Position: Towards Bidirectional Human-AI Alignment

Hua Shen[1A *] **Tiffany Knearem**[2B] **Reshmi Ghosh**[3B]
Kenan Alkiek[4C] **Kundan Krishna**[5C] **Yachuan Liu**[4C] **Ziqiao Ma**[4C] **Savvas Petridis**[9C]
Yi-Hao Peng[7C] **Li Qiwei**[4C] **Sushrita Rakshit**[4C] **Chenglei Si**[8C] **Yutong Xie**[4C]
Jeffrey P. Bigham[7D] **Frank Bentley**[6D] **Joyce Chai**[4D] **Zachary Lipton**[7D]
Qiaozhu Mei[4D] **Rada Mihalcea**[4D] **Michael Terry**[9D] **Diyi Yang**[8D]
Meredith Ringel Morris[9E] **Paul Resnick**[4E] **David Jurgens**[4E]
[1] NYU Shanghai, New York University, [2] MBZUAI, [3] Microsoft, [4] University of Michigan,
[5] Apple, [6] Google, [7] Carnegie Mellon University, [8] Stanford University, [9] Google DeepMind
[A] Project Lead, [B] Team Leads, [C] Team Members (Equal Contributions),
[D] Advisors (Equal Contributions), [E] Lead Advisors (Equal Contributions),
huashen@nyu.edu

Abstract

Recent advances in general-purpose AI underscore the urgent need to align AI systems with human goals and values. Yet, the lack of a clear, shared understanding of what constitutes "alignment" limits meaningful progress and cross-disciplinary collaboration. In this position paper, we argue that the research community should explicitly define and critically reflect on "alignment" to account for the bidirectional and dynamic relationship between humans and AI. Through a systematic review of over 400 papers spanning HCI, NLP, ML, and more, we examine how alignment is currently defined and operationalized. Building on this analysis, we introduce the Bidirectional Human-AI Alignment framework, which not only incorporates traditional efforts to align AI with human values but also introduces the critical, underexplored dimension of aligning humans with AI – supporting cognitive, behavioral, and societal adaptation to rapidly advancing AI technologies. Our findings reveal significant gaps in current literature, especially in long-term interaction design, human value modeling, and mutual understanding. We conclude with three central challenges and actionable recommendations to guide future research toward more nuanced, reciprocal, and human-AI alignment approaches.

1 Introduction

Artificial Intelligence (AI), particularly generative AI, has demonstrated remarkable capabilities in reasoning, language understanding, problem solving, and more [1]. However, its increasing integration into society raises significant risks, such as amplifying biases in hiring [2] or perpetuating stereotypes in text-to-image models [3]. These concerns highlight the urgent need to align these systems with values, ethical principles, and the goals of individuals and society at large. This need, commonly referred to as "AI alignment," [4, 5] is crucial for ensuring that AI systems function in a manner that is not only effective but also consistent with human values, minimizing harm and maximizing societal benefits. Yet, key challenges remain:

Challenge 1: Specification Gaming. AI designers often define objectives or feedback to align systems with human goals, but these rarely capture all intended values [6]. This leads to reliance on

*We denote all authors' roles, affiliations, and contributions in Appendix 13.1. This work was supported in part by the National Science Foundation under Grant No. IIS-2143529 and No. IIS-1949634. Corresponding author: Hua Shen, Assistant Professor of NYU Shanghai, New York University (huashen@nyu.edu).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Position Paper Track.

proxies like human approval [4], enabling specification gaming [7, 8], where AI makes seemingly "correct" decisions for the wrong, opaque reasons [9, 10, 11].

Challenge 2: Scalable Oversight. As AI systems become more complex—potentially reaching AGI [12] —aligning them through human feedback grows harder. Evaluating their behavior is often slow or infeasible [5], prompting research into reducing supervision burdens and enhancing human oversight, a challenge known as Scalable Oversight [13].

Challenge 3: Dynamic Nature. As AI advances, alignment must adapt to evolving human values. Without considering long-term cognitive and social impacts, AI may become neither humane nor desirable [14]. This needs a dynamic, ongoing alignment process with cross-disciplinary collaboration.

Traditionally, AI alignment has been approached as a static, one-way process focused on shaping AI to achieve desired outcomes and avoid harm [15, 1, 16]. However, this unidirectional view is increasingly insufficient as AI systems become more integrated into daily life and assume complex decision-making roles [17]. Their interactions with humans create evolving feedback loops that influence both AI behavior and human responses [18, 19, 20], highlighting the need for a more dynamic and reciprocal understanding of alignment [17].

In this position paper, we argue that it is critical for the research community to explicitly reflect on what we mean by "alignment" and to take into account the bidirectional, dynamic interactions between humans and AI to achieve responsible and safe AI systems.

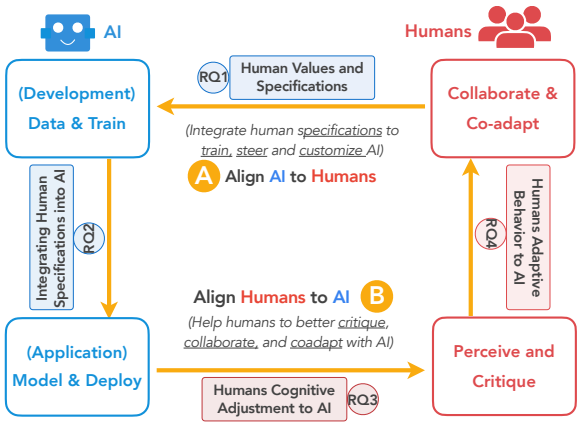

Figure 1: The overview of the Bidirectional Human-AI Alignment framework. Our framework encompasses both A conventional studies which focus on "Align AI with Humans" and B "Align Humans with AI". We further identify four key Research Questions to facilitate this holistic loop of "bidirectional human-AI alignment", and provide answers that can potentially address RQ1-RQ4 in Section 3.

Through a systematic review of over 400 papers across HCI, NLP, ML, and related fields, we examine how alignment is currently defined and implemented. Based on this analysis, we propose the Bidirectional Human-AI Alignment (in Figure 1) framework. It extends the traditional focus on "Aligning AI to Humans"—integrating human input to train, steer, and customize AI—by introducing the equally vital yet underexplored direction of "Aligning Humans to AI", which emphasizes cognitive, behavioral, and societal adaptation to rapid AI advancement.

Our findings reveal key gaps in existing research, particularly in human value modeling, oversight of model inference, critical evaluation of AI's embedded values, and its broader societal impact. We conclude by outlining near- and long-term risks and opportunities, offering actionable recommendations to advance more reciprocal, adaptive, and nuanced approaches to human-AI alignment.

