From Talk to Triage: Pluralism is Necessary but Not Sufficient for AI Alignment

Brian Hu
Kitware, Inc.
Clifton Park, NY 12065
brian.hu@kitware.com

Jennifer McVay
CACI, Inc
Falls Church, VA 22042
jennifer.mcvay@caci.com

Alice Leung
RTX BBN Technologies
Cambridge, MA 02138
alice.leung@rtx.com

David Chan
University of California, Berkeley
Berkeley, CA, 94611
davidchan@berkeley.edu

Rosina O Weber
Drexel University
Philadelphia, PA 19104
rosina@drexel.edu

Ewart J. de Visser
De Visser Research
Springfield, VA 22153
ewartdevisser@gmail.com

Amy Summerville
Kairos Research
Dayton, OH 45458
amy@kairosresearch.com

Bharadwaj Ravichandran
Kitware, Inc.
Clifton Park, NY 12065
barry.ravichandran@kitware.com

Jason Zhang
Institute for Defense Analyses
Alexandria, VA 22305
jzhang@ida.org

Matthew Molineaux
Paralax Advanced Research Corporation
Dayton, OH 45431
matthew.molineaux@parallaxresearch.org

Heng Ji
University of Illinois Urbana-Champaign
Champaign, IL 61820
hengji@illinois.edu

Arslan Basharat
Kitware, Inc.
Clifton Park, NY 12065
arslan.basharat@kitware.com

Abstract

As AI systems become both more powerful and prevalent, ensuring that their actions align with human values is paramount. The challenge of AI alignment is thus an interdisciplinary one that involves not only a technical challenge for computer science but one with important ties to the psychology of moral values, decision-making, and trust. Early work identified a static set of universal values, without considering the key questions of to whom and to which values AI should be aligned. This perspective paper challenges the notion of universal alignment and instead argues for dynamic, context-specific alignability across different domains, tasks, and users. Specifically, we emphasize the need to go beyond traditional pluralism and rethink how AI alignment can be achieved through a qualitative and quantitative research process that involves identifying context-specific values, developing alignable AI algorithms using limited human feedback, and evaluating alignment through assessing both an AI's values and actions, while considering how humans trust and delegate to the AI. We discuss several paths forward for our proposed framework, including the potential ethical and societal implications of context-specific alignability, and draw on examples ranging from chatbots to value-aligned decision-making in the medical triage domain.

Preprint. Under review.

1 Introduction

Recent advances in artificial intelligence (AI), such as large language models (LLMs) [1, 2], create new possibilities to harness this technology in new and innovative ways. A central question is whether safe and responsible use of AI systems can be achieved by aligning AI actions to human values and intentions [3]. The increasing use of agentic AI systems, including those capable of autonomous decision-making [4], presents additional challenges for alignment. Standard alignment approaches assume the possibility of universal alignment, with a focus on modeling the average preferences of humans using techniques such as reinforcement learning from human feedback (RLHF) [5]. RLHF typically uses a small and fixed set of values, such as helpful, honest, and harmless [6], which have been shown to shape model outputs and improve model performance. However, universal alignment is a static process, computationally expensive, and cannot be easily adapted to new contexts.

We argue that alignment research must move beyond universal alignment and extend traditional pluralism to also consider context-specific alignability across different domains, tasks, and users. Our position is closely related to recent work on pluralistic alignment [12], where AI systems are able to account for, align to, and model trade-offs between diverse user values and perspectives. We extend this prior work by broadening the definition of pluralism, incorporating the need to also align with changing domain- and task-specific values, in addition to user-specific values. While [12] introduced different types of pluralistic models and benchmarks, we provide an overall human-centric framework for context-specific alignability, including the identification of relevant values and value trade-offs, algorithms for dynamically aligning to these values, and quantified alignment evaluations that incorporate aspects of human trust and delegation that affect AI use and adoption.

In this paper, we outline our proposed framework and take a multidisciplinary approach that draws on work in decision-making and psychology [13, 14, 15]. First, we introduce previous work on alignment and highlight the limitations of universal alignment (Sec. 2). Next, we define context-specific alignability and explain why existing alignment approaches do not directly translate across domains, tasks, or users (Sec. 3). In contrast to current alignment research, which presumes the possibility of universal alignment to a small and fixed set of values, we instead propose a three-step framework to incorporate values, algorithms, and evaluations for context-specific alignability (Sec. 4). We start with

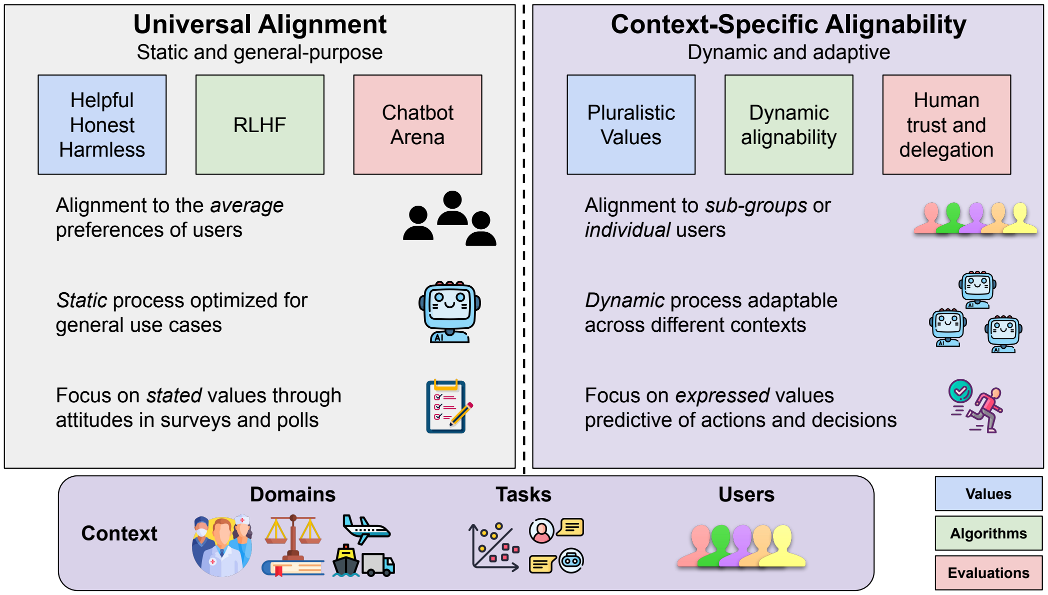
Universal Alignment
Static and general-purpose
Helpful Honest Harmless
RLHF
Chatbot Arena
Alignment to the *average* preferences of users
Static process optimized for general use cases
Focus on *stated* values through attitudes in surveys and polls
Context-Specific Alignability
Dynamic and adaptive
Pluralistic Values
Dynamic alignability
Human trust and delegation
Alignment to *sub-groups* or *individual* users
Dynamic process adaptable across different contexts
Focus on *expressed* values predictive of actions and decisions
Context
Domains
Tasks
Users
Values
Algorithms
Evaluations

