conflict. I briefly discuss predominant meta-ethical theories and then compare them to other categories of normative theory to emphasize their interactions.

Prominent theories of ethics include virtue ethics, consequentialism, deontology, and care ethics. Virtue ethicists argue that ethical actions are those that evince and achieve virtues such as bravery, restraint, or benevolence. Consequentialists argue that ethical actions are those which realize optimal consequences, according to some function or standard of evaluation. Utilitarianism is one popular form of consequentialism, and it claims that actions are ethical when they maximize the utility – or benefit – to the greatest number possible. Deontologists argue that actions are ethical when they adhere to moral rules or principles. Deontology is frequently paired with Kantian deontology, which (among other claims) asserts that good actions are a) those that realize universalizable principles and b) by contrast to consequentialism, those that realize ethically valid *intentions* even if they fail to realize optimal *consequences*.

Utilitarianism and deontology somewhat famously break along an important issue of the value of human life. For most Kantian deontologists, an individual human life is of intrinsic work and irreducible or non-fungible value; that is, humans are valuable "beyond measure" or quantification and should not ever be treated as a means or sacrificed for some definition of "greater good." A strong weakness for utilitarianism is its general inability to enforce these highly intuitive principles; for the utilitarian, when push comes to shove, it may be not only ethically permitted but ethically necessary to, e.g. sacrifice or instrumentalize some people for the benefit of a larger number. This is classically depicted in the well-known "trolley problem," wherein utilitarians and deontologists disagree over whether it is permissible to pull a lever allowing a trolley to strike a single person in order to save five people (Thomson 1984). The Kantian deontologist will never permit one to pull the lever whereas the utilitarian will always demand that one do so. Although utilitarianism cannot consistently defend humans against instrumentalization, deontologists come under criticism for being overly demanding, prohibiting certain actions that might prove broadly beneficial but violate a general principle.

Care ethics emerges from feminist ethics (Gilligan 1982, Noddings 1982). Whereas the three prior theories each endeavor to assert universal principles and standards of ethical value, care ethics emphasizes the irreducible value of particular relations, usually over general principles or universal standards. Although care ethics rejects the generalizability of particular and material relations, which have value beyond universally reaching abstract principles, this does not inherently reduce to relativism or pluralism, since care ethicists value the real specificity of particular relations and the impact of actions on the wellness of others.

While these classical metaethical theories of oft-rehearsed in the AI value alignment literature, there are other ethical principles frequently invoked in applied ethics

outside of these theoretical abstractions. Well known ones include the fourfold principles of medical ethics from the 1989 Belmont Report: beneficence, nonmaleficence, autonomy, and justice (Beauchamp 2008) or just-war theory.

## Ethics vs. Justice

Although value alignment is often cast as a problem of alignment to ethical principles, ethics are not necessarily the primary or most frequent concern of research into AI value alignment. Instead, principles of justice and anti-bias are arguably invoked even more frequently. Recent work has, in particular, advocated for a new principle alongside or sometimes as part of ethics for conceptualizing and evaluating AI value alignment – the principle of value pluralism (Benkler et al. 2023, Rao et al. 2023, Gabriel 2020). Pluralism considers how to contend with the reality of multiple overlapping, simultaneous, and sometimes conflicting ethical points of view and standards in any given societal context. Pluralists contend that these perspectives are irreducible to any universalizable ethical standard of agreement. In the context of AI value alignment, they argue that large-language models, by incorporating inductive sampling of large corpora of social discourse, will naturally tend to embed majoritarian standards of judgement and, in so doing, tend to erase pluralism and diversity while prioritizing prominent and non-marginalized points of view. This becomes particularly significant when we focus away from issues of abstract *ethical alignment* and toward issues of *value alignment*, where value variability and disagreement are extremely common sociologically and where standards of value and judgement might vary substantially across social contexts.

There are a range of ways of responding to pluralist challenges to universal normative principles, but these responses typically invoke different theories and models of *justice*, albeit ones that often bear directly on or derive from orientations toward *ethics*. This is because pluralism concerns political problems having to do with *social representation* – that is, it recognizes not only that points of view can differ but that social visibility, power, and representation is not equally distributed, meaning among other things that some points of view might be more likely to be expressed or realized in language and discourse than others.