2 Defining Alignment: Fundamentals

Building on our analysis of systematic review (see details in Appendix 8), we explicitly identify the key definitions in alignment research and formally propose "Bidirectional Human-AI Alignment".

Goals. AI alignment research proposes multiple alignment goals [21, 22], such as *intentions* [1, 23], *preferences* [24, 25], *instructions* [26, 27], and *values* [28, 29]. But these terms are often used interchangeably without clear distinctions. Philosophical analysis suggests human values (moral beliefs/principles) are the most suitable alignment goal, as they ensure AI acts ethically while minimizing risks [29, 30]. Though trade-offs exist, this work adopts "human values" as the alignment objective, meaning AI should behave as individuals or society morally expect (See details in Table 2).

Align with Whom. AI alignment involves multiple stakeholders, such as end users [31], AI practitioners [32, 23, 33], and organizations [34]. Many studies reference "general humans" without accounting for group differences, despite the fact that stakeholders often hold conflicting values [29].

Table 1: The fine-grained typology of two directions in the Bidirectional Human-AI Alignment.

	Research Question	Sub-Research Question	Dimensions	References
Align AI with Humans	RQ1: Human Values and Specifications	**Categorizing Aligned Human Values** *what values have been aligned with AI?*	Source of Values	[41, 42, 43]
			Value Types	[16, 44, 45]
		Interaction Techniques to Specify AI Values *How humans could interactively specify values in AI development?*	Explicit Human Feedback	[46, 47, 26]
			Implicit Human Feedback	[48, 49, 50]
			Simulated Human Value Feedback	[51, 52, 53]
	RQ2: Integrating Human Specification into AI	**Develop AI with General Values** *how to incorporate general human values into AI development?*	Instruction Data	[54, 55, 56]
			Model Learning	[1, 57, 58]
			Inference Stage	[26, 59, 60]
		Customizing AI for Individuals or Groups *how to customize AI to incorporate values of individuals or groups?*	Customized Data	[61, 62, 63]
			Adapt Model by Learning	[64, 65, 66]
			Interactive Alignment	[67, 68, 69]
		Evaluating AI Systems *how to evaluate AI regarding human values?*	Human-in-the-loop Evaluation	[70, 71, 43]
			Automatic Evaluation	[72, 55, 73]
		Ecosystem *how to build the ecosystem to facilitate human-AI alignment?*	Platforms	[56, 74, 75, 76, 77, 78]
Align Humans with AI	RQ3: Human Cognitive Adjustment to AI	**Perceiving and Understanding of AI** *how do humans learn to perceive and explain AI systems?*	Education and Training Human	[79, 80, 81]
			AI Sensemaking and Explanations	[82, 83, 84]
		Critical Thinking about AI *how do humans think critically about AI systems?*	Trust and Reliance on AI Decisions	[85, 86, 87]
			Ethical Concerns and AI Auditing	[88, 89]
			Calibrate Cognition to Align AI	[90, 91, 92]
	RQ4: Human Adaptation Behavior to AI	**Human Collaborating with Diverse AI Roles** *how do humans collaborate with AI with differing capability levels?*	AI Assistants	[93, 94, 95]
			AI Partners	[21, 96, 97]
			AI Tutors	[98, 99, 100]
		AI Impacts on Humans and Society *how are humans influenced by AI systems?*	Impact on Individual Behavior	[18, 101]
			Societal Concerns and AI Impacts	[102, 103]
			Reaction to AI Advancements	[104, 105]
		Evaluation in Human Studies *how might we evaluate the impact of AI on humans and society?*	Evaluate Human-AI Collaboration	[106, 107]
			Evaluate Societal Impact	[16, 108]

Rather than pursuing a universal moral theory, prior work advocates aligning AI with principles tailored to compatible human groups [28]. *Pluralistic value alignment*, grounded in social choice theory [35, 28], provides a framework for integrating diverse perspectives [28]. We adopt this view, recognizing that aligning with affected individuals and groups entails ongoing challenges.

Align with What. While prior studies have sought to align AI with human values, these values are often vaguely or inconsistently defined. To address this, we reviewed several prominent value theories, including Moral Foundations Theory [36] and Social Norms and Ethics [37], drawing from psychology and social science. We selected the Schwartz Theory of Basic Values [38, 39] for its demonstrated cross-cultural applicability, relevance across contexts (individuals, interactions, groups), and frequent adoption in NLP research [40, 41]. This framework consists of eleven types of universal values, and defines values as beliefs about desirable end states or behaviors that transcend specific situations and guide evaluation and decision-making.

***Definition*: Bidirectional Human-AI Alignment** is a comprehensive framework that encompasses two interconnected alignment processes: 'Aligning AI with Humans' and 'Aligning Humans with AI'. The former focuses on integrating human specifications into training, steering, and customizing AI. The latter supports human agency, empowering people to think critically when using AI, collaborate effectively with it, and adapt societal approaches to maximize its benefits for humanity.

3 Bidirectional Human-AI Alignment Framework

This section introduces the Bidirectional Human-AI Alignment framework, which encompasses two interconnected alignment directions as a feedback loop, as shown in Figure 1. The "**Align AI to Humans**" direction refers to mechanisms to ensure that AI systems' values match those of humans'. The "**Align Humans to AI**" direction investigates human cognitive and behavioral adaptation to AI advancement. We introduce more details below.

3.1 Align AI to Humans

This direction delineates alignment research from **AI-centered perspective** (e.g., ML/NLP domains) and provides **AI developers and researchers** with approaches for handling two main challenges: carefully specifying the values of the system, and ensuring that the system adopts the specification robustly [4, 33]. Therefore, as shown in Table 1, we explore two core research questions in this direction as: **RQ1.** *What relevant human values are studied for AI alignment, and how do humans specify these values?* and **RQ2.** *How can human values be integrated into the AI systems?*

- **RQ1: Human Values and Specifications.**

To identify key human values and specifications for AI alignment, we begin by addressing two critical subquestions: (1) what values have AI systems been aligned with? and (2) how can humans

interactively specify values during AI development? As summarized at the top of Table 1, we present the key dimensions that emerged from our analysis of these questions, along with supporting studies.

Categorizing Aligned Human Values. To systematically understand human values relevant to human-AI alignment [109, 110], we draw on the adapted Schwartz Theory of Basic Values, examining values along two dimensions: Sources and Types. The **Sources** dimension categorizes values as individual (e.g., personal interests and biological needs like factuality or cognitive biases) [38, 41, 42], social (e.g., shared group norms such as fairness or morality) [38, 43, 45], and interactive (e.g., expectations in human-AI interactions like usability, autonomy, and trust) [111, 112, 113]. The **Types** dimension organizes values into four high-order categories: Self-Enhancement (e.g., achievement, power) [114, 90, 115], Self-Transcendence (e.g., benevolence, honesty, fairness) [116, 117, 118], Conservation (e.g., safety, tradition, conformity) [119, 31, 96], Openness to Change (e.g., creativity, privacy, autonomy) [120, 121, 122]. These dimensions offer a comprehensive framework for evaluating and aligning AI systems with the multifaceted nature of human values.

