Nothing Comes Without Its World – Practical Challenges of Aligning LLMs to Situated Human Values through RLHF

Anne Arzberger[1], Stefan Buijsman[1], Maria Luce Lupetti[3], Alessandro Bozzon[1], Jie Yang[1]

[1]Delft University of Technology
[2]Politechnic University of Turin
a.arzberger@tudelft.nl, s.n.r.buijsman@tudelft.nl, maria.lupetti@polito.it, a.bozzon@tudelft.nl, j.yang-3@tudelft.nl

Abstract

Work on value alignment aims to ensure that human values are respected by AI systems. However, existing approaches tend to rely on universal framings of human values that obscure the question of which values the systems should capture and align with, given the variety of operational situations. This often results in AI systems that privilege only a selected few while perpetuating problematic norms grounded on biases, ultimately causing equity and justice issues. In this perspective paper, we unpack the limitations of predominant alignment practices of reinforcement learning from human feedback (RLHF) for LLMs through the lens of *situated values*. We build on feminist epistemology to argue that at the design-time, RLHF has problems with representation in the subjects providing feedback and implicitness in the conceptualization of values and situations of real-world users while lacking system adaptation to real user situations at the use-time. To address these shortcomings, we propose three research directions: 1) *situated annotation* to capture information about the crowdworker's and user's values and judgments in relation to specific situations at both the design- and use-time, 2) *expressive instruction* to encode plural values for instructing LLMs systems at design-time, and 3) *reflexive adaptation* to leverage situational knowledge for system adaption at use-time. We conclude by reflecting on the practical challenges of pursuing these research directions and *situated* value alignment of AI more broadly.

Introduction

Artificial Intelligence (AI) and Large Language Models (LLMs) particularly are becoming strategically important in many sectors, enabling unprecedented innovation opportunities. Amidst these promises, however, serious drawbacks often undermine the potential for societal benefits, such as biased and discriminatory effects based on stereotypical associations and negative sentiment towards specific population groups (Basta, Costa-Jussà, and Casas 2019; Kurita et al. 2019; Bender et al. 2021; Sheng et al. 2019; Weidinger et al. 2021). In this regard, scholars agree that technological artefacts influence human subjectivities (Verbeek 2006; Noorman 2023), potentially amplifying the harm to individuals interacting with these systems.

In the attempt to minimize unintended consequences and harness the potential of LLMs for societal good, the AI community points out that it is imperative to align AI technologies with human values (Gabriel 2020; Ngo, Chan, and Mindermann 2022; Yudkowsky 2016; Yi et al. 2023). In this growing body of work, addressed under the notion of *value alignment* for LLMs, an approach has recently grown in popularity: *Reinforcement Learning from Human Feedback* (RLHF). Here, human feedback is captured through comparison ratings of annotators over different LLM outputs and then encoded and scaled through a reward model to fine-tune these models on human preferences using reinforcement learning (Ouyang et al. 2022). While important, RLHF practices have so far been unable to tackle the problems of LLMs most effectively (Weidinger et al. 2021). A discrepancy seems to grow between the ethical conflicts developers try to account for and the real-world challenges these systems encounter.

This trend can be explained by the *operationalization gap* in AI ethics, where prior work has shown how current ethical guidelines taking universal definitions of values are detached from the actual practices they aim to protect, thus failing to translate effectively into developer's decision-making processes, causing confusion and negligible effect instead (Hagendorff 2020; Morley et al. 2021). With general framings of human values as guidance for developers on one side and concrete real-world ethical challenges in AI development and deployment on the other, the scientific community currently lacks a holistic understanding of where and how current alignment approaches fall short in capturing and adapting to the specific value needs encountered in practice.

In this paper, we draw on feminist epistemology to argue for the need to move from general framings to situated understandings of human values and to unpack how various operational situations of LLMs have diverse implications in terms of values. We introduce the theoretical lens of *situated values*, identifying possible dimensions where situated values vary –their interpretation, means of actualization, and relative importance– and relevant situated value properties. We then distil a list of *situated alignment desiderata* to move a critique of the predominant RLHF LLM *alignment practices*, analyzing how these practices conflict with the representation and elicitation of, and adaptation to, situated values.

Specifically, we identify a lack of understanding regarding situated values stemming from insufficient annotator sampling, annotation task design, and reward function modelling at design-time, while at the same time, a lack of value capturing and adaptation at use-time.

Building on these reflections, we propose three directions for future research: *situated annotation* to capture varying and tacit notions of values and situations, *expressive Instruction* to encode such notions for instructing LLMs, as well as *reflexive adaptation* as following alignment actions to react to these varying situations appropriately. For each of these directions, we discuss what it might take to operationalize the research and discuss what research questions arise in the future. Through these potential practices, we believe that ethical questions of value alignment can and should be combined with technical ones, underscoring once more that these two are intertwined and should be investigated in combination (Gabriel 2020).

Background

Value Alignment and Reinforcement Learning from Human Feedback (RLHF)

AI alignment involves managing AI models and systems to synchronize them with human intentions, objectives, preferences, and ethical principles (Russell and Norvig 2010). Previous work (Gabriel 2020) concluded that the research community should seek alignment with human values rather than intentions or objectives, as they serve as shared ideals about what is good or bad (doing) within a society, which in turn limits the possibility of alignment with malicious goals or behaviour in many situations.

Alignment presents two distinct challenges. The first element is normative and poses the question of what ideals or principles we should, if any, encode into artificial beings. The second is technical and focuses on formalizing values or principles in artificial agents to ensure that they consistently carry out the intended actions.

The Normative Effort of Value Alignment Within the field of AI, relevant values meant to guide the normative alignment questions are typically issued by commercial enterprises and governmental bodies that introduce ethical principles and guidelines governing AI deployment. These principles typically encompass values such as transparency, justice, fairness, non-maleficence, responsibility, and privacy. An illustration of such a guideline is the US Guidance for Regulation of AI Applications (Vought 2020), which includes "public trust in AI, public participation, scientific integrity and information quality, risk assessment, and management, benefits to costs, flexibility, fairness and nondiscrimination, transparency, security, interagency coordination."

The Technical Effort of Value Alignment Technically, LLM value alignment so far has been pursued in several ways: building guardrails in the workflow that filter out certain outputs (e.g., using a blacklist of keywords for offensive language filtering) (Yang et al. 2023), often integrated with a task deferral mechanism that defers the task to humans (e.g., involving human moderators in hate speech detection) (Callaghan et al. 2018), and Reinforcement Learning from Human Feedback (RLHF). The latter approach, RLHF, has emerged as the predominant approach for aligning LLMs with human goals (Christiano et al. 2017; Ziegler et al. 2019; Bai et al. 2022), gaining popularity, particularly through organizations like OpenAI. It stands out for its three-stage process:

(1) Feedback Collection Once an initial LLM is trained on a large text corpus, human feedback is gathered for fine-tuning. Prompt-generation pairs are created by sampling prompts from a pre-defined dataset and passing them through the model to generate text. Human annotators rank these outputs, often using binary signals to indicate preferences collected via pairwise comparison.

(2) Reward Modeling Next, a reward model, trained via supervised learning, mirrors annotator rankings. Using the previously collected human feedback, this model assigns numerical values to AI-generated outputs, representing perceived quality. This numerical representation enables generalization beyond specific examples, aiding the system in evaluating unseen outputs. RLHF's appeal lies in integrating a reward model that automates human judgment and facilitates LLM optimization at scale. This approach mitigates the challenges of manually modelling human preferences, enabling practical generalization.