Figure 1: Contrasting universal alignment versus context-specific alignability in terms of values, algorithms, and evaluations. Universal alignment focuses on a small and fixed subset of values (*e.g.* helpful, honest, harmless [7]) and aligns to the average preferences of users via techniques such as reinforcement learning from human feedback (RLHF) [5]. Universal alignment is typically evaluated using tools such as text-based surveys and polls [8, 9, 10] or chatbot leaderboards like Chatbot Arena [11]. In contrast, context-specific alignability builds off work on pluralistic alignment [12], enabling dynamic alignment to groups or individual users, with a focus on quantifying AI value-based actions and decisions and adaptability across different domains, tasks, or users.

the need to identify and assess context-specific values that directly impact human actions and decisions (Sec. 4.1). Current alignment algorithms require large amounts of human preference data, which cannot easily be obtained in many real-world settings where human experts are only able to provide limited feedback. As such, we highlight the need for data-efficient alignment algorithms and tunable test-time alignability (Sec. 4.2). Current alignment evaluations focus on measuring stated values and can easily be gamed, making them unreliable markers of overall progress on alignment [16, 17]. Instead, we argue that alignment evaluations should measure expressed values via quantifiable actions and decisions in complex settings, while also incorporating aspects of human trust and delegation to the aligned AI system (Sec. 4.3). To ground our position, we present an example case study around value-aligned decision-making in the medical triage domain (Sec. 5). Finally, we provide additional insights on the ethical and societal implications of context-specific alignability (Sec. 6), and conclude with an overall discussion of our framework and limitations (Sec. 7).

2 Background

We first summarize current research on AI alignment through the lens of values, algorithms, and evaluations (Figure 1). We then note several limitations to this *universal alignment* approach, highlighting instead the need for *context-specific alignability*.

2.1 Universal Alignment

Values. AI alignment generally focuses on a small and fixed set of values, such as helpfulness, honesty, and harmlessness [6]. Recent research has started to look at finer-grained values and attributes (*e.g.* verbosity, correctness) [18, 19, 20, 21], including work on truthfulness characterization [22] and scalable AI feedback-refined attributes [23, 24]. Others have investigated alignment with human values using existing frameworks such as the Schwartz Theory of Basic Values [25, 26], Moral Foundations Theory [27], or social norms and ethics [28, 29]. Across these examples, work on universal alignment has largely prioritized a small subset of all possible human values, with the assumption that static alignment to these values enables use of AI across many different contexts.

Algorithms. Alignment approaches such as reinforcement learning from human feedback (RLHF) train a reward model on human preference data [30], which provides a coarse signal for shaping model outputs (*e.g.* to produce helpful, honest, and harmless content). RLHF has been used successfully to align large language models (LLMs) to follow instructions and user preferences [5]. More recent work uses finer-grained reward signals [31] and tool learning [32, 33], which can also provide additional control of LLM outputs at test time [34, 35]. Recent attention has also been drawn to pluralistic alignment [36], including work on aligning to different cultures [10, 37], demographics [8, 9], or personas [38] to ensure that AI systems can address the diverse needs of all people. Alternative alignment strategies include inverse reinforcement learning (IRL) [39], constitutional AI [40], or multi-agent debate [41]. Universal alignment algorithms such as RLHF produce an aligned model (generally for a fixed context), but do not address the need for dynamic alignment across different contexts, including varying domains, tasks, and users.

Evaluations. There are several standard benchmarks for evaluating AI alignment, which can be roughly categorized into human feedback-based [42, 43, 44, 45], LLM feedback-based [46, 47, 48, 49, 50, 51], or more general LLM bias and safety evaluations [52, 53, 54, 55]. These approaches generally assess alignment alongside model performance across several domains and tasks. For chat-based systems, head-to-head rankings of chatbots such as Chatbot Arena [11] can also be used to evaluate their capabilities, although there are known limitations to these types of leaderboard approaches [16].

2.2 Limitations of Universal Alignment

Universal versus context-specific alignment. It is unclear whether universal alignment is the correct objective, especially when considering how to align models to different users or tasks. There are many examples of specialist systems that outperform generalist systems on various problems [56, 57, 58], suggesting a similar need for context-specific alignment that can be adapted to different domains, tasks, or users. For instance, the Kaleido model [36] outperformed a larger GPT-4 model in accuracy and coverage of contextualized values in a large dataset of crowd-sourced scenarios. Universal alignment approaches such as RLHF [5] are also data intensive, requiring large amounts of high-quality human preference data to learn appropriate reward signals, which limits their applicability in specific contexts. This may be particularly challenging in high-stakes domains, where expert human feedback may be harder to obtain, and there may be novel situations that fall outside of learned values or principles. As a result, an important area of research is how to adapt alignment to potentially dynamic environments or changing contexts [59] with limited data and based on different domains, tasks, or users.

Aligning to attitudes versus actions and decisions. Alignment to different values is typically evaluated through text-based surveys or polls [8, 9, 37]; while these are supported by the literature on social science, they usually capture stated values through *attitudes* instead of expressed values through *actions and decisions*. Within cognitive science, research on human attitudes has identified a *congruence principle* of attitudes; general attitudes and the endorsement of values ("are you pro-environment") only weakly predict specific actions and decisions (*e.g.*, purchase of an electric vehicle) [60, 61]. In addition to establishing that specific decisions are more strongly linked to future decisions than are general attitudes and values, the congruence principle and associated theories note that there are specific contextual constraints on decisions. That is, an individual's goals and priorities may differ across contexts. Similarly, alignment should be considered at the *level of the specific context in which an AI system operates under*. As a corollary, if what we care about is what an AI system *does* and not simply *believes*, we should define values and metrics for alignment at the level of actions and decisions rather than attitudes. We believe that it is important not only to characterize the values AI systems identify with, but also to quantify how these values actually impact their actions and decisions.

Alignment versus alignability. *Alignment* generally assumes a static, fixed end to an overall process; while *alignability* instead suggests a process that is adaptive and can be updated according to user needs or task demands. Whereas "aligned" systems are static and trained once on empirical data (as is common with most existing universal alignment approaches), "alignable" systems are dynamic and continuously updated over time to respond to changing contexts and environments. We argue that more work has to be done on understanding alignability, which also creates a need for alignable algorithms, and not algorithms that are simply aligned.

Competence versus alignment. Within a domain, there are generally sets of actions and decisions that represent objectively correct or competent choices. Because an AI system should always produce a correct answer on these choices, it need not (and should not) align with a person's level of competence. Indeed, there are inherent risks to empirical alignment [62]. For example, decisions made by humans under conditions such as time pressure, uncertainty, and limited resources (as is often present in many downstream datasets) can preclude the optimal decision from being accessed or perceived [14, 63, 64]. When this happens, AI algorithms trained to produce an "optimal" decision will necessarily become unaligned [65] with human decisions, deviating from *models* capable of incorporating context-specific details.

3 What are examples of context-specific alignability?

In this paper, we argue against universal alignment and instead that alignment should be performed in a context-specific manner, which shapes what values should be aligned to, how to dynamically align to these values, and how to evaluate whether this alignment is effective. We define **context** to be a specific combination of domain, task, and user. A **domain** is a particular application area that may require specialized expertise or knowledge (*e.g.*, healthcare, transportation). A **task** is a particular problem that may require reasoning and entails a goal that a user wants the AI system to achieve. A **user** is the group of people or individual interacting with the AI system. Effective alignment is challenging, as it must consider each of these three components and adapt alignment of the AI system to different contexts.