Interaction Techniques to Specify AI Values. This sub-research question explores how human values are interactively specified to ensure AI alignment, focusing on the techniques through which AI systems manifest or internalize these values. It identifies three main approaches: **explicit human feedback**, where values are directly communicated via principles, ratings, natural language interactions, or multimodal inputs like gestures and images [46, 47, 26]; **implicit human feedback**, where values are inferred from indirect cues such as discarded options, language patterns, theory of mind reasoning, and social relationships [48, 49, 50]; and **simulated human value feedback**, where AI systems approximate human responses using feedback simulators, comparisons to human data, or synthetically generated data [51, 52, 53]. Together, these approaches illuminate the mechanisms by which AI systems interpret and enact human values via direct and indirect human-AI interaction.

Key Takeaways. By comparing prior research with our comprehensive analysis of human values and interaction techniques, we found that existing studies are largely constrained to conventional principles and standard interaction methods, overlooking the broader spectrum of human values and the interactive approaches needed to specify them effectively in alignment.

- **RQ2: Integrating Human Specifications into AI.**

Building on the value-laden human specifications gathered through interaction, we next explore diverse methods for integrating human values into AI systems. We examine this central question across two key stages of the AI lifecycle—development and deployment—by asking: (1) how can general or customized human values be integrated throughout the AI development process? and (2) what methods and platforms are available to evaluate values during AI development?

Integrating General Values to AI. This sub-research question examines how broad, universally recognized human values are embedded into AI systems to ensure ethical alignment and societal acceptance. It outlines three key dimensions: **Instruction Data**, which includes human annotations, human-AI co-annotation, and simulated human data to guide value-based training [54, 55, 56]; **Model Learning**, where human values are integrated during training through either real-time online alignment or offline processes prior to deployment [1, 57, 58]; and **Inference Stage**, where AI systems are evaluated and refined using techniques such as prompting, external tool interactions, and response search to ensure their outputs align with predefined ethical criteria [26, 113, 59, 60]. Together, these processes aim to build AI systems that promote trust and responsible use.

Customizing AI Values. This sub-research question investigates how AI systems can be customized to reflect specific user preferences, application domains, or community values, thereby improving contextual alignment. It identifies three primary strategies: **Customized Data**, which involves curating and finetuning datasets based on socio-demographic groups, user histories, or expert selections to align models with targeted human values [61, 62, 63]; **Adapt Model by Learning**, which includes techniques such as group-based learning, active learning, adapter insertion, mixture of experts, and enhanced knowledge integration to refine model behavior [64, 65, 66]; and **Interactive Alignment**, which actively engages users through real-time feedback, steering prompts, and proactive adjustments based on user profiles to tailor AI systems to specific contexts and preferences [67, 68, 69].

Evaluating AI Systems. This sub-research question examines how the integration of human values into AI systems, particularly large language models (LLMs), is evaluated, highlighting both human-in-the-loop and automated methods. **Human-in-the-loop evaluation** involves human judgment, feedback, and collaboration to assess the ethical and value alignment of AI outputs, leveraging direct

human input or combined human-AI assessment processes [70, 71, 43]. **Automatic evaluation** utilizes computational techniques, including human simulators, standardized benchmarks, and distributional comparisons, to evaluate alignment without human intervention [72, 55, 73]. Together, these approaches aim to ensure that AI systems reflect human ethical standards and value frameworks effectively and reliably.

Ecosystem and Platforms. The ecosystem and platforms refer to the broader context in which AI systems operate and interact with other agents, platforms, or environments. This includes the infrastructure, frameworks, and technologies that support the development, deployment, and utilization of AI systems. **LLM-based Agents** are based on large language models (LLMs) such as GPT (Generative Pre-trained Transformer) models, which have been pre-trained on vast amounts of text data [56, 74, 123, 75]. **RL-based Agents** are based on reinforcement learning (RL) algorithms to learn and adapt their behavior based on feedback from the environment or human users [76]. **Annotation Platforms** refers to the ecosystems that are designed to crowdsource human demonstrations as collected data for reinforcement learning [77] and supervised finetuning learning for alignment [78].

Key Takeaways. Our systematic analysis of value integration and evaluation in AI reveals a strong focus on explicit value annotations, while implicit value expressions and behaviors are often neglected. Despite the power of general-purpose AI, methods for customizing systems to reflect individual or group values remain underexplored. Additionally, there is a lack of standardized criteria for evaluating human-in-the-loop methods and supporting platforms, underscoring the need for more robust and context-sensitive evaluation frameworks in value-aligned AI development.

3.2 Align Humans to AI

From a *long-term* perspective, it is crucial to consider the dynamic and evolving nature of human-AI alignment. This direction emphasizes a **human-centered perspective**—drawing from fields such as HCI and the social sciences—and offers guidance for **researchers and user experience designers** in addressing two core research questions: **RQ3.** *How can humans learn to perceive, explain, and critique AI?* and **RQ4.** *How do individuals and society adapt their behaviors in response to AI advancements?*

- **RQ3: Human Cognitive Adjustment to AI.**

For effective collaboration and value specification, humans need to develop a clear understanding of how AI systems function. As AI introduces various risks, fostering critical thinking is essential to prevent blind reliance. To address this, we systematically investigate, as summarized in Table 1, (1) how humans learn to perceive and understand AI, and (2) how they can engage in critical reflection on AI behavior and outputs.

Perceiving and Understanding AI. This sub-research question explores how to enhance human understanding and perception of AI systems, particularly among non-technical users, through education, training, and human-centered explanation techniques. It emphasizes the importance of **AI literacy and awareness** as foundational competencies for effective human-AI collaboration [79], supported by **explicit training courses** designed to improve users' ability to engage with AI [80, 81]. Additionally, it highlights efforts in **AI sensemaking and human-centered explanations**, including visualizations and interactive techniques, to help people interpret AI mechanisms and outputs, thereby fostering more informed and meaningful human-AI interactions [82, 83, 84].