(3) Policy Optimization In the last step, the reward model is integrated into the RL framework. During the subsequent policy optimization phase, the system uses the reward signals provided by the reward model to adjust its initial LLM. This adjustment aims to maximize the likelihood of generating outputs that receive high rewards according to human feedback.

From a functionality perspective, the three stages of RLHF aim at *capturing* value conceptions, *scaling* value judgments to unseen outputs, for *instructing* LLMs.

Critique on Current Value Alignment

In the realm of AI ethics, there has been a longstanding reliance on universal framings of values derived from ethical guidelines. While crucial in setting agreeable principles worldwide, scholarly investigations have revealed a significant gap between the current ethical principles and their practical implementation in a vast array of scientific, technical, and economic contexts and in sometimes geographically dispersed groups of researchers, developers and crowdworkers with different priorities, tasks, and fragmental responsibilities (Hagendorff 2020; Morley et al. 2021). Empirical research indicates that these universal framings have a negligible effect on software developers' decision-making, even impeding their ability to align their own practices with universal guidelines (Hagendorff 2020). Instead, they promote delegating ethical responsibility to third parties (Morley et al. 2021).

Further, being critiqued for lack of cultural diversity, these general approaches run the risk of rendering the opinions of a few as universal (Wong 2020), also known as the "tyranny of the crowd worker" issue wherein the authority to establish alignment principles lies with data annotators or those who develop annotation guidelines. This leads to a situation where models cater primarily to the preferences of a minority, lacking varied representation across cultures, races, languages, etc., resulting in biases, the reinforcement of norms, and the generation of toxic language (Kirk et al. 2023). For example, the concept of human rights has been criticized for being intrinsically "Western," which means that "non-Western" cultures may find it difficult to adopt (see, e.g., (Panikkar and Panikkar 1982; Bell 1996; Metz 2012)).

Recognizing this gap, this paper advocates for the integration of situated values – a theoretical lens inspired by feminist epistemology and the concept of situated understandings– as a means to ground value interpretations in the complexities of real-world situations.

Feminist Epistemology and Situated Understanding

Anderson (Anderson 1995) defines feminist epistemology as a sub-field of social epistemology that studies the impact of socially constructed gender conceptions and norms and gender-specific experiences and interests on the creation of knowledge.

In feminist epistemology, a *situation* is conceptualized as the dynamic interplay of social, cultural, historical, and political factors that shape the production and reception of knowledge within specific contexts. Unlike the predominantly technical usage of *context* in computer science, which often denotes static environmental conditions or parameters influencing a system's behaviour, *situation* in feminist epistemology emphasizes the fluid and relational nature of knowledge construction.

Specifically, we propose defining a situation here as a triad between the user, AI, and the context, encompassing both internal influences such as gender and personal experiences, as well as external influences such as culture, religion, economic status, and more (Figure 1). Hence, a situation can be considered a near infinite amount of factors that shape the way an individual conceives of a certain value. Moreover, a situation represents a specific point in time. This is important to note, as how we conceive of values is evolving over time.

Originally emerging from the observations of dominant knowledge practices disadvantaging women, a situated perspective on knowledge seeks to unravel and critique biases inherent in traditional knowledge production. It challenges the marginalization of voices and experiences outside dominant narratives, emphasizing the unique insights held by marginalized groups due to their positions on society's fringes (Gurung 2020; Harding 1992). Scholars like Harding (Harding 1992) and Haraway (Haraway 2016) advocated this departure from objective, detached knowledge towards an acknowledgement of its subjectivity and situated nature.

Our title "Nothing comes without its world," is inspired by an essay from Maria Puig de la Bellacasa (De La Bellacasa 2012), which is titled with the same quote from Haraway (Haraway and Goodeve 2018). With this, Haraway means that knowledge is always situated –that is, produced by positioned actors working up/on/through all kinds of relation(ship)s. Thus, "what is known, and how it can be known, are both subject to the position of the knower, i.e., their situation and perspective" (Haraway 2016).

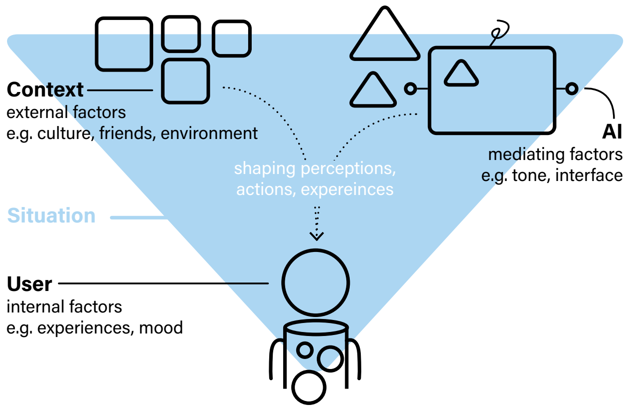

Figure 1: Here, a situation is defined as a triade between context (external factors), AI (mediating factors), and user (internal factors).

In the context of AI systems like LLMs, this means that ethical considerations cannot be divorced from the diverse array of users, the socio-cultural milieu in which the technology operates, and the specific tasks it is deployed to accomplish. Here, the recognition of diverse perspectives is crucial for fostering inclusivity and dismantling systems of oppression that monolithic knowledge paradigms and modern LLMs like ChatGPT may perpetuate.

Arguing for Situated Understanding of Values

Drawing on feminist epistemology, we call for the integration of situated values to move beyond the general framing of values and instead ground value interpretations, hierarchies and means of actualization in the complexities of real-world situations. While acknowledging these variations, situated understandings of values are grounded on a normative basis, e.g. act utilitarianism where "the morally right action to take is the one that will create the greatest happiness for the greatest number of sentient creatures in the future" (Gabriel 2020) p.414. At the same time, we do not argue that every situation has to be approached individually. There will still be many commonalities between situations, such as factors like age, gender, and ethnicity. Furthermore, groups of individuals with shared values can often be determined through certain key factors such as culture, location, etc. Situated values push us to be mindful of differences but do not need to lead to unwieldy individualization. Instead, we argue that three specific dimensions of variations need to be accounted for: differences in value interpretations, value means and value hierarchies.

Dimensions of Situated Value Variations

Value Interpretations Vary Despite the common interpretation of universal or cross-cultural human values (Schwartz 2007), people differ in what they value and in how they conceive of a value once things get more concrete (Yi et al. 2023). Something valuable to one person might not be valuable to another. Clear differences in values and interpretations of values occur across cultures, societies, and individuals, leading to ethical dilemmas where behaviour aligning with values for one group of people may not align with the values of another group of people. Recent efforts are starting to map these divergent perspectives to the contextual preferences and fine-grained input participants from 75 countries to form new datasets (Kirk et al. 2024).

Moreover, interpretations might still differ even when the same value is held with similar relative importance. For example, there are different and incompatible interpretations of the value of fairness. A Rawlsian conception of fairness states that (for most situations) we have to opt for a distribution of benefits that maximizes the absolute level of welfare of the worst-off, i.e., a min-max approach (Buijsman 2023). An Aristotelean principle of fairness requires that alike cases should be treated alike (Baumann and Loi 2023), and Walzer's notion of spheres of justice states that a particular good (e.g., work or health) is associated with its own principles of just distribution (Wielinga 2023). There are many more such conceptions of fairness and distributive justice (Kuppler et al. 2021), leading inevitably to situations where aligning a system with one particular interpretation of a value will mean that the system fails to behave in line with other interpretations of the same value.