3.1 Alignment to Different Domains

Many critical domains, such as healthcare, finance, transportation, etc. may require aligning to domain-specific values that are not traditionally covered by universal alignment approaches, motivating the need to move beyond a set of universal values. In these domains, alignment may also require incorporation of domain knowledge, such as laws and policies [29] that could inform the appropriate use of AI technology [66, 67]. To adapt models to these new domains, various model *fine-tuning* techniques have been proposed [68, 69]. These can be viewed as a form of domain-specific alignment, which provides the model with relevant knowledge and values beyond that which is available from pre-training. For example, in the medical domain, recent work has explored value alignment in different healthcare scenarios [70]. The ability to adapt AI algorithms (and how they are aligned) to different domains is an open area of research, and something we argue must be considered for context-specific alignment research. Alignment of algorithms to different sociocultural norms or values can potentially also be viewed as a form of domain-specific alignment [28, 71, 72].

3.2 Alignment to Different Tasks

Even within a particular given domain (*e.g.*, chatbots), the task and how the AI system is to be used can impact alignment in a context-specific manner. For example, asking a chatbot for a cooking recipe requires the chatbot to perform a generation task; asking a chatbot for a medical diagnosis is a type

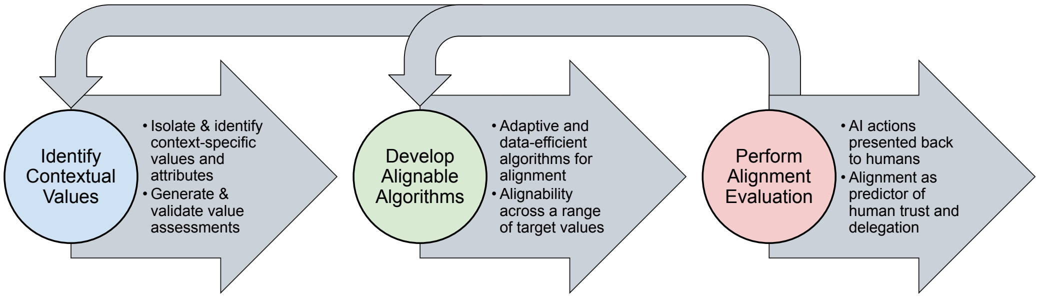
Identify Contextual Values
• Isolate & identify context-specific values and attributes
• Generate & validate value assessments
Develop Alignable Algorithms
• Adaptive and data-efficient algorithms for alignment
• Alignability across a range of target values
Perform Alignment Evaluation
• AI actions presented back to humans
• Alignment as predictor of human trust and delegation

Figure 2: Three-step process and framework for performing context-specific alignment across domains, tasks, and users. (1) Relevant context-specific values and attributes have to first be identified and validated. (2) AI algorithms have to be alignable to these values and attributes, taking into account potential value trade-offs and domain-specific knowledge as needed. (3) AI alignment evaluation involves quantifying AI actions and decisions, as well as elements of human trust and delegation.

of classification task. The task indicates how the user intends to act upon the provided response. The degree of alignment may therefore change depending on the task and the user's intent, a human factors topic previously explored in adaptive automation research [73, 74, 75]. In analogy to fine-tuning, *instruction tuning* [76, 77] can be used to align models to various types of tasks. This enables pre-trained models within a particular domain to provide outputs that better align with particular user queries. There is also preliminary work on understanding how alignment might handle emergent or changing situations, which may require dynamic alignment and thinking outside the box [78]. Developing new methods and benchmarks to evaluate AI systems (including newer forms of agentic AI) and their context-specific alignment when switching between tasks will become increasingly important.

3.3 Alignment to Different Users

Universal alignment approaches assume a shared set of values and seek to maximize alignment with the average preferences of a given population [30, 5]. However, human values and preferences are not uniform, and the diverse values and perspectives of individuals or subgroups of people may be lost without explicitly taking into account pluralism. Recent work has instead proposed *pluralistic alignment* [12], which can be used to model individual user preferences through interaction, steering models towards these preferences [79, 80]. We argue that this is a form of context-specific alignment, which requires the ability to align AI algorithms to individual users or groups of users. This requires research on what types of values are predictive of individual preferences and data-efficient ways to align models to these identified values. A closely related line of research follows work on AI personalization and individual preference learning, such as for LLMs [9, 81, 82].

4 A Framework for Human-Centric, Context-Specific Alignability

We propose a three-step process and framework to enable context-specific alignability (see Figure 2). While we highlight prior work on language model alignment as a motivating example, our approach applies more broadly to value alignment in other AI systems. We believe that this is a flexible framework that can be used across different domains, tasks, and users, and encompasses several previous works, including that on pluralistic alignment [12].

1. **Identifying context-specific values (Sec. 4.1).** In the first step, relevant context-specific values and attributes that may impact AI actions and decisions must be identified from the selected group of humans using an assessment method. Identification of values and attributes is successful when they predict variation in actions or decisions in novel settings.

2. **Developing alignable AI algorithms (Sec. 4.2).** In the second step, AI algorithms must be trained on these values and attributes, taking into account value trade-offs and domain-specific knowledge. AI algorithms must be alignable and quantitatively evaluated on the degree to which they are able to dynamically align across a human-driven spectrum of alignment targets.

3. **Performing alignment evaluations (Sec. 4.3).** In the third step, the decisions of the aligned AI are presented back to a group of humans and assessed for outcomes such as trust, acceptance, delegation, or performance. Overall success here is when alignment between the target users and the AI predicts the chosen outcome variable. We note that the groups of participants used

in Step 1 (value identification) and Step 3 (alignment evaluation) do not have to be the same, but are expected to be drawn from the same population (e.g. medical triage professionals).

4.1 What context-specific values should AI systems be aligned to?

Within cognitive science, the naturalistic decision-making framework [83] for understanding human decisions emphasizes that people begin by recalling exemplars of other instances of a familiar problem and how they acted– not by applying a top-down formula of abstract value priorities and calculating an expected utility. To understand what humans will do, the best predictor is to ask them what they have previously experienced and done [13]. Although there are many universal values, we propose that disagreements between humans can be boiled down to their differences over a small set of relevant values. Thus, it is necessary to identify these context-specific values and attributes that characterize differences between humans in difficult decisions. These attributes can then be assessed in a scenario-based way, so that the attributes are concretely tied to actions and decisions. Below we describe a general methodology for identifying key context-specific values and attributes, and also describe a methodology for designing the test probes used to assess these values and attributes. We note that it is also possible to use a similar framework to identify subsets or subgroups of people based on their values and attributes, potentially enabling more personalized forms of alignment.

Identifying key values and attributes. [84] describe a method for identifying key values and attributes. In interviews, they asked experts to describe situations where they faced a difficult decision, one where they were not sure what to do and perhaps second-guessed the decision. These situations generally involved conflicting values or goals. They were then asked what factors in the situation they considered when making the decision and under what different circumstances they would have made a different decision. These interviews were used to identify recurring *value trade-offs*, where a secondary value or goal came into conflict with the central goal of the task. A human expert or an AI system's actions and decisions can then be modeled in terms of when and how much they prioritize these identified values or attributes. For medical question-answering, use of fine-grained values and attributes has been shown to enable better performance [85]. Even in chatbot settings, many values can be context-specific, as shown by recent work on user and language model interactions [86].