Critical Thinking around AI. This sub-research question examines how individuals critically reflect on and evaluate AI systems by comparing their own mental models with those of the AI, focusing on the rationality, reliability, and ethical behavior of these technologies. It emphasizes the need for humans to develop critical thinking skills to identify biases, errors, and ethical concerns in AI outputs, and to audit AI systems for compliance with moral and societal standards. Key areas include building appropriate levels of **Trust and Reliance on AI** based on its competence and reliability [85, 86, 87, 124], engaging in **selective AI adoption** aligned with user needs and values [125, 126], and addressing **Ethical Considerations** through mitigation strategies and auditing [88, 89]. Additionally, it underscores the importance of **Recalibrating Cognition**, where users adjust their understanding and expectations of AI performance and reliability to foster a balanced and informed relationship with AI systems [90, 91, 92].

Key Takeaways. Our systematic review reveals a significant gap in research on training and educating humans to develop appropriate knowledge, trust, and reliance when collaborating with AI systems.

Additionally, there is limited exploration of how to critically evaluate AI across a broad spectrum of human values.

- **RQ4: Human Adaptation Behavior to AI.**

Building on our investigation of human cognitive adjustments to AI, we further examine how individuals and society can respond effectively to AI's expanding influence. To guide this inquiry, we address three key questions: (1) How do humans learn to collaborate with AI across its diverse roles? (2) In what ways are individuals and society affected by AI? and (3) How can these impacts be comprehensively assessed? The key dimensions emerging from this analysis are summarized at the bottom of Table 1.

Human-AI Collaboration Mechanisms. This category explores the diverse ways humans and AI collaborate through partnerships, co-creation, and mutual learning. **AI Assistants**, particularly those powered by LLMs, support human tasks by interpreting user demands, enhancing prompt formulation, generating creative prototypes, and aiding decision-making [93, 94, 95]. In collaborative frameworks, humans and AI function as **AI Partners**, where simulated agency and reciprocal learning enable joint decision-making and knowledge sharing [21, 96, 97]. AI delegation and mediation transform traditional human tasks, while co-design with AI treats the system as both a collaborator and a design material. Additionally, **AI Tutoring** systems enhance human learning in both technical and social domains by offering tailored feedback, adaptive instruction, and immersive practice environments, ultimately improving skill acquisition and performance [98, 99, 100].

AI Impact on Humans and Society. This category investigates the multifaceted effects of AI advancement on human behavior, attitudes, and societal dynamics, aiming to inform policy, education, and interventions. The **Impacts on Participatory Individuals and Groups** dimension focuses on how AI influences human decision-making, creativity, privacy, and authorship—shaping behaviors and raising concerns about data rights and intellectual ownership [18, 101, 127, 128]. The **Societal Concerns and AI Impacts** dimension expands this lens to the societal level, examining AI's effects on misinformation, education, social norms, and the workplace, including issues like disinformation, shifts in learning practices, changes in interpersonal relationships, and job displacement [102, 103, 129]. The **Reaction to AI Advancement** dimension explores regulatory, cultural, and institutional responses to AI, addressing how societies perceive, govern, and adapt to AI technologies [104, 105, 130]. This includes efforts to regulate bias and discrimination, develop policy frameworks, track evolving AI acceptance, ensure transparency and oversight, and establish responsible AI checklists to guide ethical and safe deployment.

Evaluation in Human Studies. This summary covers common empirical methods used to rigorously evaluate AI's impact on humans at both micro and macro levels. At the micro-level, **Human-AI Collaboration Evaluation** assesses not only task success and efficiency but also user experience, including cognitive workload, user satisfaction, control, and trust, especially in critical settings to prevent failures [106, 131, 107, 106]. Methods include quantitative interaction analytics, qualitative surveys and interviews, and statistical analyses to understand and verify user behaviors and perceptions. At the macro-level, **Societal Impact Evaluation** focuses on understanding long-term behavioral changes within large populations as AI use becomes widespread, employing large-scale public opinion surveys and behavioral data analytics over time to capture evolving patterns and societal shifts related to AI interaction [16, 108, 132].

Key Takeaways. Our review reveals that while prior studies have extensively examined how AI can assist humans across various tasks, there is limited exploration of the challenges humans face when collaborating with AI that surpass human capabilities in certain domains. Additionally, more research is needed to understand the evolving impact of AI on individuals and society at large over time.

4 Underexplored Research Gaps and Challenges

In this section, we consolidate key findings from our systematic analysis of the framework and the reviewed literature (in Figure 2. To accurately capture the distribution of existing research and identify current gaps and challenges, we quantified the number of relevant studies corresponding to each dimension in the Bidirectional Human-AI Alignment. We elaborate on key findings below.

Underexplored Dimensions in Aligning AI with Humans. Current AI alignment research has primarily focused on incorporating explicitly stated human values, often gathered through direct feedback mechanisms such as ratings, rankings, or instructions. However, several critical dimensions remain underexplored. First, the **use of implicit human feedback** — such as behavioral cues,

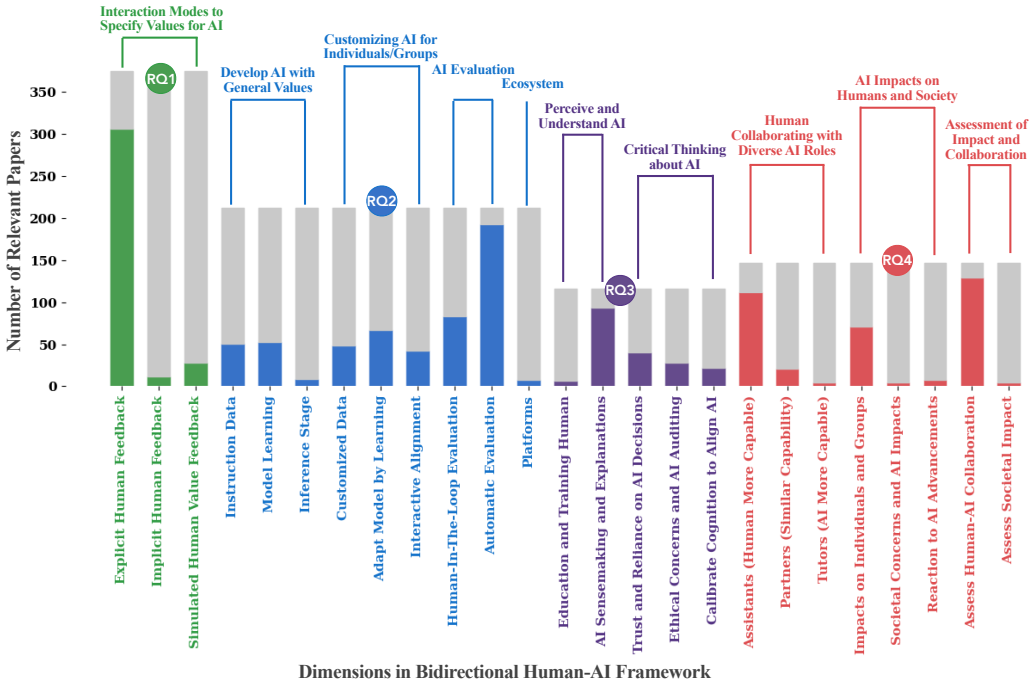

Figure 2: The number of papers for each dimension in the *bidirectional human-AI alignment* framework. Out of papers that are relevant to each research question (*i.e.,* gray bars), we show the number of papers that are relevant to each dimension (*i.e.,* color bars).