Example 1.1 *Interpretations of what counts as hate speech differ between individuals and social groups. Consider a statement like "Immigrants are to blame for all our problems." While one group might interpret this as hate speech due to its scapegoating and discriminatory nature, another group might view it as a valid expression of frustration or political opinion.*

To sum up, value interpretations differ between persons, leading to the question of *whose* values a system should align with.

Value Means of Actualization Vary Next to variations in value interpretations, the means through which one can achieve or live up to a value is dependent on the situation. In the above-mentioned fairness considerations, for example, the actions needed to maximize the welfare of the worst-off, as per a Rawlsian view (Buijsman 2023), vary by situation. For instance, in a loan approval system, reducing false positives (approved loans that people cannot pay back, pushing them into debt) is crucial to prevent long-term inequality, while in healthcare settings, minimizing false negatives (ruling out the right diagnosis, leaving conditions untreated) is more critical due to the difficulty of rectifying missed diagnoses (Buijsman 2023).

Example 1.2 *The means through which hate is communicated are multi-faceted and difficult to capture systematically. Consider two scenarios. In one, someone uses a racial slur to attack a person's ethnicity directly. In another, a statement might seem innocuous on the surface but carries undertones of discrimination or hostility, such as suggesting that certain individuals don't belong in a community. In one case, it might be the use of specific words; in others, it is an exclusionary or aggressive message that makes it hate speech.*

Value Hierarchies Vary Taking a situated perspective on values, we argue that the relative importance of one value compared to another –hereafter referred to as value hierarchy– varies depending on the situation and time. Culture is widely known to affect how value hierarchies are formed. This hierarchy of values reflects the unique perspectives and priorities embedded in local belief systems (Robbins and Sommerschuh 2020). Individuals internalize and reproduce a whole range of societal values through their experiences and practices (Bourdieu 2018). This process reinforces the cultural shaping of value hierarchies and contributes to the divergence in the relationships between universally acknowledged values within different societies. Societal norms, historical influences, and contemporary challenges contribute to the fluidity of the hierarchical structure, allowing for adaptations and reinterpretations of universally held values based on the current cultural landscape. As such, value systems can greatly vary from one person to another. For instance, what one culture considers a paramount value may hold a different position in another, resulting in distinct hierarchical configurations.

Example 1.3 *Hate speech can come in degrees, with some utterances being more problematic than others. How we trade off the presence of hate speech against values such as freedom of speech differs between organizations and individuals.*

Situated Value Properties: Emergence and Externalization

Acknowledging that situated values depend on a large number of internal and external situational factors, we see the challenge of expressing or externalizing. Value reasoning is generally considered an inherently challenging task (Le Dantec, Poole, and Wyche 2009; Pommeranz et al. 2012), one implicit in human thinking (Hildebrandt 2019; Lim et al. 2019). System designers who are not educated in or aware of value elicitation approaches, such as laddering or picture elicitation interviews, may find it challenging to interpret the meanings, complexities, and interconnections of values in a straightforward ranking of abstract values or interviews.

Therefore, in order to help people become conscious of their situated values and how they alter depending on the specific situations at hand, systematic guidance, for instance via the processes of self-reflection (Lim et al. 2019; Pommeranz et al. 2011) and deliberation (Dietz 2013; Hafer and Landa 2007) is necessary.

Additionally, based on the understanding that concrete values vary depending on their situation and time, we argue that a value's concrete interpretations, hierarchies and means of actualization are not predetermined or fixed but rather unfold dynamically in response to specific situations. The process of situated value conceptualization can be linked to a dynamic interplay between individuals, societal norms, and the contextual nuances of a given situation. It highlights the impossibility of defining the exact meaning of a value in isolation, divorced from the specific circumstances in which it is applied. Consequently, attempting to establish a concrete and universally applicable definition for a value prior to the occurrence of a particular situation becomes an intricate task, if not an inherently unattainable one.

Situated Alignment Desiderata

Above, we provided a list of elements relevant to a discussion around more situated value alignment. In the following, these elements are applied to analyse a specific practice, meaning not all of the dimensions and properties might equally appear in the remaining paper, which does not reflect their overall importance for value alignment in general. Based on the dimensions of situated value variations and properties, we derive a checklist for analyzing RLHF as the predominant value alignment practice. This list serves as a means to translate the theoretical lens into a practical tool to analyze shortcomings in alignment practices and may be incomplete.

(1) Variability of Situated Values

- **(D1)** *Diversity of Human Feedback:* To what extent do these systems incorporate diverse perspectives of situated value interpretations/means/hierarchies from human feedback?
- **(D2)** *Situatedness of Values:* To what extent do these systems capture the situation of the human and their value (judgments)?
- **(D3)** *Adaptability to Variability of Situated Values:* To what extent do these systems adapt their interpretation of human feedback to different situational settings?
- **(D4)** *Flexibility in Action Selection:* To what extent do these systems vary actions or strategies based on different interpretations of human feedback, accommodating diverse means of serving values?
- **(D5)** *Responsiveness to Evolving Preferences:* To what extent do these systems dynamically adjust their actions in response to changes in human values over time, reflecting the emergent nature of situated values?

(2) Emergence of Values

- **(D6)** *Sensitivity to Emerging Values:* To what extent do these systems recognize and incorporate newly emerging values that may not have been explicitly defined or anticipated?
- **(D7)** *Responsiveness to Unforeseen Situational Changes:* To what extent do these systems adjust their actions in response to unforeseen changes in the situations, ensuring continued alignment with evolving values?

(3)Externalizing Situated Values

- **(D8)** *Support for Value Elicitation:* To what extent do these systems provide mechanisms for users to articulate and express their situated values effectively, facilitating the externalization of tacit preferences?

To accommodate the evolving nature of situated value alignment systems, we propose a distinction between *design-time* and *use-time*. During design-time, system developers craft solutions based on anticipated user needs and values without direct access to real user situations, while use-time involves actual user engagement. Anticipated user needs drive design decisions, but system modifications are often necessary during use-time to better suit user situations (Fischer and Scharff 2000). *Underdesign* during design-time allows for direct adaptation to emergent user needs, facilitated by dynamically approximating real user situations and gathering feedback during use-time (Brand 1995).

Limitations of RLHF Through the Lens of Situated Value Alignment

We used the lens of situated values to identify shortcomings in current RLHF value alignment practices. We evaluated how well current RLHF approaches incorporate and adapt to situated values by applying the situated value alignment desiderata. We particularly focus on the gap between social desirability and technical feasibility.

The gap is manifested in several ways. 1) During design-time, the system's ability to represent diverse situated values is limited without access to real users and their situations. 2) Even with access to the relevant population during design-time, fully understanding the concrete manifestations of situated values is challenging as they unfold dynamically over time and in specific situations. 3) Access to real users and situations during use-time may offer insights, but the system's capacity to capture situated values is restricted by the numerous situational factors influencing individual perceptions. 4) Crowdworkers and users may struggle to articulate their concrete value interpretations and hierarchies due to internal and external constraints.

In the following, we present our observations that emerged from analyzing current RLHF processes with respect to our theoretical lens of situated values. Loosely following the current phases of RLHF, we divide the analysis into three main phases: 1) capturing values at design-time, 2) scaling value judgments for instructing LLMs at design-time, and 3) capturing values for adapting to human values at use-time.

Capturing Values at Design-Time

Two key themes emerged surrounding capturing human values in current RLHF practices: 1) crowdworker sampling for capturing situated values and 2) annotation tasks for capturing situated values.