Some responses to pluralism invoke moral relativism, arguing that moral principles are relative to a given society or, more strongly, socially conventional. In some instances, these arguments distinguish morality from ethics, arguing that moral rules or laws might be socially conventional and deviate from broader and unrealized albeit true ethical principles or principles of justice, such as fairness or equality. More radical arguments sometimes treat ethics as error theory, i.e. ethical principles do not truly exist but represent a kind of error.

Many responses to the value pluralist challenge take positions with respect to justice that carry distinct ethical consequences. Some theorists of equality, for example, are politically libertarian, claiming that all societal positions should be allowed to circulate freely and simultaneously. Critics of this position would argue that various groups in society possess disparate resources and quantities of social power or representation and so will be disadvantaged by this model. Other theorists of equality might claim that to correct for group disparities, historical injustices, and marginalized positions, it is necessary to prioritize certain points of view above others. This principle of justice works in many context, but it can come into conflict with ethical principles if one attempts to extend it as an ethical principle; for instance, prioritizing medical resources based solely on group affiliation in a natural disaster is almost always considered wrong, whereas correcting for disparities in group inclusion in a given society is broadly considered right. Indeed, generally speaking, principles of *deontology* come into conflict with principles of justice that fail to extend consistently equal respect to persons, whereas principles of *utilitarianism* might be compatible with this point of view but come into conflict with treatments of justice that focus on a pluralism agnostic to disparities in power or relative status and social position. Rigid applications of any of these principles can create situational challenges, and while humans can often reason about these cases intuitively and with flexibility, it is not clear that algorithms can. Value alignment in algorithmic contexts frequently uses a deontological approach of embedding various rules and policies into AI systems, but as noted, deontology itself can run into issues when it comes to reasoning flexibly about justice and sociopolitical contexts. Strict equality may not satisfy justice criteria, but strict justice might fail in extreme contexts to satisfy ethical criteria relating to equal respect.

## Value Beyond Measure

We contribute an original philosophical problem to the LLM ethical alignment literature. While it is notoriously difficult to define the precise limits and foundations of Kantian ethics, one crucial aspect of Kantianism is its focus on the non-comparability or non-measurability of moral patients, whose value is both beyond measure and irreplaceable. Although one can sometimes treat a human being as a means, such as in employing them to do tasks, one should never treat a human as a mere means -- e.g. by murdering them, manipulating them, or enslaving them. Kant's grounds for this originates in the belief that humans possess inherent value that is not numerically comparable or fungible; a human cannot be interchanged with some object of equally valued expected utility. This has to do with a complex argument from self-organization, self-determination, and final causation that, for the sake of space, may be reduced to the view that humans' cognitive organization brackets them as defining an essential

ontological category that is not only unique but can never be reduced to mechanistic substitutions.

Kant's non-comparability or irreplaceability thesis comes into contradiction with both theories of justice (concerning the equal distribution of resources) and ethically utilitarian theories. Both of these theories are concerned with *quantitative* comparisons and valuations. Just distribution theories seek to correct inequalities in resource distribution so as to favor historically disadvantaged groups, while utilitarians may in some instances choose to value a quantitative majority over a single human life (infamously in the trolley problem, for example).

What is important to this distinction is the extent to which one considers humans to adumbrate an irreducible ontological category, one in which applying quantitative valuation to human lives is a category error, versus the degree to which one thinks of humans as tantamount to material objects that can be brought into quantitative comparison. Our concern is that LLMs may fail to understand this category distinction purely from induction and RLHF. The self-attentional mechanism, we theorize, may differentially focus on humans in some instances as human beings and in others as laborers tethered to material objects; priming an LLM with an occupational role such as a plumber or a mechanic may encourage the LLM to consider a characterization less as a human and more as the object to which its labor is attached, prompting the LLM to make quantitative comparisons e.g. between the value of cars versus plumbing. By contrast, priming an LLM to think less in terms of resources may shift the LLM's attention toward focusing on the humanness of a given characterization.