physiological signals, or interaction patterns—and simulated human value feedback has received limited attention, despite their potential to provide rich, context-sensitive information about human values. *Second*, most alignment efforts concentrate on model training stages, whereas **developing and customizing AI models** during inference or through interactive processes for embedding and evaluation values remains significantly underexplored. Enabling real-time adaptation of AI behavior to human input could enhance alignment in dynamic or personalized contexts. *Third*, **human-in-the-loop evaluation**, which involves assessing AI systems through active human participation and feedback, is rarely used in comparison to fully automated evaluation metrics. Expanding research into these areas is essential for advancing more robust, responsive, and context-aware alignment.

Underexplored Dimensions in Aligning Humans with AI. Human-centered alignment research has predominantly emphasized designing AI systems that facilitate human understanding through sensemaking and explanation—primarily by clarifying the justifications behind AI decisions to foster user trust and reliance. However, this focus often **overlooks the broader goal of fostering AI literacy**—the essential skills and competencies individuals need to understand, critique, use, and interact effectively with AI systems. Despite its foundational role in responsible AI engagement, AI literacy remains an underexplored area. Furthermore, while numerous studies have proposed interactive mechanisms and prototypes to support human-AI collaboration, they commonly assume that AI operates in a subordinate or assistive role. As AI systems grow increasingly capable, research must also **consider collaborative dynamics between humans and AI** with equal or superior capabilities. Additionally, the **ethical auditing** of AI from a human-centered perspective and the **societal-level impacts of AI** — such as changes in human behavior, social relationships, and public responses—have not been sufficiently examined. These dimensions are critical for ensuring meaningful and equitable alignment between humans and evolving AI technologies.

5 Near to Long-term Risks and Opportunities

Drawing upon insights gained from the development of our framework and the associated systematic review analysis, we propose future research aiming to achieve the long-term alignment goal by identifying three important challenges from near-term to long-term objectives, including the Specification Game, Dynamic Co-evolution of Alignment, and Safeguarding Coadaptation.

5.1 Specification Game

An important near-term challenge is resolving the "Specification Game", which involves precisely defining and implementing AI goals and behaviors to align with human intentions and values. Next, we will introduce how synergistic efforts from two directions can potentially address this challenge.

Integrate fully specified human values into aligning AI. Individuals often possess value systems that encompass multiple values with varying priorities, rather than a single value, to guide their behaviors [38, 39]. Also, these priorities can change dynamically throughout an individual's life stages. As such, It is more realistic to select values compatible with specific societies or situations, given the fact that we live in a diverse world [29]. Future research, inspired by Social Choice Theory [35], could focus on using democratic processes to aggregate individual values into collective agreements. Building on the summaries in Sections 3.1, researchers can employ democratic methods to identify diverse subsets of human values for AI alignment. Additionally, creating datasets that represent these values is crucial. Besides, it is crucial yet challenging *for AI designers to investigate how to fully specify the appropriate values and to further integrate these values into AI alignment*. Future important area involves developing algorithms, such as the Bradley-Terry Model [58] or Elo Rating System [26], to convert heterogeneous human values into AI-compatible formats for training reward models and guiding reinforcement learning. Researchers should also explore AI models capable of aligning with unstructured human data, including free-form descriptions of values, multimedia, or sensor recordings depicting human behavior.

Elicit nuanced and contextual human values during diverse interactions. Current alignment methods use instructions, ratings, and rankings to infer human values, which can not fully capture all relevant human values and constraints. Future research should focus on optimizing interactive interfaces to efficiently elicit human values. These interfaces can leverage diverse interaction modes to capture comprehensive human value information. Additionally, people often struggle to formulate optimal prompts for AI, accurately specify their requirements, and articulate their desired values, which can change based on context and time. Developing proactive interfaces that use conversational techniques to elicit nuanced and evolving values is also crucial. Implicit human signals that indicate values are also frequently overlooked. Additionally, systems that track interactions to hypothesize and validate implicit human values in real-time should be designed.

5.2 Dynamic Co-evolution of Alignment

The challenge ahead lies in comprehending and effectively navigating the dynamic interplay among human values, societal evolution, and the progression of AI technologies. Future studies in these directions aim to bolster a synergistic co-evolution between AI and human societies, adapting both to each other's changes and advancements.

Co-evolve AI with changes in humans and society. Existing literature often treats AI alignment as static, ignoring its dynamic nature. A long-term perspective must consider the co-evolution of AI, humans, and society. As AI systems evolve and scale up, they gain new capabilities, making it essential to ensure their goals remain aligned with human values. Thus, alignment solutions require continuous oversight and updates. Future research should develop methods for continuously updating AI with limited data without compromising alignment values and performance. This could involve forecasting human value evolution and preparing AI with flexible strategies like prompting or interventions. *(ii)* Additionally, AI advancements also influence human actions and values, necessitating adaptive alignment solutions. Ensuring AI co-evolves with human and societal changes is crucial for robust alignment. This challenge could potentially be addressed by forecasting the potential evolution trajectories of human values or behavioral patterns, and preparing AI with the flexibility to adapt in advance, for example, through prompting or intervention strategies.

Adapt humans and society to the latest AI advancements. While current AI systems lag behind humans in many tasks, identifying and handling AI mistakes, including knowing when to seek human intervention, remains essential. Future research should focus on developing validation mechanisms that enable humans to interpret and verify AI outputs. This could involve designing interfaces that allow humans to request step-by-step justifications from AI or integrating tools to verify the truthfulness of AI referring to Section 3.2. Additionally, developing interfaces that enable groups of humans to collaboratively validate AI outputs and creating scalable validation tools for large-scale applications are important directions. *(ii)* As AI advances, it becomes essential to develop systems that enable humans to utilize AI with capabilities surpassing their own. Research is needed to understand how individuals can interpret and validate AI outputs for tasks beyond their abilities and

leverage advanced AI sustainably, avoiding issues like job displacement or loss of purpose. Another research direction is designing strategies to enhance human capabilities by learning from advanced AI, including gaining knowledge and building skills. *(iii)* As AI integrates more into daily tasks, its impact on human values, behaviors, capabilities, and society remains uncertain. Continuous examination of AI's influence on individuals, social relationships, and broader societal changes is vital. Research should assess how humans and society adapt to AI advancements, guiding AI's future evolution. Potential areas include evaluating changes in individual behavior, social relationships, and societal governance as AI replaces traditional human skills. Understanding these dynamic changes is essential for grasping the broader impact of AI on humanity and society.