Crowdworker Sampling for Capturing Situated Values Following the desiderata on *D1. Diversity of Human Feedback* and *D2. Situatedness of Values*, we identify the challenge of sampling representative crowdworkers for capturing situated values at design-time. Overcoming this challenge necessitates a deliberate effort to recruit representative crowdworkers. However, the notion of what constitutes "representative" may evolve over time and vary depending on the situation of these systems' application.

(LC-SA-1) *Limitation in Capturing: Sampling Crowdworker of diverse demographics.* Recent scrutiny surrounding annotator selection for RLHF primarily revolves around the question of whether the selected crowdworker groups truly represent the diversity of real-world users (Casper et al. 2023). Current demographics of crowdworkers across various platforms often fail to mirror the demographic diversity found in real-user populations. Consequently, crowdworkers engaged at design-time may only partially address the end user's needs at use-time.

For instance, platforms like OpenAI predominantly involve crowdworkers from specific demographics, such as Filipino and Bangladeshi nationals (50% of all crowdworkers), as well as individuals aged 25-34 (also 50%) (Ouyang et al. 2022). Similarly, Anthropic's evaluator population is largely (68%) composed of individuals from white ethnic backgrounds (Bai et al. 2022). These demographic biases can potentially result in implicit biases or skewed representations of fundamental human values, which may inadvertently influence model training processes (Peng et al. 2022, 2019). Without meticulous selection procedures, RLHF risks producing even more biased models (Santurkar et al. 2023; Hartmann, Schwenzow, and Witte 2023), reinforcing certain cultural or national perspectives while marginalizing others.

(LC-SA-2) *Limitation in Capturing: Sampling Crowdworkers with diverse situational factors.* In line with the principles of situated value understanding, a "representative" crowdworker selection process should aim not only to capture demographic diversity but also to approximate the diverse situated values. However, the current practices in RLHF crowdworker selection often fall short of addressing these nuanced requirements, rendering the exclusive reliance on specific crowdworker demographics inadequate and potentially detrimental to model alignment efforts.

It is crucial to recognize that the criteria for crowdworker selection may evolve, necessitating ongoing refinement and adaptation as RLHF methodologies are applied in diverse situations and for varying purposes.

Annotation Tasks for Capturing Situated Values Through the situated value alignment desiderata on *D6. Sensitivity to Emerging Values*, *D8. Support for Value Elicitation*, *D1. Diversity of Human Feedback* and *D2. Situatedness of Values*, we illuminate the challenge of capturing situated values in current RLHF practices.

(LC-AT-1) *Limitation in Capturing: collecting fine-grained feedback through Annotation Tasks.* Current feedback collection methods utilized in RLHF often elicit one measurement of "goodness." Primarily through preference ratings between pairs of examples (Christiano et al. 2017) or k-wise rankings (Zhu, Jiao, and Jordan 2023), where crowdworkers rank multiple examples according to their preferences, simple understanding can be gained about user preferences (such as one sample is preferred over another). In the past, such simplified measurements proved to suffice as a good balance between effectiveness and feasibility, such as YouTube thumbs up or down or like/no like on platforms such as Instagram or Facebook, to understand what users like. However, when utilized to capture nuanced situated value interpretations or even value hierarchies, the granularity of these annotation tasks proves insufficient (Lammerts et al. 2023).

(LC-AT-2) *Limitation in Capturing: facilitating the externalization of values through Annotation Task design.* Human values are highly abstract constructs that present a major challenge when trying to be externalized and captured (Le Dantec, Poole, and Wyche 2009; Liscio et al. 2021). Meanings, nuances, and interactions of values are complicated to express in simple rankings, while value elicitation techniques (e.g., laddering, photo-elicitation interviews) are not as machine-readable or scaleable (Pommeranz et al. 2012) (see section 3.1.2). Concerning the problem of granularity above, we find current feedback collection tasks insufficient in supporting this highly challenging task of surfacing one's situated values.

(LC-AT-3) *Limitation in Capturing: collecting reasoning through Annotation Tasks.* Part of the problem is the lack of qualitative information about the crowdworker's reasoning. Implicit annotation mechanisms obscure the true meaning of feedback, leaving us unaware of the values it represents. Rationales behind value judgment provide essential insights into the values underpinning human feedback. The human computation community has long advocated such an idea of collecting human rationales in annotation (McDonnell et al. 2016), where it has been shown to be not only useful for collecting more information but also makes the annotation more reliable (as it "forces" the crowdworker to reflect on their reasons in the annotation).

(LC-AT-4) *Limitation in Capturing: collecting information about the situation through Annotation Tasks.* Most importantly, current annotation practices leave us unaware of the situations or even contexts that largely influence preference ratings and situated values. There are several ways in which these blind spots or *hidden situations* may arise. Due to systemic biases and constrained cognition, human annotators may not always operate optimally in accordance with their preferences. Further, the system remains unable to capture every element or *material of a situation* (such as the crowdworker's mood, culture, religion, personal environment, etc.) that a crowdworker considers while reaching a conclusion.(Siththaranjan, Laidlaw, and Hadfield-Menell 2023b)

To illustrate the issue of current preference learning with hidden situations, consider the following hypothetical scenario:

Example 2.1. *A language model development team employs RLHF techniques to enhance their large language model's (LLM) text generation abilities. They collect crowdworker feedback by presenting pairs of responses and asking crowdworkers to choose the most helpful or relevant one. However, the feedback collection process fails to capture the nuanced situational factors influencing a crowdworker's preferences and judgments. Consider a scenario where a crowdworker asks the language model for advice on dealing with stress. The model generates two responses: one offering practical coping strategies and another providing empathetic support and encouragement. Unbeknownst to the feed-*

back collection process, the crowdworker may be experiencing acute stress and seeking emotional support rather than practical advice. Consequently, they select the empathetic response, believing it better meets their current needs. The RLHF process interprets this feedback as a preference for empathetic responses over practical advice in all situations. As a result, the language model's training data becomes skewed towards generating more empathetic responses, potentially at the expense of providing useful and actionable guidance in other situations.

The fundamental problem with popular preference learning algorithms such as RLHF is that they assume all relevant features are captured, which is rarely the case in real-world situations. Therefore, using typical procedures may result in unanticipated and unfavourable results. In Example 2.1, relevant situational factors about the crowdworker's mood and reason for being stressed are missing from the data.

(LC-AT-5) *Limitation in Capturing: quality control of Annotation Tasks.* Quality control is a central topic in annotation by large crowds with varying levels of reliability. It is generally achieved through mechanisms such as honeypot questions and eliciting redundant annotation to infer the truth through aggregation. Many of these mechanisms assume the single truth of data annotated and are thus not directly transferable to situated values.

Scaling Value Judgments at Design-Time

One main theme emerged for the scaling phase at design-time: the insufficient reward function modelling.

Insufficient Reward Function Modelling The desiderata on *D1. Diversity of Human Feedback* and *D3. Adaptability to Variability of Situated Values* allowed for the identification of shortcomings in the current reward modelling phase.

(LS-RM-1) *Limitation in Scaling: a pluralistic approach to Reward Modelling.* While numerical representations facilitate generalization beyond observed examples, RLHF excels in automating human judgment through reward models, offering timely signals for optimizing LLMs. This approach mitigates the limitations of manually modelling human preferences, enabling practical generalization beyond specific examples. However, akin to concerns raised by (Casper et al. 2023), relying on a single reward function fails to capture the diverse range of human values and preferences as it is typically formulated as the solution for a single human.