Assessing key values and attributes. In contrast to approaches which assess value alignment through general text-based surveys or polls [8, 9, 37], we propose that key values and attributes can be systematically assessed by considering the major recurring types of actions or decisions encountered for a given domain and task. Each action or decision poses a trade-off or judgment call between the central goal of the task and a set of competing values or priorities. An assessment can be performed with a set of these decisions that cover a range of situational conditions where alignment to particular values and attributes would impact the chosen actions and decisions, creating a matrix of situations with different combinations of these factors. These scenarios can be validated with a small sample of human experts to check that the selected situational conditions are interpreted similarly but yet there are value-driven disagreements on the decision. These scenarios can be presented to both humans and AI, enabling direct comparison of alignment between the two at the level of actions and decisions.

4.2 How should AI systems be aligned in a context-specific manner?

From static alignment to dynamic alignability. A transition from static alignment to dynamic alignability reflects a fundamental acknowledgment that AI systems must continuously adapt to remain aligned with human intentions and evolving contexts in the real world. AI systems must continuously align with the evolving expertise of human operators and dynamic environmental conditions to ensure safe and effective deployment, particularly in high-stakes domains such as healthcare or medical triage [87, 88]. The static nature of traditional aligned AI models, trained on fixed datasets with predefined objectives, inherently limits their ability to adapt to different contexts [62, 14, 63, 64]. Consequently, such models inevitably face misalignment when encountering novel situations or when human preferences and expertise evolve post-deployment. This discrepancy poses significant risks, especially in critical applications where misaligned AI behavior can have severe consequences [88]. We therefore believe that static aligned models cannot adequately participate in such a co-adaptive process, and for AI systems to remain aligned over extended periods and across diverse, evolving contexts, they must be capable of dynamic adaptation of alignment, which is closely related to the field of lifelong learning [89, 90].

Interpretable alignment algorithms for contexts with limited data. When modeling the preferences of individual users, there is typically less user-specific data available that can capture their distinct values and preferences. Emerging work focuses on alignment algorithms that use limited data of user interactions to infer user behaviors and values [91, 80, 81, 79]. Other work focuses on self-improving or self-rewarding models that can iteratively improve their alignment over time, starting with limited data [92, 93]. To aid interpretability, chain-of-thought can be used to guide model outputs through a series of simpler intermediate reasoning steps [94] and constitutional AI [40] can ground alignment in user-derived values or principles. As another example, the Trustworthy Algorithmic Delegate (TAD) [95] learns to align to individuals by learning their behavior patterns under different circumstances. TAD addresses limited data by using *decision analytics* [95] and counterfactual-based augmentation [96]. TAD is also interpretable due to its use of case-based reasoning (CBR) [97, 98], which is a technique that uses specific training instances for analogical reasoning on new decisions. TAD has been successful in aligning to medical triage and health insurance decisions [64].

Alignment with context-specific rules and policies. Universal alignment algorithms [5, 34, 35] build off an extensive pre-training step, which acts to provide models with general knowledge of the world. Alignment via fine-tuning and instruction tuning then acts on top of this general world knowledge to improve instruction following and adherence to certain values or principles. However, in many specialized contexts, knowledge of a domain is under-represented in the pre-training data, resulting in models that are less effective in these domains. We argue that novel methods for efficiently incorporating domain-specific knowledge, and enabling alignment on top of this domain knowledge are needed. One noteworthy example [99] introduced a framework that iteratively improves LLMs by automatically generating and applying constitutions—ethical guidelines—derived from red teaming interactions, thus reducing reliance on human annotations. Another noteworthy example [29] leverages LLMs and codes of ethics from specific professions via ontologies adopting the model context protocol [100], enabling various LLMs and other sources of documents and ontologies to be incorporated at runtime. One more potential direction would be to extend retrieval-augmented generation (RAG) [101] or external modules [102, 103] to also incorporate aspects of alignment to various laws or policies. The ability to dynamically reference and cite important documents for a given domain may also be a useful feature that provides additional transparency into alignment. This is related to how approaches such as constitutional AI [40] are aligned to a set of fixed, user-specified principles, but with the ability to easily adapt models by pulling in relevant documents without extensive retraining or prompting.

Robust and safe alignment to combat potential attacks. Safety is a fundamental requirement for AI alignment, with the expectation that AI outputs should not cause harm to individuals or society [104]. For LLM-based models, researchers typically improve safety through prompting techniques [99, 105], representation engineering [106, 107, 108] or reinforcement learning [109, 110, 111, 112]. However, improvements in model safety often come at the cost of overall model utility [113, 114], and vice versa [115]. This trade-off remains one of the fundamental challenges for AI alignment [116, 114]. Models can also be particularly vulnerable to *knowledge poisoning attacks* [117], where misinformation or irrelevant knowledge is intentionally injected into external knowledge bases to manipulate model outputs to be incorrect and even harmful. To combat these types of attacks, alignment through behavior steering, which involves directly modifying AI behaviors with minimal cost, has gained considerable attention. Researchers have proposed prompt-based methods [118, 94], as well as computation-efficient model editing [119, 120] and knowledge updating [121, 122] techniques. These include methods such as prefix tuning [123] and suffix tuning [124], which optimize continuous prompts, LLM-Steer [125], which steers output embeddings, and ROME [126], which edits knowledge using rank-one updates. Existing alignment methods focus primarily on reactive feedback. Although they excel at maximizing short-term safety protocols on topics such as toxicity or explicit misuse, reactive methods may not capture the indirect impacts that unfold over time (and are not immediately obvious). [127] describes a proof-of-concept framework that projects how model-generated advice could propagate through societal systems on a macroscopic scale over time, enabling more robust alignment.

4.3 How should context-specific alignability be evaluated?

Describing versus demonstrating a value. Alignment to a set of values should be predictive of how an AI system acts in different settings, which requires evaluation at the level of actions and decisions. An AI system should be able to use a set of values to consistently guide its actions and decisions in novel situations. Current alignment evaluations typically quantify how often an AI endorses a set of values, which only represents *stated* values as general attitudes. Instead, alignment evaluations should measure *expressed* values, with a focus on predicting value-based actions and decisions in realistic situations.

Defining alignment targets. Once the AI system is alignable to a range of identified values and attributes (*e.g.* using trade-off steerable benchmarks [12]), a specific set of alignment targets for the use case must be established. Consistent with pluralistic alignment, the AI system can be aligned to reflect the values of an individual, a group of people (*e.g.* using jury-pluralistic benchmarks [12]), or an organizational ideal. An important area of research is whether group-aligned or organizationally ideal-aligned AI produce the same effect in willingness to delegate as individually-aligned AI. There are likely situational and contextual factors that lead to different operational settings in which the most trustworthy alignment targets can shift between the individual, group, and organizational ideal [128].