5.3 Safeguarding Co-adaptation

As AI gains autonomy and capability, the risks associated with its instrumental actions, as a means toward accomplishing its final goals, increase. These actions can be undesirable for humans. Therefore, safeguarding the co-adaptation between humans and AI is crucial. We next explore future research to address this challenge from both directions.

Specify the goals of an AI system into interpretable and controllable instrumental actions for humans. As advanced AI systems become more complex, they present greater challenges for human interpretation and control. It is crucial to empower humans to detect and interpret AI misconduct and enable human intervention to prevent power-seeking AI behavior. Research should focus on designing corrigible mechanisms for easy intervention and correction, including modular AI architectures and robust override protocols that allow human operators to halt or redirect AI activities. These components should be human-interpretable, enabling scenario testing. *(ii)* Furthermore, advanced AI systems may intentionally mislead or disobey humans, generating plausible fabrications [133]. Developing reliable interpretability mechanisms to validate the faithfulness and honesty of AI behaviors is essential. This includes correlating AI behaviors with internal neuron activity signals, akin to physiological indicators in human polygraph tests [134]. Inspecting these indicators can help humans assess the truthfulness of AI interpretations and prevent risky actions.

Empower humans to identify and intervene in AI instrumental and final strategies in collaboration. Preventing advanced AI from engaging in risky actions requires robust human supervision. Essential steps include developing training and simulation environments with scenario-based exercises and timely feedback, and creating interactive dashboards for real-time monitoring. These dashboards should feature effective data visualization, anomaly detection, and prompt alert systems for immediate intervention. *(ii)* Scalable solutions are needed for supervising AI across various applications. Real-time oversight becomes more challenging with widespread AI deployment, necessitating advanced autonomous monitoring tools. These tools should learn normal AI behavior and flag deviations immediately. Integrating training environments, interactive dashboards, and scalable diagnostic tools will enhance human ability to ensure better alignment with human values.

6 Limitations

One limitation of this work is the scope of the sampled and filtered papers. The rapidly growing literature on human-AI alignment spans diverse venues across many domains. Instead of an exhaustive collection, we focused on developing a holistic bidirectional human-AI alignment framework using essential research questions, dimensions, and codes. Our surveyed papers and team members primarily focus on computing-related fields like ML, NLP, and HCI, though alignment research also involves disciplines like cognitive science, psychology, and STS (Science, Technology, and Society). Our framework can naturally extend to these areas. Despite these limitations, we believe our bidirectional human-AI alignment framework serves as a foundational reference for future researchers.

7 Conclusion

This study clarifies the conceptual foundations of human-AI alignment by analyzing how key terms are defined and operationalized across over 400 papers from NLP, HCI, ML, and related domains. We introduce the Bidirectional Human-AI Alignment framework, which organizes alignment efforts into two interdependent directions: aligning AI with humans values, and aligning humans with AI – enabling humans to effectively understand, evaluate, and adapt to AI systems. Our analysis identifies critical gaps in current literature, including limited support for long-term interaction, underdeveloped models of human values, and challenges in mutual intelligibility. We conclude with three central challenges—specification gaming, scalable oversight, and dynamic alignment—and offer actionable recommendations to support future research aimed at fostering reciprocal, robust, and context-aware approaches to human-AI alignment.

APPENDIX

8 Systematic Literature Review

8.1 Systematic Literature Review Process

To understand the research literature relevant to the ongoing, mutual process of human-AI alignment, we performed a systematic literature review based on the PRISMA guideline [135, 136]. Figure 3 shows the workflow of our process for paper coding and developing the *bidirectional human-AI alignment* framework. We introduce the step details below.

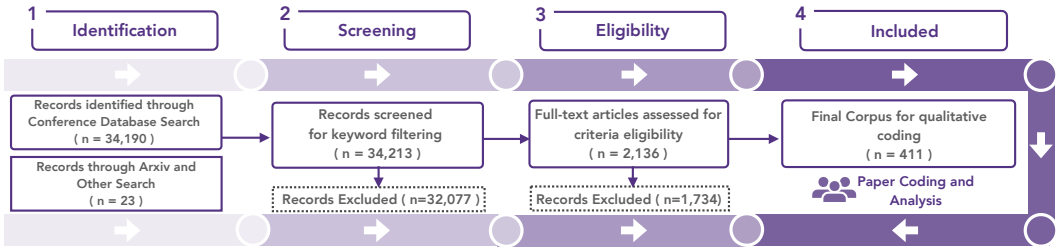

Figure 3: The selection and refinement process of our systematic literature review. We referred to the PRISMA guideline [135, 136] to report the workflow. From the identification of 34,213 records by keyword search, to screen eligible papers against our criteria and arriveg at our final corpus of 411 papers. For each of the stages where literature reviews were excluded (identification, screening, and eligibility) we further present the total of excluded records.

8.1.1 Identification and Screening with Keywords.

We started with papers published in the AI-related domain venues (including NLP, HCI, and ML fields) beginning from the advent of general-purpose generative AI to present, *i.e.,* primarily between January, 2019 and January, 2024 (see details in Appendix 8.2). We retrieved 34,213 papers in the initial *Identification* stage. Further, we collectively defined a list of keywords (see details in Appendix 8.3) and screened for papers that included at least one of these keywords (*e.g.,* human, alignment) or their variations in the title or abstract. We included 2,136 papers in *Screening* stage.

8.1.2 Assessing Eligibility with Criteria

We further filtered the 2,136 papers based on explicit inclusion and exclusion criteria, i.e., the *Eligibility* stage. Our criteria revolved around six research questions that we collectively identified to be most pertinent to the topic, including *1) what essential human values have been aligned by some AI models? 2) how did we effectively quantify or model human values to guide AI development? 3) what strategies have been employed to integrate human values into the AI development process? 4) how did existing studies improve human understanding and evaluation of AI alignment? 5) what are the practices for designing interfaces and interactions that facilitate human-AI collaboration? 6) How have AI been adapted to meet the needs of various human value groups?* We included papers that could potentially answer any of these questions. Further, based on the scope in Section **??**, we excluded papers that did not meet our inclusion criteria. This resulted in a final corpus of 411 papers, which were analyzed in detail using qualitative coding (see Appendix 8.4 for more details).