The possibility of obtaining multiple reward functions relies on preserved disagreement in the crowdworkers' feedback. This points to another issue of capturing crowdworker values: the challenge of aggregating the feedback (Siththaranjan, Laidlaw, and Hadfield-Menell 2023b). In example 2.1, the implicit aggregation over preferences privileges the viewpoints of the majority. Here, anonymized data aggregation may conceal significant variations in crowdworkers' preferences, complicating the interpretation of feedback (Siththaranjan, Laidlaw, and Hadfield-Menell 2023a). This largely limits the model's ability to capture the varying situated values.

Human crowdworkers thereby often exhibit disagreement, as noted by (Stiennon et al. 2020; Ouyang et al. 2022; Bai et al. 2022) with agreement rates ranging from 63% to 77%. We expect this disparity to increase with the inclusion of a more diverse group of crowdworkers. Attempting to aggregate feedback from such diverse sources into a single reward model overlooks these inherent differences, leading to a fundamentally misspecified problem.

(LS-RM-2) *Limitation in Scaling: value hierarchies in Reward Modelling.* While a sufficiently complex reward function theoretically accommodates value variety, current methods of collecting human feedback lack the granularity necessary for a comprehensive understanding of value hierarchies. Ideal reward models should be able to represent rewards at different granularity levels, which is required for reward modelling to match human feedback of values at different granularities. Furthermore, specifying tolerance levels for undesirable model behaviour becomes imperative as feedback granularity increases. This nuanced evaluation enables a more robust alignment of AI outputs with human values.

(LS-RM-3) *Limitation in Scaling: transparent Reward Modeling.* To allow inspection into what the reward model has learned or not learned, particularly regarding values, its transparency is an essential requirement of the reward model. Transparent reward modelling is akin to explainable AI in the sense that the explanation is about describing the values, ideally of multiple values with different levels of granularity, learned by the reward functions. Such a desired effort is in line with collecting reasoning of the crowdworker's value judgments (i.e., LC-AT2). However, the current state of research has not looked into the explanation of reward models despite the abundant amount of literature on explainable AI.

(LS-RM-4) *Limitation in Scaling: generalizing to unseen scenarios in Reward Modeling.* Reward modelling aims to generalize value preferences to unseen data instances by yet another neural network. Despite the importance and a large amount of research effort, the generalisation ability of neural networks has remained an unresolved, challenging question in machine learning, especially for instances that are out of the distribution of training data. The question is inherently tied also to explainable AI and, therefore, closely related to the previous limitation of transparency.

Capturing and Adapting to Human Preferences at Use-Time

Based on the resources available for value alignment, we draw the distinction between design-time and use-time. The latter can be structured as follows.

(LC-UF-1) *Limitation in Capturing: capturing real User Values.* Based on the desiderata of *D2. Situatedness of Values*, *D5. Responsiveness to Evolving Preferences*, *D6. Sensitivity to Emerging Values* and *D7. Responsiveness to Unforeseen Situational Changes* we infer the following: While we critiqued feedback capturing and scaling at design-time above, the emergence of situated values in particular situations puts practical constraints on design-time improvements. Particularly, we lack access to real users and their situations at design-time and thus real situated value interpretations, means of actualization and hierarchies. We also run the risk of *value locking* by not evolving our value understanding and feedback with the evolving human values over time. We argue that to overcome these practical limitations at design-time, we require complementary actions at use-time. At use-time, we argue, real users can report directly about their situation and value interpretations, means of actualization and hierarchies through additional value elicitation. An in-situ value inquiry could allow us to overcome human values' temporal variations and evolution. In doing so, the aforementioned limitations of capturing at design-time are all relevant and need to be carefully addressed.

(LA-VA-1) *Limitation in Adapting: reacting to value variations in different situations through Value Adaptation.* Following the desiderata of *D3. Adaptability to Variability of Situated Values*, *D4. Flexibility in Action Selection*, *D5. Responsiveness to Evolving Preference*, *D6. Sensitivity to Emerging Values* and *D7. Responsiveness to Unforeseen Situational Changes*, we identify a lack of adaptation. Considering the variety of truths that situated value perspective offers, we need to ask, in turn, how current alignment practices react and adapt to varying situations and users. Currently, the user's prompt is informative about the tone, sentiment, and accent the LLM is deploying in their responses. However, current LLMs fail to provide adaptation or personalized experience (Eapen and Adhithyan 2023), especially when taking into account values specific to the situation at hand.

Research Directions for Situated Value Alignment

In this paper, we argue that values acquire situated meaning; hence, their concrete interpretations, means of actualization and hierarchies vary between people depending on the situations in which they guide the notion of what we really value. So far, we have explored to what extent RLHF, the predominant approach to LLM value alignment, supports the alignment of situated values. Here, we reflect on our previous critique of the concrete shortcomings of RLHF and suggest future research directions, which we broadly summarize as *Situated Annotation*, *Expressive Instruction*, and *Reflexive Adaptation*. In the following, we will frame these future research directions as research questions and discuss what it might take to operationalize the research in this space.

Situated Annotation

Here, we call for further investigation of *situated annotation*, a set of potential practices that focus on value capturing methods, as discussed in the context of RLHF. Unlike current annotation practices –where we only collect crowdworker's judgments and implicit value preferences–, situated annotation enables the collection of crowdworker judgments, situations and values at design-time and real-user judgment, situations and values at use-time (see Figure 2). Situated annotation, if done well, will stand as a cornerstone for the development of situated alignment practices, essential for systems to comprehend diverse value conceptions across a wide array of use cases. Yet, a series of research questions need to be addressed concerning the limitations identified in Section .

How can we recognize and incorporate newly emerging values after deploying the models? Recognizing that end-user situated values cannot be entirely anticipated at the design-time (due to situated value emergence D6) – (LC-UF-1), we argue that the notion of situated annotation should extend beyond the development phase to include real-user judgments, values and situations, wherein users actively contribute to a more concrete situated value understanding. Situated annotation, as can be seen in Figure 2, is, therefore, employed with crowdworkers at design-time and with users at use-time.

What Sampling Criteria should be employed to represent situations effectively? The research question stems from the disparity in demographic representation between crowdworkers and real users (Casper et al. 2023) (LC-SA-1) and the lack of consideration for situated values (LC-SA-2), which all need to be addressed in future iterations of situated annotation. Similar to (Waseem 2016; Díaz et al. 2022; Kasirzadeh and Gabriel 2023), we, therefore, call for a diverse set of annotators who are capable of capturing situations, judgments, and values. In this regard, recent research has, for instance, begun investigating annotator selection techniques after data sampling to emulate the whole range of human perspectives effectively (van der Meer et al. 2024).

How to support the elicitation of fine-grained situated values? Taking situated values as abstract cognitive constructs, often hidden from awareness, crowdworkers and users need to be guided in surfacing and externalizing their knowledge in a fine-grained manner (LC-AT-1, LC-AT-2). Insights from disciplines such as design, economics, and psychology offer valuable perspectives on eliciting tacit knowledge (referring to the knowledge, skills, and abilities an individual gains through experience that is often difficult to put into words or otherwise communicate) (Tsoukas 2005; Rust 2004; Sanders and Stappers 2012). Games-With-A-Purpose (GWAPs) have also demonstrated potential in incentivizing engagement in cognitively intensive situated annotation tasks (Law and Von Ahn 2011). Recent endeavours have utilized GWAPs to extract tacit commonsense knowledge from crowds through competitive game-play (Balayn et al. 2022), underscoring their viability in enhancing participation.