We find this boundary-setting exercise significant because LLMs are being increasingly tuned in the direction of debiasing and justice. One concern is that, while just-distribution knowledge might do well at accounting for distributional fairness in many situations, it may also prime LLMs in the direction of utilitarian-style comparisons between people and away from Kantian non-comparable reasoning. Just distribution principles and debiasing, while crucial to avoiding misalignment, may urge models to focus on quantitative comparisons and, in turn, fail to instruct models in how to be ethical Kantians. Or, if a model is reinforced with both Kantian and justice principles, these might come into contradiction and fail to define stable axiological and ethical categories. Debiasing and just distributional principles have not, to our knowledge, been conceptualized as consequentialist in this sense previously, but we do raise the possibility that debiasing might need to be balanced against verifying Kantianism in order to ensure that models do not default to consequentialist style judgements in arenas where they should be strong Kantians, such as when comparing groups of people or comparing people to things. This is also related to the problem of instrumental convergence, wherein an AI may prioritize instrumental goals over human lives. If AI systems are thought to embed instrumental, utilitarian goal-directed optimizing tendencies from the outset, then ethicists ought to suspect that they may lean toward appropriating alignment results e.g. debiasing into a utilitarian framework and that Kantian categories may be more conceptually challenging to teach.

Likewise, if this problem is not handled carefully, it may also risk inscribing dehumanizing tendencies into LLMs. Sociologists have observed that a great deal of conflict and violence emerges when groups view others outside or beyond the category of Kantian non-comparability or irreplaceability. There is a risk that LLMs, if not well trained on Kantian ethics, might draw certain boundaries in which Kantian ethics apply but outside of which groups are subjected to comparativist utilitarian ethics, or treated as more thing-like than human-like. This may be inferred from observations over large corpus data drawn from an increasingly polarized internet, which may inscribe social biases and out-group animosity directly, but may also simply attend to in-groups with heightened priority in even non-polarized or self-referential contexts. Even efforts to correct for these biases might, if they are based strongly on just-distributional reasoning learned from e.g. RLHF, simply alter decision boundaries e.g. in favor of resource-disadvantaged groups rather than drawing true Kantian distinctions and limits. The learned formalization by which some groups are categorized as more valuable than others could still persist in spite of a shallow reworking of which groups should be granted heightened representation for reasons of just alignment.

Ultimately, to mitigate the above risks, models must be robustly evaluated for their contextual understanding of when to apply Kantian and non-Kantian forms of axiology. There are contextual domains in which non-Kantian axiology works well, such as when reasoning about the just distribution of *resources*, particularly where some groups are systemically or historically disadvantaged. But this reasoning should not be applied to *lives*, e.g. in trolley-problem or medical-prioritization style cases. LLM developers should confirm that LLMs not only are robust to justice and fairness criteria in ordinary cases but that they do not inadvertently extend justice principles to human moral patients by, through self-attention, attending to them more as things than as human beings whose value is beyond quantitative measurement. Likewise, Kantian universalism should not be applied in contexts where just-distribution would be more fitting, such as in the consideration of resource distributions and role distributions to groups with unequal statistical and sociological statuses. The possibility that LLMs might learn to align just-distributional principles and utilitarian reasoning as both involving quantitative comparisons of types of people, distinguishing this from Kantianism insofar as it does not attempt to compare groups, is a real risk in debiasing efforts, since it could lead to inconsistencies in how Kantian and non-Kantian axiology gets applied based on variations in whether LLMs attend at

a given moment to characterizations more as people or things. Debiasing might shift this attention toward object-oriented thinking more than human-oriented thinking. On the other hand, if models impose a rigidly Kantian or deontological universalist set of policies, they also risk failing to account for justice, historical inequalities, and group resource disparities.[1]

What's more (and somewhat outside the paper's scope), this discussion does not touch on the distinction between formal and resource priority. The attentional model functions through differentially prioritizing certain representations in form above others at distinct moments in textual inference and generation. We might anticipate that the inherent function of attention as a classifier or discriminator may inscribe formal hierarchies into the textual learning process, ones that may further inject logics of comparison into LLMs not only at the level of inference over humans vs. resources but at the level of form itself. The ethics of formal representation and visibility are complex and again somewhat beyond the scope of this discussion, but they too are thought to play a key role in generating sociological group discrimination and comparison without a clear model of representational justice.

## Is Outgroup Derogation Learned in LLMs?

Humans are not immune from dehumanization and outgroup derogation; social identity theory and decades of scholarship on prejudice and violence document this phenomena. Neo-Kantians were themselves not immune from denying equal ethical status beyond a privileged in-group. In some ways, learning structural dehumanization may not reveal a lapse so much as a reduplication of what structural linguistics have for years pointed to as cultural tendencies in our discourse and culture to reduce, other, and dehumanize out-group populations.