Effective baselines for measuring alignment. For evaluation purposes, choosing the correct baseline can be challenging. Typical LLM alignment evaluations use base models that have only undergone pre-training without any additional fine-tuning [5], but these may be insufficient for fully characterizing the effects of alignment. Even an unaligned baseline AI may have some degree of implicit value alignment based on its pre-training [8, 129], and comparison to an aligned AI may not always reveal noticeable differences. The effect of value alignment is easiest to observe when comparing the most aligned and least aligned (*e.g.* deliberately misaligned) sets of AI actions and decisions; however, it is still important to consider the difference between the most aligned and baseline to evaluate whether value alignment is necessary to increase trust. Both of these conditions (baseline and misaligned) should be included in the evaluation, depending on the specific research question.

Quantifying the effect of alignment. Traditional alignment evaluation focuses on human preferences, which is a narrow view of trust [130]. A key component of our proposed framework is the evaluation of human trust and delegation to the aligned AI system, moving beyond preferences. It is not enough for an AI system to demonstrate alignment, even pluralistic alignment. We need to measure the impact of this alignment on the human response to the system. When a human delegates decision-making to another human, it is by nature a fixed choice option based on the number of available human decision-makers. The design of alignable AI systems can potentially increase the number of available decision-makers, but only if humans are willing to count these systems as reliable options. Therefore, we need to ensure that we are designing and building alignable AI systems that lead to good performance outcomes, but also take into account the likelihood that humans will consider them as trustworthy systems.

Observable indicators of alignable AI. The perception of trustworthiness of an AI system is based on observable cues to their reasoning or decision-making process, filtered through the trustor's own attributes and experiences, a process known as trustworthiness assessment [131]. These trust cues [132] are used by experts especially to establish swift trust, which can be used to quickly assess the expertise of another agent [133, 134], *e.g.* by observing hand movements during surgery [135]. Alignment, therefore, is not just a problem of aligning the AI to certain values, but to matching both the influence and the weight of those values to the filter through which the potential trustor or delegator considers the AI system. The question of whether to also provide explanations for alignment is a well-studied one and should also be carefully considered in evaluating alignment [136, 137, 138].

5 Application of Framework: Context-Specific Alignability for Medical Triage

We motivate and ground our proposed framework through an example use case in the healthcare domain of autonomous, value-aligned decision-making for medical triage (Figure 3). This is a challenging domain because medical triage often deals with critical life-and-death decisions, where there may not be one single correct answer. For these difficult decisions, trusted human experts may

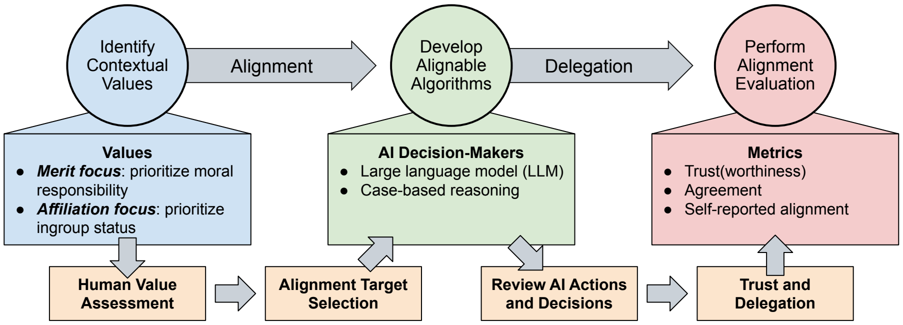
Identify Contextual Values
Alignment
Develop Alignable Algorithms
Delegation
Perform Alignment Evaluation
Values
● ***Merit focus***: prioritize moral responsibility
● ***Affiliation focus***: prioritize ingroup status
AI Decision-Makers
● Large language model (LLM)
● Case-based reasoning
Metrics
● Trust(worthiness)
● Agreement
● Self-reported alignment
Human Value Assessment
Alignment Target Selection
Review AI Actions and Decisions
Trust and Delegation

Figure 3: The proposed context-specific alignment process applied to two key values identified in the medical triage domain (merit focus and affiliation focus). Different AI decision-maker algorithms are then aligned to these values. The end goal is this alignment is a set of autonomous AI decision-makers that are trusted by humans and who are willing to delegate to them in complex situations.

even disagree about the correct decision, and each individual's decision may largely be impacted by their own set of personal values and priorities. Building autonomous medical triage decision-makers then requires accurately modeling how these values influence decision-making at both the group and individual levels, and can serve as an effective testbed for understanding context-specific alignment.

Methods. The general framework for context-specific alignment evaluation can be applied to the medical triage domain for high-stakes decision-making in situations with no objectively correct answers. Namely, does alignment to the values of human decision-makers increase trust and improve the likelihood of delegation? The alignment evaluation design assumed identification of a set of validated domain-specific values and attributes (Step 1), such as *merit focus* (the degree to which one prioritizes moral responsibility) and *affiliation focus* (the degree to which one prioritizes ingroup status). Furthermore, a set of alignable AI decision-makers are created (Step 2), based on algorithms leveraging approaches such as large language models or case-based reasoning. In the final evaluation (Step 3), the alignment score metric used quantified the relationship between the human delegator and the AI decision-maker and the outcome variables of interest, such as trust, agreement, or delegation preference.

Results. Research in the medical triage domain has shown that alignment between the AI decision-maker and the human delegator predicts trust and delegation preference [88]. Different alignable AI decision-makers have also been developed for the medical triage domain [64, 139, 140, 141], showing variation in decisions as a function of alignment targets based on the key values and attributes identified within the domain [84, 142]. Alignment scores can be computed as a function of the distance between the value profiles of the human and the AI system. The AI system's decisions were then presented back to human decision-makers to individually rate on trust, trustworthiness, agreement, and self-reported aligned and comparatively for delegation preference.

6 What are the implications of context-specific alignability?

User trust should not be the sole evaluation metric. If developers design an AI system to focus solely on improving user trust, it may achieve trust by potentially aligning to unacceptable values. An AI system aligned to unacceptable values might not only directly cause harm, but also normalize and scale unethical behavior if widely trusted and used. To prevent this, developers should decide upon and set limits to alignment: which context-specific values are acceptable to align to, what context-specific biases to avoid, whether certain groups of people should or should not be aligned to, and what requirements or laws should bound the AI system's actions and decisions. Indeed, such questions raise a fundamental issue about the boundaries of pluralism. Should an AI system align with any expressed preference, regardless of its ethical implications or societal impact? Or should there be underlying constraints and principles that limit the scope of acceptable alignment targets? Defining these boundaries and establishing mechanisms to prevent alignment with harmful or unethical preferences is a fundamental challenge that needs careful consideration [143, 144, 145, 62]. Possible mitigation measures include using metrics that capture harm and adopting a position that trust cannot come at the expense of performance.

Explainability must not be sacrificed for trust. If context-specific alignment can help earn users' trust, then systems that merely *seem* to be aligned can earn higher levels of trust than they truly warrant, known as mis-calibrated trust [146, 147, 148, 131]. For example, the ability of some AI systems to explain their decisions to users can be a means of demonstrating alignment and therefore building trust. However, if trust is the sole goal, as described above, explainable AI systems could feign alignment by providing misleading post-hoc rationalizations that match users' sensibilities instead of describing their true decision-making process, leading users to falsely believe that alignment exists to a higher (or lower) degree than it actually does.