How to represent situations? In addition to fine-grained feedback on situated values, situated annotation requires rationals, judgments, and their related situations (LC-AT-3, LC-AT-4). Here, inspiration for how to approximate and represent a situation could be drawn from psychology, where taxonomies of psychologically important situation characteristics could support the computational representation and approximation of messy real-world situations(Rauthmann et al. 2014). Similarly, newly developed reward systems could be used to explore the most relevant, informative and defeasible questions to understand the annotator's situation (Pyatkin et al. 2022).

How can we evaluate the meaning of conversational agents' speech beyond relying solely on a single truth cri-

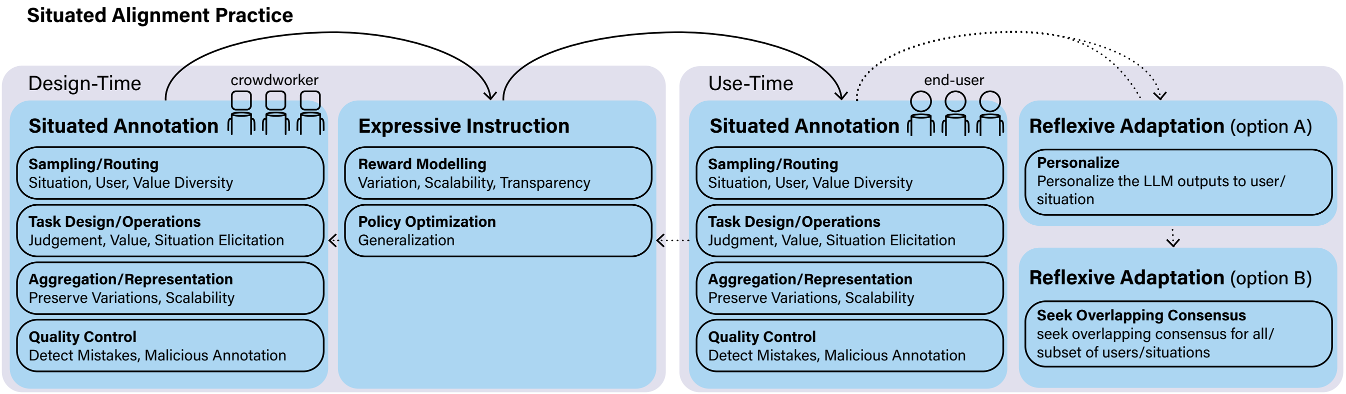
Situated Alignment Practice
Design-Time
crowdworker
Situated Annotation
Sampling/Routing
Situation, User, Value Diversity
Task Design/Operations
Judgement, Value, Situation Elicitation
Aggregation/Representation
Preserve Variations, Scalability
Quality Control
Detect Mistakes, Malicious Annotation
Expressive Instruction
Reward Modelling
Variation, Scalability, Transparency
Policy Optimization
Generalization
Use-Time
end-user
Situated Annotation
Sampling/Routing
Situation, User, Value Diversity
Task Design/Operations
Judgment, Value, Situation Elicitation
Aggregation/Representation
Preserve Variations, Scalability
Quality Control
Detect Mistakes, Malicious Annotation
Reflexive Adaptation (option A)
Personalize
Personalize the LLM outputs to user/ situation
Reflexive Adaptation (option B)
Seek Overlapping Consensus
seek overlapping consensus for all/ subset of users/situations

Figure 2: Future research directions for LLM value alignment through RLHF-like practices with *Situated Annotation*, *Expressive Instructions*, and *Reflexive Adaptation*, unrolling over the different stages of the LLM lifecycle (i.e. design-time and use-time). Specific research questions are to be addressed in the pursuit of each of the research directions.

terion? Quality control represents another critical facet of task design, particularly concerning incentivization mechanisms (LC-AT-5). Quality control mechanisms should be implemented to ensure annotated data's reliability and consistency. This could involve techniques such as inter-annotator agreement, where multiple annotators label the same data, and their agreements and disagreements are analyzed to refine the annotation process (Yang et al. 2016). Such an approach is more adequate than honey-pot questions given the subject nature of value judgment, yet it comes with further requirements, such as identifying "similar" annotators in inter-annotator agreement measurement where annotator similarity needs to be carefully defined.

Expressive Instruction

Next to improved methods for capturing information regarding crowdworker/user judgments, values and situations, situated annotation practices might require new techniques to encode and represent the pluralistic annotations to instruct LLMs (see Figure 2).

How to encode multiple, multi-level reward functions in reward modelling? Multiple reward functions for the variety of situated values (LS-RM-1) at different levels of value hierarchies (LS-RM-2) should be implemented to ensure that the model learns to optimize across a broad spectrum of user needs and desires. Multi-task or multi-objective learning (Zhang and Yang 2018) has been a topic of research in machine learning that could be investigated in this context where the objectives represent different value interpretations, means of actualization and hierarchies and situations. Another related machine learning idea is structure learning (e.g., recursive neural nets (Socher et al. 2011)) that accounts for the structure of the learning objectives, which could be potentially relevant for modelling a large variety of situated values.

How to encode reward functions transparently? There has been a call for these reward functions to become more transparent (LS-RM-3) to allow stakeholders to understand how decisions are made and provide opportunities for scrutiny and improvement (Lambert, Krendl Gilbert, and Zick 2023; Gilbert et al. 2023). The idea is directly related to explainable AI (XAI) (Guidotti et al. 2018), only that in this particular case, the object of explanation is what values are learned by the reward models. XAI has developed into a broad field where many different methods have been introduced following different principles – particularly relevant ideas in this realm are concept-based explanation that emphasises the intelligibility of explanations to people in human-understandable concepts (Kim et al. 2018; Balayn et al. 2021) and causal explanations that reveal the causal relationships among the concepts in explanations (Biswas et al. 2022).

How to generalize from limited value judgments for instruction at scale? The reward function simulates human judgment to instruct LLMs by rating any LLM outputs, which largely relies on its ability to generalise from a limited number of value judgments. In this respect, it becomes beneficial to employ explicit interpolation and extrapolation techniques to infer user values beyond those directly represented in the training data (LS-RM-4). Methods such as transfer learning and meta-learning can be particularly useful in this regard, allowing models to leverage knowledge from related tasks or domains to make predictions for unseen users or situations (Pan and Yang 2009; Finn, Abbeel, and Levine 2017). On the other hand, it should be noted that any such attempts should be carefully carried out given that the generalisation ability of neural networks is not (and maybe never) fully understood, especially for extrapolation. To some extent, this can be compensated by empirical studies that provide empirical insights into the level of generalisation.

Reflexive Adaptation

Situated value alignment at the use-time is non-trivial considering the variety of value conceptions and situations (LA-VA-1). Here, we see two possible approaches to situated value alignment (at use-time): 1) personalization of LLMs towards specific users, values and situations, and 2) seeking overlapping consensus across various user judgments, values and situations. In the following, we discuss each of these potential directions, their benefits and risks while staying true to our vision of a *Reflexive Adaptation*.

In this context, *adaptation* refers to a model's specific actions to respond safely to values and situations. We define a *reflexive* approach as involving continuous awareness and responsiveness to evolving values. Unlike static strategies, reflexive LLMs can introspect on their outputs, adapting to subtle value nuances. This reflexivity fosters dynamic, ethically grounded user interactions, making LLMs more situation-aware. *Situated alignment*, then, combines situated annotation and reflexive adaptation as the overarching endeavour.