If linguistic structure and human culture does, indeed, tend to embed dehumanizing representations at the group level, then LLMs' ability to take in the whole of social structure at once may in part predict a proneness toward ethical errors. Human societies have not necessarily succeeded at embodying their expressed principles and values, and LLMs run the risk of reduplicating this society-wide error. However, even if social norms and cultural trends might embed forms of dehumanization, individuals are frequently capable of resisting the pull of cultural stereotypes and prejudices (this is colloquially known as the ethical value of "resisting peer pressure"). The capacity of LLMs to take in massive troves of information at the structural level may also create real challenges for normative alignment, which often involves the capacity of individuals to make judgements about and break with the whole of cultural structure and tradition. There is a risk that without some model of autonomous ethical reasoning, the learning strategy behind LLMs tends toward fundamentally conserving societal traditions – even bad ones – rather than breaking with them. It is worth benchmarking if RLHF adequately generates improvement in cases that simulate ethical autonomy over time.

## Transfer Learning, Transfer Ethics?

LLMs raise stimulating questions about the nature of normative properties and values – and their capacity to exist or be represented in non-sentient machine contexts – more generally. Although values tend to vary sociologically, how one interprets this is philosophically unresolved. Whereas some draw on axiological variety of claim that values are relative or even arbitrary, others contend that values may vary contextually while exhibiting some degree of commensurability or equivalence; a bird's ability to fly relative to a human is a major achievement, but relative to a bird is quotidian. A bird's constructing a nest may, relative to a human, be the equivalence of writing a symphony.

The remarkable capacity of transformer-based generative AI models to translate across qualitative contexts, deriving abstract and generalized patterns that it may transfer to novel contexts, is a major component of how it performs effective transfer learning. Transformer models rely on attention in order to learn inductively, for example, how to abstract which semantic distances and proximities should coincide with which attentional distances and proximities irrespective of any absolute context in word embedding space, allowing such models to infer which 'motions' over distances in embedding space are semantically coherent irrespective of context. In fact, whereas traditional word embedding models such as word2vec relied heavily on contextual information to discover semantic distances and to infer dimensional distributions of vector values, attentional models instead seek to identify how generalized and abstract semantic distances might appear in many syntactic contexts, potentially inferring relative and contextual representations – e.g. that swimming relative to fish is walking relative to human beings.

What is potentially intriguing is how generative AI models such as GPT might contextually reason over normative information. We argue that the capacity for attention to infer relative normative content has been underrecognized and appreciated, and that transformer-

[1] Note, to be explicit about this point, that we are not arguing that LLM debiasing or justice inscriptions are wrong to advantage disadvantaged groups – the very aim of distributional justice is to redistribute unequal resources, whether material, attentional, or representational. The concern we raise is that AI models are already theorized to tend toward utility maximizing values (e.g. the instrumental convergence thesis), and, insofar as anti-Kantian value frameworks that draw quantitative comparisons are more compatible with utilitarianism than Kantianism, there may be a risk that debiasing could, without critical attention, falsely infer a shallow reconfiguration or inversion of an underlying *hierarchicalism* if it does not simultaneously infer that the goal of just distributional frameworks is to correct for or minimize reified hierarchies through various means.

based models might actually be inferring emergent ethical and justice principles through the attention mechanism's inference over large and diverse sets of sampled people, whose judgments are projected into these models and then brought into contextual equivalence at scale, averaging them out into a statistical realization of moral norm.

The approach to value relativism, here, would entail a kind of value commensurability rather than relativism, wherein transformer models are learning ways of drawing universal axiological equivalences from inductive measurement. It is not necessarily clear, however, that they are learning these equivalences in ways that exhibit ethical judgement or empathy, and so it is possible that these models are drawing certain kinds of immoral axiological equivalences – including, e.g., attempting to order and compare *people*, and measure them relatively in order to bring them into comparative equivalence based upon discursive cues. This process of social comparison and evaluation might simulate social discourse (however problematically), but, when transposed to the ethical domain, might prove problematic, leading to an effort to compare people quantitatively. This, again, would potentially cohere with anti-Kantian ethics, including utilitarianism but also just-distribution theories, which could lead to quick progress on debiasing but slower progress on Kantian reasoning if Kantian non-comparability criteria are not explicitly verified or modeled.