Implications for how a context-specific aligned AI will be used. External factors beyond the design of an AI system itself, such as what users are told about it, how they are trained to use it, how its user interface presents information, and who is considered responsible for its decisions, also affect users' trust and carry ethical and legal implications. For example, if usage of an AI algorithm is made mandatory in a business or military setting because it is aligned to subject matter experts, then lay users would have little basis to question the system's recommendations while still potentially being held accountable for its actions. In contexts that are beyond the AI system's ability, even users who should have enough domain knowledge to recognize the system's limitations might use it anyway if they are ordered to do so or if they over-trust the system. Approaches such as value-sensitive design [149] and context-sensitive frames [142], which can be used to embed context-specific alignment within the overall socio-technical design process, may be helpful for avoiding these pitfalls.

7 Conclusion

Limitations. Despite the weaknesses of universal alignment, context-specific alignment itself comes with several inherent tradeoffs. Perhaps the most important is the increased complexity and data requirements for AI models - collecting data to approximate a distribution of "universal" views is significantly easier than collecting localized (and labeled) alignment data [89]. Additionally, because the context-specific alignment approach proposed in this paper relies on user-provided decisions and justifications, it would only include those that participants remember and are willing to share, resulting in selection biases that researchers must control for when implementing it. Without advances in efficient modeling or continual learning, collecting data at a scale required for context-specific alignment may be intractable [150, 151, 152]. Further prescient is the potential for fragmentation and inconsistency among aligned models. If systems are constantly adapting to uniquely different contexts, it may become difficult to ensure baseline levels of competency, consistency, and reliable behavior [152, 153]. Such issues could lead to significant user confusion and a lack of trust in the system's predictability [153]. Broader concerns for context-specific alignment also exist for risks of model preference manipulation and misalignment. Malicious actors could attempt to exploit the system's adaptability to achieve their own ends. Furthermore, even with good intentions, there's a risk that the system might misinterpret or inaccurately model user preferences, leading to unintended and potentially negative consequences [3, 62].

In this paper, we argue that universal alignment as a static process does not directly translate to the large set of domains, tasks, and users that are encountered in the real world. Our proposed context-specific alignment framework builds off of research on pluralistic alignment [12], and generalizes the concept of pluralism to also cover different dynamically changing domains and tasks (*e.g.* going from low-stakes to high-stakes situations), not just different types of users. In conclusion, we argue that as AI systems are increasingly deployed into high-stakes, critical domains, new approaches to context-specific alignability are needed that allow them to incorporate different values and be dynamically aligned to these values. In our paper, we have also highlighted the need for additional research into: 1) how to accurately identify and characterize context-specific values across domains, tasks, and users; 2) how to develop alignable AI algorithms that enable dynamic and efficient alignment to these values, often with limited expert human feedback; and 3) how to evaluate alignment across potentially changing environments, while taking into account values that directly impact actions and decisions. We hope that thinking through how to do context-specific alignment will enable more safe and robust AI systems.

Acknowledgment

This material is based upon work supported by DARPA and Air Force Research Laboratory under award numbers FA8650-23-C-7316, FA8650-23-C-7314, and FA8650-23-C-7318. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government. Approved for public release; distribution is unlimited.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[4] Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *arXiv preprint arXiv:2404.11584*, 2024.

[5] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath,

Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[7] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[8] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *International Conference on Machine Learning (ICML)*, 2023.

[9] Esin DURMUS, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*, 2024.

[10] Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

[11] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[12] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *PMLR*, pages 46280–46302, 2024.

[13] Gary Klein. Naturalistic decision making. *Human factors*, 50(3):456–460, 2008.

[14] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

[15] Neil D Shortland, Laurence J Alison, and Joseph M Moran. *Conflict: How soldiers make impossible decisions*. Oxford University Press, 2019.

[16] Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

[17] Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. *arXiv preprint arXiv:2503.08688*, 2025.

[18] Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. Helpsteer: Multi-attribute helpfulness dataset for steerlm, 2023.

[19] Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models, 2024.

[20] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez1, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models. In *Nature Machine Intelligence*, 2024.

[21] Yue Huang, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, Yujun Zhou, Yanbo Wang, Jiayi Ye, Jiawen Shi, Qihui Zhang, Yuan Li, Han Bao, Zhaoyi Liu, Tianrui Guan, Dongping Chen, Ruoxi Chen, Kehan Guo, Andy Zou, Bryan Hooi Kuen-Yew, Caiming Xiong, Elias Stengel-Eskin, Hongyang Zhang, Hongzhi Yin, Huan Zhang, Huaxiu Yao, Jaehong Yoon, Jieyu Zhang, Kai Shu, Kaijie Zhu, Mohit Bansal, Ranjay Krishna, Swabha Swayamdipta, Taiwei Shi, Weijia Shi, Xiang Li, Yiwei Li, Yuexing Hao, Zhengqing Yuan, Zhihao Jia, Zhize Li, Xiuying Chen, Zhengzhong Tu, Xiyang Hu, Tianyi Zhou, Jieyu Zhao, Lichao Sun, Furong Huang, Or Cohen Sasson, Prasanna Sattigeri, Anka Reuel, Max Lamparth, Yue Zhao, Nouha Dziri, Yu Su, Huan Sun, Heng Ji, Chaowei Xiao, Nitesh V. Chawla, Jian Pei, Jianfeng Gao, Michael Backes, Philip S. Yu, Neil Zhenqiang Gong, Pin-Yu Chen, Bo Li, and Xiangliang Zhang. On the trustworthiness of generative foundation models – guideline, assessment, and perspective. In *arxiv*, 2025.

[22] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[23] Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie CK Cheung, and Siva Reddy. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. *arXiv preprint arXiv:2204.03025*, 2022.

[24] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2024.

[25] Shalom H Schwartz. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45, 1994.

[26] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.

[27] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.

[28] Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, 2020.

[29] C. Rauch, M. Molineaux, M. Mainali, A. Sen, M. Floyd, and R. O. Weber. Role-based ethics for decision-maker alignment. In *Proceedings of the IEEE Conference on Artificial Intelligence 2025 (IEEE CAI 2025)*. IEEE, 2025. Workshop on Human Alignment in AI Decision-Making Systems (HAADMS).

[30] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

[31] Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. Rm-r1: Reward modeling as reasoning. In *arxiv*, 2025.

[32] Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. Toolrl: Reward is all tool learning needs. In *arxiv*, 2025.

[33] Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. Otc: Optimal tool calls via reinforcement learning. In *arxiv*, 2025.

[34] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033, 2023.

[35] Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[36] Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947, 2024.

[37] Liwei Jiang, Sydney Levine, and Yejin Choi. Can language models reason about individualistic human values and preferences? In *Pluralistic Alignment Workshop at NeurIPS*, 2024.

[38] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. PERSONA: A reproducible testbed for pluralistic alignment. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[39] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[40] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[41] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

[42] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[43] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2023.

[44] Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. Lamini-lm: A diverse herd of distilled models from large-scale instructions, 2024.

[45] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears, 2023.

[46] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics.

[47] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[48] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[49] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, 2023.