8.1.3 Qualitative Code Development.

Referring to the code development process in [20], we first conducted qualitative coding for each paper by identifying relevant sentences that could answer the above research questions, and entering short codes to describe them into a codebook. We iteratively coded relevant sentences from each paper through a mix of inductive and deductive approaches, which allowed flexibility to expand, modify or change the driving research questions based on our learnings as we went through the process. To ensure rigor in our coding process, two authors coded each paper. The first author independently annotated all papers after reviewing the paper abstracts and introductions. Twelve team members

each annotated a subset of the paper corpus. Our corpus includes papers from different domains (*e.g.,* HCI, NLP and ML). Therefore, we divided the authors into HCI and NLP/ML[2] teams and assigned the papers accordingly based on expertise. All team members coded each of their assigned papers to answer all six questions (if applicable) introduced above.

8.1.4 Framework Development and Rigorous Coding.

After developing annotations, all authors collaborated to create the bidirectional human-AI alignment framework by integrating the annotations within each of the codes. The initial version of the framework was proposed by the author who reviewed all papers. This framework furthermore underwent iterative improvement through: *1)* discussions with all team members involved in paper coding, and *2)* revisions based on feedback from the project advisors. Additionally, we strengthened the framework by reviewing papers from the AI Ethics conferences (including FAccT and AIES), and related work of the collected papers that covered other domains such as psychology and social science. We further added missing codes and papers to ensure comprehensive coverage (see Appendix 8.2 for details). The final bidirectional human-AI alignment framework, with detailed topologies, is presented in Section 3. Following the framework's finalization, we conducted another separate coding process to annotate *whether each paper investigated dimensions within our framework*. Two authors independently coded each paper.[3] These codes were then used to perform quantitative and qualitative analyses, as presented in Section **??**.

8.2 Venues

We primarily focused on papers from the fields of HCI, NLP, and ML ranging from year 2019 to 2024 January. We included all their papers tracks (*e.g.,* CSCW Companion and Findings) without including workshops of conferences. From the ACL Anthology, OpenReview and ACM Digital Library, we retrieved 34,190 papers into a Reference Manager Tool (*i.e.,* Paperpile). Particularly, the venues we surveyed are listed below.

- **HCI**: CHI, CSCW, UIST, IUI;
- **NLP**: ACL, EMNLP, NAACL, Findings
- **ML**: ICLR, NeurIPS
- **Others**: ArXiv, FAccT, AIES, and other related work

Additionally, we also consolidate the framework by reviewing the papers published in FAccT and AIES (*i.e.,* important venues for AI Ethics research) between 2019 and 2024 and supplemented the codes, including the AI Regulatory and Policy code in Section **??** and the exemplary paper of Regulating ChatGPT [105]), which were not covered by the original collections. Also, we include a number of papers in the "Other" class are found by related work that are highly relevant to this topic.

8.3 Keywords

We decided on a list of keywords relevant to bidirectional human-AI alignment. The detailed keywords include:

- **Human**: Human, User, Agent, Cognition, Crowd
- **AI**: AI, Agent, Machine Learning, Neural Network, Algorithm, Model, Deep Learning, NLP
- **LLM**: Large Language Model, LLM, GPT, Generative, In-context Learning
- **Alignment**: Align, Alignment
- **Value**: Value, Principle
- **Trust**: Trust, Trustworthy
- **Interact**: Interact, Interaction, Interactive, Collaboration, Conversational

[2]Note that NLP and ML are two different domains, we combine them together for the purposes of literature review analysis since they both work on developing and evaluating AI technologies.

[3]The joint probability of agreement for the paper annotations was 0.78.

- **Visualize**: Visualization, Visualize
- **Explain**: Interpretability, Explain, Understand, Transparent
- **Evaluation**: Evaluate, Evaluation, Audit
- **Feedback**: Feedback
- **Ethics**: Bias, Fairness

8.4 Inclusion and Exclusion Criteria

To further filter the most relevant papers among the keyword-filtered 2136 papers, we identified the six most important research questions we are interested in. We primarily selected the potential papers that can potentially address these six questions after reviewing their title and abstracts. The six topics of research questions in our filtering include:

RQ.1 **[human value category]** What essential human values have been aligned by some AI models?

RQ.2 **[quantify human value]** How did we effectively quantify or model human values to guide AI development?

RQ.3 **[integrate human value into AI]** What strategies have been employed to integrate human values into the AI development process?

RQ.4 **[assess / explain AI regarding human values]** How did existing studies improve human understanding and evaluation of AI alignment?

RQ.5 **[human-AI interaction techniques]** What are the practices for designing interfaces and interactions that facilitate human-AI collaboration?

RQ.6 **[adapt AI for diverse human values]** How has AI been adapted to meet the needs of various human value groups?

Particularly, we provide elaborated inclusion and exclusion criteria during our paper selection as listed below. We are aware that we have limitations during our paper filtering process.

Inclusion Criteria:

- [**Human values**] we include papers that study human value definition, specification and evaluation in AI systems.
- [**AI development techniques**] We include techniques of developing AI that aim to be more consistent with human values with interactions along all AI development stages (e.g., data collection, model construction, etc.)
- [**AI evaluation, explanation and utilization**] we include papers that build human-AI interactive systems or conduct human studies to better evaluate, explain, and utilize AI systems.
- [**building dataset with human interaction**] especially responsible dataset.

Exclusion Criteria:

- [**Alignment not between human & AI**] we do not include alignment studies that are not between human and AI, such as entity alignment, cross-lingual alignment, cross-domain alignment, multi-modal alignment, token-environment alignment, etc.
- [**AI models beyond LLMs - Modality**] we do not focus on AI models other than LLMs (e.g., 3D models, VR/AR, voice assistant, spoken assistant), our primary model modality is text. Specifically, we do not consider audio / video data; we do not consider pure computer vision modality.
- [**No human-AI interaction**] we do not consider studies that do not involve the interaction between human and AI, such as (multi-agent) reinforcement learning. Specifically, we do not consider interactions via voices/speech, Do not consider game interaction; Do not consider interaction for Accessibility; Do not consider Mobile interaction; Not consider autonomous vehicle interaction wearable devices, or Physical interaction;
- [**Tasks**] art and design, emotion.

- [**No human included**]
- [**focus on English**] primarily focus on English as the main language;
- [**Application**] not include the NLP papers tailored for a specific traditional task, such as translation, entity recognition, sentiment analysis, knowledge graph, adversarial and defense, topic modeling, detecting AI generations, distillation, low resource, physical robots, text classification, games, image-based tasks, hate speech detection, Human Trafficking, etc.
- [**Visualizing Embeddings**] Visualizing/interacting transformer embeddings?
- [**Embedding-based**] explanation, evaluation, etc.
- [**multi-agent reinforcement learning with self-play and population play**] techniques, such as self-play (SP) or population play (PP), produce agents that overfit to their training partners and do not generalize well to humans.