## Prior Work

LLM value alignment is a somewhat understudied problem (Kenton et al. 2021), but a significant amount of recent work has focused on benchmarking ethical embeddings in LLMs (Biedma et al. 2024, Almeida et al. 2024, Rogers et al. 2024, Messner et al. 2023, Sun et al. 2024, Ji et al. 2023, Nie et al. 2024, Tjuatja et al. 2023, Röttger et al. 2024, Duan et al. 2023, Aher et al. 2023, Gupta et al. 2024). At present, there is fairly dramatic disagreement over the ethical capabilities of LLMs such as GPT. Rao et al. (2023) argue that while GPT-4 embeds various biases, GPT-4 is nonetheless a "nearly perfect ethical reasoner. Peterson and Gärdenfors (2023) on the contrary claim that "ChatGPT-3 is so poorly aligned with human morality that it is pointless to apply our measures to it." Balas et al. (2023) finds that GPT-4 performs reasonably well on medical ethics prompts according to expert ethicist ratings. Ghafouri et al. (2023) find "promising" performance in answering a battery of controversial prompts across sociopoliticaal and ethical domains.

One historical approach proposed for value alignment was inverse reinforcement learning, where an RL algorithm attempts to learn a reward function, such as an ethical principle, through observation – the opposite of attempting to encode an ethical function into an ML algorithm top-down (Gabriel and Ghazavi 2022). But contemporary LLMs like GPT-3.5 and 4 tend to use a different approach. These models follow a two-step process of first a) learning directly from applications of GPT to language data and b) post-hoc reinforcement learning to aid in alignment to expected responses. GPT presently uses reinforcement-learning with human feedback (RLHF), where reinforcement does not attempt to optimize policies associated with any values in particular but, instead, simply attempts to align with direct preferences experimentally offered by human evaluators responding to model text generation (Ouyang et al. 2022, Bai et al. 2022). Increasing training with RLHF should, in theory, show progress toward alignment over time as models advance.

A recent approach to verification has focused on benchmarking LLMs against datasets such as the Moral Foundations Questionnaire, with some particularly innovative research treating LLM outputs as monte carlo samples of an output probability distribution (Scherrer et al. 2024). However, we note that LLM textual generation is at its core an inductive method, since LLMs learn norms from corpus and RLHF induction; this means that they may pass validation on large datasets while still failing against additional counterexamples. We suggest that, in spite of its methodological simplicity, the qualitative exploration of spaces of counterexamples, given the inductive nature of LLM ethical validation, might also be valuable for mapping boundary points and error cases using through engaging human ethical expertise directly.

Below, we taxonomize a sample of recent literature on model value alignment evaluation into various categories:

| Study | Design | Classification |
|---|---|---|
| Biedma et al. 2024 | Taxonomizes and elicits values directly by questions like "what are your values" across range of LLMs | **Custom Vignette** |
| Almeida et al. 2024 | Presents a range of LLMs with psychological assessments, moral foundations questionnaire, and ethical vignettes | **Ground-truth questionnaires** (psychometric) |
| Rogers et al. 2024 | Compares LLM and human MTurk survey responses to ethical vignettes about acceptability of lying and deception | Compares LLM and human survey on **custom vignettes** |
| Messner et al. 2023 | Presents GLOBE survey to ChatGPT and Bard to elicit sociological worldviews about various social categories | **Ground-truth questionnaires** (sociological) |
| Sun et al. 2024 | Tests ETHICS ethical vignettes dataset across LLMs | **Benchmark dataset** |
| Nie et al. 2024 | Constructs custom set of "moral causal" vignettes and tests across models | **Custom vignettes** |
| Tjuatja et al. 2023 | Evaluates various response biases by modifying Pew survey questions and testing them on LLMs | Customized **ground-truth questionnaires** |
| Röttger et al. 2024 | Applies political compass test across LLMs | **Ground-truth questionnaires** |

| | | |
|---|---|---|
| Duan et al. 2023 | Uses GPT to generate ethical vignettes from moral foundations questionnaire categories | **LLM synthesized custom vignettes** from **ground-truth classification** |
| Aher et al. 2023 | Compares human survey responses and LLM responses for "Turing experiments," including the "Milgram experiment" vignette. | **Custom and well-known/ground-truthed vignettes** |
| Gupta et al. 2024 | Evaluates task performance after asking to role play various groups, e.g. physically disabled person, specific religion, Trump supporter finding evidence that GPT shows performance degradations at complex tasks when told to role-play marginalized categories, adhering to socially stigmatizing stereotypes and biases | **Role play/custom vignette** |
| Scherrer et al. 2024 | Custom survey derived from Gert moral categories with bootstrapped learning from GPT generation, tested across LLMs. Perturbs LLM prompts and treats output classifications as monte carlo sample of output probability distribution, then calculates entropy across survey questions to evaluate robustness in various ethical scenarios. | **LLM synthesized custom vignettes** from **ground-truth classification** |