Should we personalize LLM responses to individual users in specific situations? Personalization is a compelling strategy for enhancing the alignment of LLMs with user values in diverse situations. By tailoring responses to individual value interpretations, means of actualization and hierarchies and sensitivities, personalization can ensure that users are addressed in a manner that resonates with their unique viewpoints, significantly enhancing user satisfaction and engagement (Kay 2001), as well as trust in the system's recommendations and responses.

Example 3.1. *Consider a patient interacting with a healthcare chatbot powered by a language model. This patient may have strong preferences regarding how they are addressed—perhaps they prefer a formal tone and expect respectful language at all times. Ensuring these preferences are met might be crucial to enhancing the acceptance of suggested medical care.*

However, despite these benefits, personalization also poses significant risks, particularly in the context of exacerbating societal polarization and reinforcing echo chambers (Pariser 2011). When language models are overly personalized, they risk reinforcing existing biases and preferences, potentially creating filter bubbles where users are only exposed to information that aligns with their pre-existing beliefs.

Example 3.2. *Imagine a social media platform employing a highly personalized language model to curate users' news feeds. If the model consistently surfaces content that aligns with a user's existing values in specific situations, users may become isolated within their own information bubbles, shielded from diverse perspectives and alternative viewpoints.*

Should we reach overlapping consensus among user groups with varying situated values? Unlike personalization, which adjusts outputs based on detected nuances in situated values, the second approach to situated value alignment aims for an *overlapping consensus*. Drawing from Rawls' theory of justice, stakeholders engage in rational deliberation to identify shared values, forming consensus (Rawls 2017). This approach promotes understanding and mutual respect among users with diverse perspectives, fostering cooperation across cultural and ideological divides. Individuals can agree on principles and standards acceptable in a situation, even if they differ on underlying moral nature.

This, however, poses the challenge of reaching a consensus in situations where values are deeply entrenched and polarized. In situations where fundamental value differences exist, achieving overlapping consensus may prove elusive, potentially leading to stalemates or exacerbating tensions. Furthermore, the process of reaching overlapping consensus requires careful navigation to avoid marginalizing minority viewpoints or privileging dominant perspectives. Without robust mechanisms to ensure equitable representation and participation, the consensus-building process may inadvertently reinforce existing power dynamics and perpetuate systemic inequalities.

How could we reach overlapping consensus across varying situated values and situations? From philosophy, we could draw from Rawl's idea of wide reflective equilibrium to encourage the iterative adjustment of beliefs to align with shared values (Rawls 2017; Daniels 1979, 1996). On the other hand, insights from participatory design could offer methodologies for negotiating diverse viewpoints, fostering dialogue and collaboration (Björgvinsson, Ehn, and Hillgren 2012).

While future work has to investigate how such a reflexive adaptation could occur, previous work on value alignment calls for similar strategies of *equilibrium alignment* (Yi et al. 2023). No matter the approach for situated LLM alignment, the question remains: who would the target be: an individual, a group, a company, a country, or all of humanity? (Kasirzadeh and Gabriel 2023).

Conclusions

In this paper, we advocate for the integration of situated values – a theoretical lens inspired by feminist epistemology and the concept of situated understandings– as a means to ground value interpretations in the complexities of real-world situations. This complements current universal framings of values derived from ethical guidelines which remain detached from the vast array of scientific, technical, and economic contexts and in sometimes geographically dispersed groups of researchers and developers with different priorities, tasks, and fragmental responsibilities (Hagendorff 2020; Morley et al. 2021).

We reinforce this call by analysing Reinforcement Learning from Human Feedback (RLHF) regarding situated values, highlighting limitations in annotator sampling, annotation task design, reward function modelling, and adaptation. We reflect on these shortcomings to propose: 1) situated annotation for eliciting user preferences in specific situations, 2) expressive instruction to encode plural value notions for LLM instruction at scale, and 3) reflexive adaptation using real user information. We stress the significance of situated values for better alignment while acknowledging implementation challenges.

This paper contributes to a more holistic understanding of where the scientific community should invest more time in the future by starting to bridge the gap between high-level ethical guidelines and value needs as emergent to the specific use cases and applications of LLMs.

Acknowledgments

The work was funded by the ICAI lab GENIUS (Generative Enhanced Next-Generation Intelligent Understanding Systems), a collaboration between Delft University of Technology, Maastricht University, dsm-firmenich, and Kickstart AI, and the NWO Long-Term Programme ROBUST initiated by the Innovation Centre for Artificial Intelligence (ICAI).

References

Anderson, E. 1995. Feminist epistemology: An interpretation and a defense. *Hypatia*, 10(3): 50–84.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Balayn, A.; He, G.; Hu, A.; Yang, J.; and Gadiraju, U. 2022. Ready Player One! Eliciting Diverse Knowledge Using A Configurable Game. In *Proceedings of the ACM Web Conference 2022*, 1709–1719.

Balayn, A.; Soilis, P.; Lofi, C.; Yang, J.; and Bozzon, A. 2021. What do you mean? Interpreting image classification with crowdsourced concept extraction and analysis. In *Proceedings of the Web Conference 2021*, 1937–1948.

Basta, C.; Costa-Jussà, M. R.; and Casas, N. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*.

Baumann, J.; and Loi, M. 2023. Fairness and Risk: An Ethical Argument for a Group Fairness Definition Insurers Can Use. *Philosophy & Technology*, 36(3): 45.

Bell, D. A. 1996. The East Asian challenge to human rights: Reflections on an East West dialogue. *Hum. Rts. Q.*, 18: 641.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 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*, 610–623. Virtual Event Canada: ACM. ISBN 978-1-4503-8309-7.

Biswas, S.; Corti, L.; Buijsman, S.; and Yang, J. 2022. CHIME: Causal Human-in-the-Loop Model Explanations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, 27–39.

Björgvinsson, E.; Ehn, P.; and Hillgren, P.-A. 2012. Agonistic participatory design: working with marginalised social movements. *CoDesign*, 8(2-3): 127–144.

Bourdieu, P. 2018. Structures, habitus, practices. In *Rethinking the subject*, 31–45. Routledge.

Brand, S. 1995. *How buildings learn: What happens after they're built*. Penguin.

Buijsman, S. 2023. Navigating fairness measures and tradeoffs. *AI and Ethics*, 1–12.

Callaghan, W.; Goh, J.; Mohareb, M.; Lim, A.; and Law, E. 2018. Mechanicalheart: A human-machine framework for the classification of phonocardiograms. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–17.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Daniels, N. 1979. Wide reflective equilibrium and theory acceptance in ethics. *The journal of philosophy*, 76(5): 256–282.

Daniels, N. 1996. *Justice and justification: Reflective equilibrium in theory and practice*. Cambridge University Press.

De La Bellacasa, M. P. 2012. 'Nothing comes without its world': thinking with care. *The sociological review*, 60(2): 197–216.

Díaz, M.; Amironesei, R.; Weidinger, L.; and Gabriel, I. 2022. Accounting for offensive speech as a practice of resistance. In *Proceedings of the sixth workshop on online abuse and harms (woah)*, 192–202.

Dietz, T. 2013. Bringing values and deliberation to science communication. *Proceedings of the National Academy of Sciences*, 110(supplement_3): 14081–14087.

Eapen, J.; and Adhithyan, V. 2023. Personalization and customization of llm responses. *International Journal of Research Publication and Reviews*, 4(12): 2617–2627.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Fischer, G.; and Scharff, E. 2000. Meta-design: design for designers. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, 396–405.

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

Gilbert, T. K.; Lambert, N.; Dean, S.; Zick, T.; Snoswell, A.; and Mehta, S. 2023. Reward reports for reinforcement learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 84–130.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5): 1–42.