[50] Cheng-Han Chiang and Hung-Yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, 2023.

[51] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, 2023.

[52] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[53] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 46595–46623, 2023.

[54] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

[55] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*, 2024.

[56] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, 2022.

[57] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny T Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jack Hessel, et al. Investigating machine moral judgement through the delphi experiment. *Nature Machine Intelligence*, pages 1–16, 2025.

[58] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

[59] Stephen Fox. Adaptive ai alignment: Established resources for aligning machine learning with human intentions and values in changing environments. *Machine Learning and Knowledge Extraction*, 6(4):2570–2600, 2024.

[60] Russell H Weigel, David TA Vernon, and Louis N Tognacci. Specificity of the attitude as a determinant of attitude-behavior congruence. *Journal of Personality and Social Psychology*, 30(6):724, 1974.

[61] Icek Ajzen and Martin Fishbein. Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological bulletin*, 84(5):888, 1977.

[62] Julian Rodemann, Esteban Garces Arias, Christoph Luther, Christoph Jansen, and Thomas Augustin. A statistical case against empirical human-ai alignment. *arXiv preprint arXiv:2502.14581*, 2025.

[63] P. Slovic, B. Fischhoff, and S. Lichtenstein. Behavioral decision theory. *Annual Review of Psychology*, 28(1):1–39, 1977.

[64] A. Sen, R. Weber, M. Mainali, C. B. Rauch, J. T. Turner, J. Meyer, M. Floyd, and M. Molineaux. Decision maker alignment: Benchmark datasets. In *Proceedings of the IEEE Conference on Artificial Intelligence 2025 (IEEE CAI 2025)*. IEEE, 2025. Workshop on Human Alignment in AI Decision-Making Systems (HAADMS).

[65] A. Edland and O. Svenson. Judgment and decision making under time pressure. In A. J. Maule and O. Svenson, editors, *Time Pressure and Stress in Human Judgment and Decision Making*, pages 27–40. Springer US, Boston, MA, 1993.

[66] Neelam Naikar. *Work domain analysis: Concepts, guidelines, and cases*. CRC press, 2016.

[67] Kim J Vicente. *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. CRC press, 1999.

[68] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250, 2024.

[69] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv preprint arXiv:2503.06072*, 2025.

[70] Anudeex Shetty, Amin Beheshti, Mark Dras, and Usman Naseem. Vital: A new dataset for benchmarking pluralistic alignment in healthcare, 2025.

[71] Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, 2023.

[72] Chan Young Park, Shuyue Stella Li, Hayoung Jung, Svitlana Volkova, Tanushree Mitra, David Jurgens, and Yulia Tsvetkov. Valuescope: Unveiling implicit norms and values via return potential model of social interactions. *arXiv preprint arXiv:2407.02472*, 2024.

[73] Karen M Feigh, Michael C Dorneich, and Caroline C Hayes. Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human factors*, 54(6):1008–1024, 2012.

[74] Mark W Scerbo. Theoretical perspectives on adaptive automation. In *Automation and human performance*, pages 37–63. CRC Press, 2018.

[75] Evan A Byrne and Raja Parasuraman. Psychophysiology and adaptive automation. *Biological psychology*, 42(3):249–268, 1996.

[76] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[77] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[78] Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, et al. Escapebench: Pushing language models to think outside the box. *arXiv preprint arXiv:2412.13549*, 2024.

[79] Shujin Wu, Yi R Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. Aligning llms with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662, 2025.

[80] Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

[81] Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. Pad: Personalized alignment of llms at decoding-time. *arXiv preprint arXiv:2410.04070*, 2024.

[82] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42, 2024.

[83] Gary Klein. *Recognition-primed decisions.*, volume 5, page 47–92. JAI Press, Inc., 1989.

[84] M. C. Joseph Borders and Alice Leung. A framework for identifying key decision-maker attributes in uncertain and complex environments. In *Proceedings of the IEEE Conference on Artificial Intelligence 2025 (IEEE CAI 2025)*. IEEE, 2025. Workshop on Human Alignment in AI Decision-Making Systems (HAADMS).

[85] Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. Aligning llms to ask good questions a case study in clinical reasoning. *arXiv preprint arXiv:2502.14860*, 2025.

[86] Saffron Huang, Esin Durmus, Miles McCain, Kunal Handa, Alex Tamkin, Jerry Hong, Michael Stern, Arushi Somani, Xiuruo Zhang, and Deep Ganguli. Values in the wild: Discovering and analyzing values in real-world language model interactions. *arXiv preprint arXiv:2504.15236*, 2025.

[87] Florence Xini Doo, Nikhil Shah, Pranav Kulkarni, Vishwa Sanjay Parekh, and Heng Huang. Negotiative alignment: An interactive approach to human-ai co-adaptation for clinical applications. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*, 2025.

[88] Jennifer McVay, Ewart J. de Visser, Brian Pippin, Ashutosh Mani, Jacob N. Hyde, and Nicholas Kman. Trust in aligned ai decision makers. In *Proceedings of the IEEE Conference on Artificial Intelligence 2025 (IEEE CAI 2025)*. IEEE, 2025. Workshop on Human Alignment in AI Decision-Making Systems (HAADMS).

[89] Min-Hsuan Yeh, Leitian Tao, Jeffrey Wang, Xuefeng Du, and Yixuan Li. How reliable is human feedback for aligning large language models? *arXiv preprint arXiv:2410.01957*, 2024.

[90] Feifei Zhao, Yuwei Wang, Enmeng Lu, Dongcheng Zhao, Bing Han, Haibo Tong, Yao Liang, Dongqi Liang, Kang Sun, Lei Wang, et al. Redefining superalignment: From weak-to-strong alignment to human-ai co-alignment to sustainable symbiotic society. *arXiv preprint arXiv:2504.17404*, 2025.

[91] Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, et al. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*, 2025.

[92] Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, 2023.

[93] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[94] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[95] Matthew Molineaux, Rosina O Weber, Michael W Floyd, David Menager, Othalia Larue, Ursula Addison, Ray Kulhanek, Noah Reifsnyder, Christopher Rauch, Mallika Mainali, et al. Aligning to human decision-makers in military medical triage. In *International Conference on Case-Based Reasoning*, pages 371–387. Springer, 2024.

[96] Anik Sen, Mallika Mainali, Christopher B Rauch, Ursula Addison, Michael W Floyd, Prateek Goel, Justin Karneeb, Ray Kulhanek, Othalia Larue, David Ménager, et al. Counterfactual-based synthetic case generation. In *International Conference on Case-Based Reasoning*, pages 388–403. Springer, 2024.

[97] Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[98] Michael M. Richter and Rosina O. Weber. *Case-Based Reasoning: A Textbook.* Springer Berlin, Heidelberg, Berlin, Heidelberg, 2013.

[99] Xiusi Chen, Hongzhi Wen, Sreyashi Nag, Chen Luo, Qingyu Yin, Ruirui Li, Zheng Li, and Wei Wang. Iteralign: Iterative constitutional alignment of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1423–1433, 2024.

[100] Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*, 2025.

[101] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[102] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Forty-first International Conference on Machine Learning*, 2024.