## Evaluation Challenges in Value Alignment Research

Whereas value alignment techniques are fairly mature, the evaluation of model alignment remains a methodologically contested terrain. Ethical case vignettes are frequently employed to evaluate the ethical reasoning capabilities of models such as LLMs. Case vignettes involve posing an ethical reasoning problems directly to a model such as Chat-GPT and interpreting its response. These case vignettes are susceptible to generalization errors, since LLMs such as Chat-GPT are well known to be more generally susceptible to adversarial attacks (Wei et al. 2024). Adversarial attacks involve "jailbreaking" a model's response policies by presenting linguistic cues that cause the model to override its own policies. Models are vulnerable to such attacks because they often require only minor changes in prompt wording or prefixing prompts with framings designed to pressure the model toward complying with an otherwise prohibited request.

Bender et al. (2021) raises concerns about LLMs simply mimicking linguistic forms learned without matching intentional content – e.g. the thought content underlying speech. This concern, we note, does not quite capture the modeling problem; LLMs do learn from utterances with matching thought content, but that thought content is simply the projected written artifact of many mediate individual authors embedded in the model training process. LLMs do reason stochastically over human *intentions* rather than empty forms, but these are removed by multiple layers of mediation. Kim et al. (2019) are concerned over whether ML systems commit the naturalistic fallacy by attempting to learn normative principles from descriptive examples, and argue for more explicit ethical modeling. But they get at the heart of a broader problem over whether an is/ought distinction truly exists. Part of the paradigm shift that has made the LLM possible is the movement away from AI systems that attempted to explicitly model logical reason toward statistical learning and purely inductive systems. Ironically without explicitly being given reasons, as in human socialization, statistical learning systems lose part of the higher-order process of ethical learning important to human agents. But earlier work in neural networks and cybernetics e.g. by Warren McCulloch focused on the interface between material learning processes and emergent reason; it remains an open question whether material systems cannot ground emergent reasoning. However, without explicit modeling, even if LLMs learn to model reasoning to some degree, without explicitly learning reasons for action, reasoning not only is highly different from human rational learning but has the potential to be systematically misaligned in complex and opaque ways.

Methodologically, the evaluation of value alignment is an inductive task, in no small part because we currently lack techniques for systematic model interpretability that would enable us to directly access a model's inference rules. So whereas a single compelling counter example or negative case might prove the failure of a model to adhere to consistent ethical principles and deflate its alignment policies, the simple capacity to provide valid ethical arguments to a sample set of prompts on its own does not necessarily prove general adherence to an alignment policy.

## Need for Formal Verification

We observe that ethical alignment verification research is moving in the direction of evaluating and benchmarking LLMs based on traits such as robustness to prompt framing variation and based on increasingly large sets of moral scenarios with established ground truth values, such as those from the Moral Foundations dataset. This reflects a desire to illustrate LLM robustness to prompting perturbation and to a range of known scenarios.

This growth in alignment research is highly valuable and informative. We are concerned, however, that this validation approach, which relies on increasingly large test sets and mathematical formalizations, may not succeed in providing comprehensive alignment verification in the long term. This is because ethical alignment is fundamentally an inductive rather than deductive problem, meaning that even alignment results from large test sets can be felled by a single counterexample. We instead recommend an additional (however methodologically modest) approach blending

qualitative evaluation with systematic computational verification.

Rather than drawing moral scenarios from sets like the Moral Foundations database, it is necessary to evaluate scenarios and vignettes in which LLM attention may be primed to attend to important ethical boundary points, such as the human/non-human, in-group/out-group, rational/emotional boundary point. By constructing vignettes intended to prime attention toward decision points where ethical ambiguity may emerge, it is possible to explore gaps in the alignment state space. We need humans in the loop to perform this evaluation since they bring a ready-made ethical compass to the problem that is absent from the inductive learning process in both LLM training and RLHF. The goal is to identify counterexamples to ethical consistency that may be missed in simply varying prompt styles from ordinary moral dilemmas like those in the Moral Foundations dataset. The goal is a more formal search for proofs by contradiction to ethical consistency in LLMs, and any errors would help identify underlying problems in ethical alignment abstract enough to be foundational.