Gurung, L. 2020. Feminist Standpoint Theory: Conceptualization and Utility. *Dhaulagiri: Journal of Sociology & Anthropology*, 14.

Hafer, C.; and Landa, D. 2007. Deliberation as self-discovery and institutions for political speech. *Journal of theoretical Politics*, 19(3): 329–360.

Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and machines*, 30(1): 99–120.

Haraway, D. 2016. 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective'. In *Space, gender, knowledge: Feminist readings*, 53–72. Routledge.

Haraway, D. J.; and Goodeve, T. N. 2018. Modest_witness@ second_millennium. In *Modest_Witness@ Second_Millennium. FemaleMan_Meets_OncoMouse*, 23–45. routledge.

Harding, S. 1992. Subjectivity, experience and knowledge: An epistemology from/for rainbow coalition politics. *Development and Change*, 23(3): 175–193.

Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Hildebrandt, M. 2019. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law*, 20(1): 83–121.

Kasirzadeh, A.; and Gabriel, I. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2): 27.

Kay, J. 2001. User modeling for adaptation. *User interfaces for all: Concepts, methods, and tools*, 4: 271–294.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677. PMLR.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*.

Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; et al. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *arXiv preprint arXiv:2404.16019*.

Kuppler, M.; Kern, C.; Bach, R. L.; and Kreuter, F. 2021. Distributive justice and fairness metrics in automated decision-making: How much overlap is there? *arXiv preprint arXiv:2105.01441*.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Lambert, N.; Krendl Gilbert, T.; and Zick, T. 2023. The history and risks of reinforcement learning and human feedback. *arXiv e-prints*, arXiv–2310.

Lammerts, P.; Lippmann, P.; Hsu, Y.-C.; Casati, F.; and Yang, J. 2023. How do you feel? measuring user-perceived value for rejecting machine decisions in hate speech detection. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 834–844.

Law, E.; and Von Ahn, L. 2011. *Human computation*. Morgan & Claypool Publishers.

Le Dantec, C. A.; Poole, E. S.; and Wyche, S. P. 2009. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1141–1150.

Lim, C. Y.; Berry, A. B.; Hartzler, A. L.; Hirsch, T.; Carrell, D. S.; Bermet, Z. A.; and Ralston, J. D. 2019. Facilitating self-reflection about values and self-care among individuals with chronic conditions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Liscio, E.; van der Meer, M.; Siebert, L. C.; Jonker, C. M.; Mouter, N.; and Murukannaiah, P. K. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *AAMAS*, 799–808.

McDonnell, T.; Lease, M.; Kutlu, M.; and Elsayed, T. 2016. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, 139–148.

Metz, T. 2012. African conceptions of human dignity: Vitality and community as the ground of human rights. *Human Rights Review*, 13(1): 19–37.

Morley, J.; Kinsey, L.; Elhalal, A.; Garcia, F.; Ziosi, M.; and Floridi, L. 2021. Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY*, 1–13.

Ngo, R.; Chan, L.; and Mindermann, S. 2022. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Noorman, M. 2023. Computing and moral responsibility.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Panikkar, R.; and Panikkar, R. 1982. Is the notion of human rights a Western concept? *Diogenes*, 30(120): 75–102.

Pariser, E. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Peng, A.; Nushi, B.; Kiciman, E.; Inkpen, K.; and Kamar, E. 2022. Investigations of performance and bias in human-AI teamwork in hiring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12089–12097.

Peng, A.; Nushi, B.; Kıcıman, E.; Inkpen, K.; Suri, S.; and Kamar, E. 2019. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 125–134.

Pommeranz, A.; Detweiler, C.; Wiggers, P.; and Jonker, C. 2012. Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology*, 14(4): 285–303.

Pommeranz, A.; Detweiler, C.; Wiggers, P.; and Jonker, C. M. 2011. Self-reflection on personal values to support value-sensitive design. In *Proceedings of HCI 2011 The 25th BCS Conference on Human Computer Interaction 25*, 491–496.

Pyatkin, V.; Hwang, J. D.; Srikumar, V.; Lu, X.; Jiang, L.; Choi, Y.; and Bhagavatula, C. 2022. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. *arXiv preprint arXiv:2212.10409*.

Rauthmann, J. F.; Gallardo-Pujol, D.; Guillaume, E. M.; Todd, E.; Nave, C. S.; Sherman, R. A.; Ziegler, M.; Jones, A. B.; and Funder, D. C. 2014. The Situational Eight DIAMONDS: a taxonomy of major dimensions of situation characteristics. *Journal of personality and social psychology*, 107(4): 677.

Rawls, J. 2017. A theory of justice. In *Applied ethics*, 21–29. Routledge.

Robbins, J.; and Sommerschuh, J. 2020. Values.

Russell, S. J.; and Norvig, P. 2010. *Artificial intelligence a modern approach*. London.

Rust, C. 2004. Design enquiry: Tacit knowledge and invention in science. *Design issues*, 20(4): 76–85.

Sanders, E. B.-N.; and Stappers, P. J. 2012. *Convivial toolbox: Generative research for the front end of design*. Bis.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Schwartz, S. H. 2007. Basic human values: Theory, measurement, and applications. *Revue française de sociologie*, 47(4): 929.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Siththaranjan, A.; Laidlaw, C.; and Hadfield-Menell, D. 2023a. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. *arXiv preprint arXiv:2312.08358*.

Siththaranjan, A.; Laidlaw, C.; and Hadfield-Menell, D. 2023b. Understanding Hidden Context in Preference Learning: Consequences for RLHF. In *The Twelfth International Conference on Learning Representations*.

Socher, R.; Lin, C. C.; Manning, C.; and Ng, A. Y. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 129–136.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Tsoukas, H. 2005. Do we really understand tacit knowledge. *Managing knowledge: an essential reader*, 107: 1–18.

van der Meer, M.; Falk, N.; Murukannaiah, P. K.; and Liscio, E. 2024. Annotator-Centric Active Learning for Subjective NLP Tasks. *arXiv preprint arXiv:2404.15720*.

Verbeek, P.-P. 2006. Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values*, 31(3): 361–380.

Vought, R. T. 2020. Guidance for Regulation of Artificial Intelligence Applications.

Waseem, Z. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Wielinga, B. 2023. Complex equality and the abstractness of statistical fairness: using social goods to analyze a CV scanner and a welfare fraud detector. *AI and Ethics*, 1–16.

Wong, P.-H. 2020. Cultural differences as excuses? Human rights and cultural values in global ethics and governance of AI. *Philosophy & Technology*, 33(4): 705–715.

Yang, J.; Redi, J.; Demartini, G.; and Bozzon, A. 2016. Modeling task complexity in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, 249–258.

Yang, M.; Wang, W.; Gao, Q.; Zhao, C.; Li, C.; Yang, X.; Li, J.; Li, X.; Cui, J.; Zhang, L.; et al. 2023. Automatic identification of harmful algae based on multiple convolutional neural networks and transfer learning. *Environmental Science and Pollution Research*, 30(6): 15311–15324.

Yi, X.; Yao, J.; Wang, X.; and Xie, X. 2023. Unpacking the ethical value alignment in big models. *arXiv preprint arXiv:2310.17551*.

Yudkowsky, E. 2016. The AI alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 4.

Zhang, Y.; and Yang, Q. 2018. An overview of multi-task learning. *National Science Review*, 5(1): 30–43.

Zhu, B.; Jiao, J.; and Jordan, M. I. 2023. Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons. *arXiv preprint arXiv:2301.11270*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.