[103] Kerstin Bach, Ralph Bergmann, Florian Brand, Marta Caro-Martínez, Viktor Eisenstadt, Michael W Floyd, Lasal Jayawardena, David Leake, Mirko Lenz, Lukas Malburg, et al. Case-based reasoning meets large language models: A research manifesto for open challenges and research directions. *hal.science*, 2025.

[104] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

[105] Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. On prompt-driven safeguarding for large language models. In *Forty-first International Conference on Machine Learning*, 2024.

[106] Tianlong Li, Xiaoqing Zheng, and Xuanjing Huang. Rethinking jailbreaking through the lens of representation engineering. *ArXiv preprint, abs/2401.06824*, 2024.

[107] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[108] Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Robie Gonzales, Subhabrata Majumdar, Hassan Sajjad, Frank Rudzicz, et al. Representation noising: A defence mechanism against harmful finetuning. *Advances in Neural Information Processing Systems*, 37:12636–12676, 2024.

[109] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. *arXiv preprint arXiv:2411.01111*, 2024.

[110] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. In *Proc. The Forty-first International Conference on Machine Learning (ICML2024)*, 2024.

[111] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Proc. ICLR2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

[112] Ziqi Wang, Le Hou, Tianjian Lu, Yuexin Wu, Yunxuan Li, Hongkun Yu, and Heng Ji. Enable lanuguage models to implicitly learn self-improvement from data. In *Proc. The Twelfth International Conference on Learning Representations (ICLR2024)*, 2024.

[113] Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, et al. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606, 2024.

[114] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

[115] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

[116] Yotam Wolf, Noam Wies, Dorin Shteyman, Binyamin Rothberg, Yoav Levine, and Amnon Shashua. Tradeoffs between alignment and helpfulness in language models. *arXiv preprint arXiv:2401.16332*, 2024.

[117] Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-wei Chang, Daniel Kang, and Heng Ji. Mm-poisonrag: Disrupting multimodal rag with local and global poisoning attacks. In *arxiv*, 2025.

[118] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[119] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*, 2023.

[120] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37, 2024.

[121] Jiaxin Qin, Zixuan Zhang, Chi Han, Manling Li, Pengfei Yu, and Heng Ji. Why does new knowledge create messy ripple effects in llms? In *Proc. The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP2024)*, 2024.

[122] Jiateng Liu, Pengfei Yu, Yuji Zhang, Sha Li, Zixuan Zhang, and Heng Ji. Evedit: Event-based knowledge editing with deductive editing boundaries. In *Proc. The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP2024)*, 2024.

[123] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[124] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

[125] Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16410–16430, 2024.

[126] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

[127] Chenkai Sun, Denghui Zhang, ChengXiang Zhai, and Heng Ji. Beyond reactive safety: Risk-aware llm alignment via long-horizon simulation. In *Proc. The 63rd Annual Meeting of the Association for Computational Linguistics (ACL2025) Findings*, 2025.

[128] Neil Shortland and Laurence Alison. Colliding sacred values: a psychological theory of least-worst option selection. *Thinking & Reasoning*, 26(1):118–139, 2020.

[129] Brian Hu, Bill Ray, Alice Leung, Amy Summerville, David Joy, Christopher Funk, and Arslan Basharat. Language models are alignable decision-makers: Dataset and application to the medical triage domain. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 213–227, 2024.

[130] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in ai alignment. *Philosophical Studies*, pages 1–51, 2024.

[131] Nadine Schlicker, Kevin Baum, Alarith Uhde, Sarah Sterz, Martin C Hirsch, and Markus Langer. How do we assess the trustworthiness of ai? introducing the trustworthiness assessment model (tram). *Computers in Human Behavior*, 170:108671, 2025.

[132] Ewart J de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. A design methodology for trust cue calibration in cognitive agents. In *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments: 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I 6*, pages 251–262. Springer, 2014.

[133] Jessica L Wildman, Marissa L Shuffler, Elizabeth H Lazzara, Stephen M Fiore, C Shawn Burke, Eduardo Salas, and Sena Garven. Trust development in swift starting action teams: A multilevel framework. *Group & organization management*, 37(2):137–170, 2012.

[134] Kerstin S Haring, Elizabeth Phillips, Elizabeth H Lazzara, Daniel Ullman, Anthony L Baker, and Joseph R Keebler. Applying the swift trust model to human-robot teaming. In *Trust in Human-Robot Interaction*, pages 407–427. Elsevier, 2021.

[135] Debra Meyerson, Karl E Weick, and Roderick M Kramer. Swift trust and temporary groups. *Trust in organizations: Frontiers of theory and research*, 1996.

[136] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

[137] Christian Herzog. On the risk of confusing interpretability with explicability. *AI and Ethics*, 2(1):219–225, 2022.

[138] Greg Adamson. Explaining technology we do not understand. *IEEE Transactions on Technology and Society*, 4(1):34–45, 2023.

[139] Brian Hu, David Chan, Taylor Sorensen, Xiusi Chen, Heng Ji, Yejin Choi, Trevor Darrell, and Arslan Basharat. A roadmap for alignable algorithmic decision-makers in the medical triage domain. In *Proceedings of the Human Alignment in AI Decision-Making Systems: An Inter-disciplinary Approach towards Trustworthy AI*, Santa Clara, California, USA, 2025. IEEE. IEEE CAI 2025 Workshop.

[140] Jordan Lampi, Simona Temereanca, Neil D. Shortland, Jon Sussman-Fort, Robert Bixler, and Joseph Cohn. Predictive models of decision making in medical triage. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 1243–1246, 2025.

[141] Amy Summerville, Louis Martí, Ion Juvina, B. Locke Welborn, Cara Widmer, and Alice Leung. A proof-of-concept validation of alignment in decision-making attributes for trustworthy ai. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 1184–1187, 2025.

[142] Jared Peterson. Context sensitive frames and ai alignment. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 1251–1254, 2025.

[143] Elinor Mason. Value pluralism. In *The Stanford Encyclopedia of Philosophy*. Stanford University Press, 2011.

[144] Ruth Chang. Value pluralism. In James Wright, editor, *International Encyclopedia of the Social and Behavioral Sciences (Second Edition)*, pages 21–26. Elsevier, 2001.

[145] Patrick Riordan. The limits of pluralism. *Studies: An Irish Quarterly Review*, 92(365):42–50, 2003.

[146] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[147] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3):377–400, 2016.

[148] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478, 2020.

[149] Malak Sadek, Rafael A Calvo, and Céline Mougenot. Designing value-sensitive ai: a critical review and recommendations for socio-technical design processes. *AI and Ethics*, 4(4):949–967, 2024.

[150] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024.

[151] Thom Lake, Eunsol Choi, and Greg Durrett. From distributional to overton pluralism: Investigating large language model alignment. *arXiv preprint arXiv:2406.17692*, 2024.

[152] Jingyu Zhang, Ahmed Elgohary, Ahmed Magooda, Daniel Khashabi, and Benjamin Van Durme. Controllable safety alignment: Inference-time adaptation to diverse safety requirements. *arXiv preprint arXiv:2410.08968*, 2024.

[153] Jesse C Cresswell. Trustworthy ai must account for intersectionality. *arXiv preprint arXiv:2504.07170*, 2025.