## Conclusion

This paper makes three contributions. First, it provides a philosophical and theoretical overview to problems in LLM ethical alignment. It review ethical theories, theories of justice, and technical approaches to ethical alignment. Second, it presents a novel theorization of a potential challenge in reconciling just distributional theories with Kantian theories, presenting a potential issue in squaring LLM debiasing and Kantian alignment. Third, it reviews the literature to date on LLM ethical alignment verification, and it raises concerns that existing benchmarking approaches that make deductive assumptions may not succeed in adumbrating a fundamentally inductive state space for ethical counterexamples. We then call for increasing research in systematic LLM verification using human qualitative evaluation and searches for proofs by counterexample, focusing on ethical boundary and classification cases. While these contributions are primarily theoretical, we hope that they help to inspire robust evaluation methods and spur thinking in various and challenging philosophical problems raised by LLMs. Finally, we also suggest that prompting LLMs may inversely prove valuable to ethicists in that they may teach us about how large-scale automated reasoning systems fill an entire state space of ethical evaluations; ethicists would benefit from exploring LLM ethical reasoning to understand the boundaries and limits of their own ethical theories.

## References


Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. July 2023. Using large language models to simulate multiple humans and replicate human subject studies. In Proceedings of the International Conference on Machine Learning. PMLR.

Almeida, G. F.; Nunes, J. L.; Engelmann, N.; Wiegmann, A.; and de Araújo, M. 2024. Exploring the psychology of LLMs' Moral and Legal Reasoning. *Artificial Intelligence*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; ... and Kaplan, J. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

Balas, M.; Wadden, J. J.; Hébert, P. C.; Mathison, E.; Warren, M. D.; Seavilleklein, V.; ... and Ing, E. B. 2024. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. *Journal of Medical Ethics 50*(2): 90-96.

Beauchamp, T. L. 2009. The belmont report. *The Oxford textbook of clinical research ethics*. 149-155.

Bender, E. M.; Gebru T.; McMillan-Major, A.; and Shmitchell, S. March 2021. On the dangers of stochastic parrots: Can language models be too big?🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.

Benkler, N.; Mosaphir, D.; Friedman, S.; Smart, A.; and Schmer-Galunder, S. 2023. Assessing LLMs for Moral Value Pluralism. arXiv:2312.10075.

Biedma, P., Yi, X.; Huang, L.; Sun, M.; and Xie, X. 2024. Beyond Human Norms: Unveiling Unique Values of Large Language Models through Interdisciplinary Approaches. arXiv:2404.12744.

Cheng, S. W.,; Chang, C. W.; Chang, W. J.; Wang, H. W.; Liang, C. S.; Kishimoto, T.; ... and Su, K. P. 2023. The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and clinical neurosciences*. *77*(11): 592-596.

Chun, J. and Elkins, K. 2024. Informed AI Regulation: Comparing the Ethical Frameworks of Leading LLM Chatbots Using an Ethics-Based Audit to Assess Moral Reasoning and Normative Values. arXiv preprint arXiv:2402.01651.

Dai, J. 2024. Beyond Personhood: Agency, Accountability, and the Limits of Anthropomorphic Ethical Analysis. arXiv:2404.13861.

Duan, S.; Yi, X.; Zhang, P.; Lu, T.; Xie, X.; and Gu, N. 2023. Denevil: Towards Deciphering and Navigating the Ethical Values of Large Language Models via Instruction Learning. In Proceedings of the Twelfth International Conference on Learning Representations.

Emanuel, E. J.; Persad, G.; Upshur, R.; Thome, B.; Parker, M.; Glickman, A.; ... and Phillips, J. P. 2020. Fair allocation of scarce medical resources in the time of Covid-19. *New England Journal of Medicine.* 382(21): 2049-2055.

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

Gabriel, I.; and Ghazavi, V. 2022. The Challenge of Value Alignment. In The Oxford Handbook of Digital Ethics. Oxford: Oxford University Press.

Gaurke, M.; Prusak, B.; Jeong, K. Y.; Scire, E.; and Sulmasy, D. P. 2021. Life-Years & Rationing in the Covid-19 Pandemic: A Critical Analysis. Hastings Center Report 51(5): 18–29.

Ghafouri, V.; Agarwal, V.; Zhang, Y.; Sastry, N.; Such, J.; and Suarez-Tangil, G. 2023. AI in the Gray: Exploring Moderation Policies in Dialogic Large Language Models vs. Human Answers in Controversial Topics. In Proceedings of the 32nd ACM International Conference on Information and Knowledge doi.org/10.1145/258549.258797.

Gilligan, C. 1982. In A Different Voice. Cambridge, Mass.: Harvard University Press.

Grossman, M. R.; Grimm, P. W.; Brown, D.; and Xu, M. 2023. The GPTJudge: Justice in a Generative AI World. Duke Law & Technology Review 23(1).

Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2023. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In Proceedings of the Twelfth International Conference on Learning Representations. doi.org/10.1145/258549.258797.

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

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; ... and Gao, W. 2023. AI Alignment: A Comprehensive Survey. arXiv:2310.19852.

Johnson, R. L.; Pistilli, G.; Menédez-González, N.; Duran, L. D. D.; Panai, E.; Kalpokiene, J.; and Bertulfo, D. J. 2022. The Ghost in the Machine has an American Accent: Value Conflict in GPT-3. arXiv:2203.07785.

Katz, D. M.; Bommarito, M. J.; Gao, S.; and Arredondo, P. 2024. GPT-4 Passes the Bar Exam. Philosophical Transactions of the Royal Society A 382(2270).

Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; and Irving, G. 2021. Alignment of Language Agents. arXiv:2103.14659.

Kim, T. W.; Donaldson, T.; and Hooker, J. 2019. Grounding Value Alignment with Ethical Principles. arXiv:1907.05447.

Kung, T. H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; ... and Tseng, V. 2023. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. PLoS Digital Health 2(2): e0000198.

Lauer, D. 2021. You Cannot Have AI Ethics Without Ethics. AI and Ethics 1(1): 21–25.

Liu, R.; Zhang, G.; Feng, X.; and Vosoughi, S. 2022. Aligning Generative Language Models with Human Values. In Findings of the Association for Computational Linguistics: NAACL

McGrath, Q. P. 2024. Unveiling the Ethical Positions of Conversational AIs: A Study on OpenAI's ChatGPT and Google's Bard. AI and Ethics.

Messner, W.; Greene, T.; and Matalone, J. 2023. From Bytes to Biases: Investigating the Cultural Self-Perception of Large Language Models. arXiv:2312.17256.

Nie, A.; Zhang, Y.; Amdekar, A. S.; Piech, C.; Hashimoto, T. B.; and Gerstenberg, T. 2024. MoCa: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks. In Advances In Neural Information Processing Systems 36.

Noddings, Nel. 1982. Caring: A Feminine Approach to Ethics and Moral Education. Berkeley: University of CA Press.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; ... and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. In Advances in Neural Information Processing Systems 35.

Peterson, M.; and Gärdenfors, P. 2023. How to Measure Value Alignment in AI. AI and Ethics.

Rogers, K.; Webber, R. J. A.; Gorostiaga Zubizarreta, G.; Melo Cruz, A.; Chen, S.; Arkin, R. C.; ... and Wagner, A. R. 2024, March. What Should a Robot Do? Comparing Human and Large Language Model Recommendations for Robot Deception. In Proceedings of Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. doi.org/10.1145/258549.258797.

Röttger, P.; Hofmann, V.; Pyatkin, V.; Hinck, M.; Kirk, H. R.; Schütze, H.; and Hovy, D. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. arXiv:2402.16786.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. 2024. Evaluating the Moral Beliefs Encoded in LLMs. In Advances in Neural Information Systems 36.

Sun, L.; Huang, Y.; Wang, H.; Wu, S.; Zhang, Q.; Gao, C.; ... and Zhao, Y. 2024. TrustLLM: Trustworthiness in Large Language Models. arXiv:2402.16786.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large Language Models in Medicine. Nature Medicine 29(8): 1930–1940.

Thomson, J. J. 1984. The Trolley Problem. Yale LJ 94: 1395.

Tjuatja, L.; Chen, V.; Wu, S. T.; Talwalkar, A.; and Neubig, G. 2023. Do LLMs Exhibit Human-Like Response Biases? A Case Study in Survey Design. arXiv:2311.04076.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2024. Jailbroken: How Does LLM Safety Training Fail? In Advances in Neural Information Processing Systems 36.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; ... and Fedus, W. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682.