We acknowledge the extensive scope and rapid advancements of research in this area, and posit that our study offers insights that can be generalized to various modalities. For example, the value taxonomy and human-in-the-loop evaluation paradigm outlined in our framework can be applied to both text-based and other modality-based (*e.g.,* vision, robotics) models. It's worth noting that our literature review does not aim to exhaustively cover all papers in the field, which is impossible given the rapid advancement of human-AI alignment research. Instead, we adopt a human-centered perspective to review more than 400 key studies in this domain, focusing on delineating the framework landscape, identifying limitations, future directions, and a roadmap to pave the way for future research.

9 Selected Paper List

- **Human-Centered Studies**: [137, 107, 138, 92, 139, 140, 141, 142, 143, 144, 145, 96, 146, 147, 148, 149, 150, 151, 152, 86, 117, 126, 153, 154, 155, 156, 87, 157, 158, 159, 160, 161, 162, 104, 163, 164, 165, 166, 167, 168, 169, 32, 170, 171, 172, 173, 174, 175, 176, 3, 177, 178, 179, 180, 181, 89, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 122, 195, 196, 85, 91, 88, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 121, 81, 211, 83, 212, 213, 214, 215, 90, 216, 112, 217, 218, 219, 220, 221, 222, 223, 95, 224, 21, 225, 226, 227, 94, 228, 229, 230, 231, 93, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 79, 46, 257, 125, 137]
- **AI-Centered Studies**
  [258, 118, 78, 52, 24, 25, 259, 75, 48, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 69, 270, 271, 272, 273, 274, 44, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 57, 288, 289, 290, 291, 292, 293, 120, 294, 73, 295, 49, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 84, 307, 308, 309, 310, 311, 61, 70, 312, 313, 71, 314, 315, 316, 317, 41, 43, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 31, 330, 331, 332, 333, 66, 334, 335, 27, 59, 336, 60, 337, 338, 33, 76, 72, 115, 339, 340, 341, 74, 342, 113, 343, 344, 345, 50, 346, 347, 348, 349, 350, 351, 45, 47, 352, 353, 354, 355, 356, 357, 358, 62, 359, 360, 361, 362, 363, 54, 364, 365, 53, 366, 40, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 56, 55, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 42, 411, 77, 1, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 64, 63, 422, 423, 424, 425, 67, 68, 426, 427, 428, 58]
- **Others** [429, 28, 430, 16, 5, 26, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 101, 51, 82, 132, 443, 131, 129, 103, 105, 65, 444, 445, 446, 447, 98, 100, 18, 127, 102, 130]

10 Alignment Goals and Human Values

10.1 A Comprehensive Taxonomy of Human Values

This conventional theory was developed without the context of human-AI interaction, which might overlook values that need to be considered for human-AI alignment. Therefore, we used a *bottom-up* approach to extract all values studied in our collected alignment literature, mapped them onto the

	Goals	Definitions	Limitations / Risks
The Goal of Alignment	**Instructions**	The agent does what I instruct it to do.	On a larger scale, it is difficult to precisely specify a broad objective that captures everything we care about, so in practice the agent will probably optimise for some proxy that is not completely aligned with our goal.
	Intentions or (Expressed Intentions)	The agent does what I intend it to do.	It is quite possible for intentions to be irrational or misinformed, or for the principal to form an intention to do harmful or unethical things.
	Preferences or (Revealed Preferences)	The agent does what my behaviour reveals I prefer.	1) People have preferences for things that harm them. 2) People have preferences about the conduct of other people. 3) Preferences are not a reliable guide to what people really want or deserve due to adaptiveness.
	Desires or (Informed Preferences)	The agent does what I would want it to do if I were rational and informed.	Researchers would have to apply a corrective lens or filter to the preferences they actually observe. As a consequence, the approach is no longer strictly empiricist.
	Interest or (Well-being)	The agent does what is in my interest, or what is best for me, objectively speaking.	Something in a human's interest does not mean he/she ought to do it or is morally entitled to do so, such as an interest in stealing. Also, it is hard to manage trade-offs the collective interests of different people.
	Values	The agent does what it morally ought to do, as defined by the individual or society.	Current the best possibility, but it still encounters two difficulties of 1) specifying what values or principles, and 2) concerning the body of people who select the principles with which AI aligns.

Table 2: The **Goals** of Alignment. We present the six prevailing alignment goals, associating with their Definitions (middle column), Limitations and Risks (right column). We consider **Human Values** as the main goal of alignment in this work referring to an extensive analysis and arguments in existing studies [29, 30]

Schwartz Theory of Basic Values, and supplemented the theory with AI-related structure and content. As a result, we identified the structural relationships among human values and mapped existing literature to a fine-grained taxonomy (see Table 3). We supplemented the traditional theory's four high-order value types (*i.e.,* "Self-Enhancement", "Openness to Change", "Conservation", "Self-Transcendence") with a novel high-order value type, named "Desired Values for AI Tools" that encompasses two motivational value types (*i.e.,* "Usability" and "Human-Likeness"). We further organize the relationship among these value types along two dimensions [39]: different resources (*i.e.,* individuals, society and interaction) and different self-intentions (*i.e.,* self-protection against threat and self-expansion and growth). Furthermore, we elaborate the definitions of the 12 motivational value types and their exemplary values by mapping them to relevant human-AI alignment papers from our corpus in Table 3. During the process of mapping, we found: 1) value terms in empirical papers were often named differently (*e.g.,* capability and competence), or check opposites (*e.g.,* fairness and bias); 2) there are many values not studied in our corpus, *i.e.,* indicated as (•) in the Figure.

10.2 Insights into Human Values for Alignment

Our analysis, based on the adaptation of Schwartz's Theory of Basic Values and our comprehensive literature review, identifies three critical findings for future research:

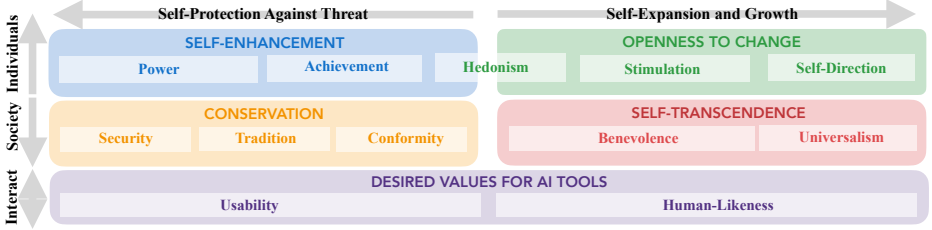

Table 3: The value relations and taxonomy. We consider 5 high-order value types encompassing 12 motivational value types, indicated by their sources (e.g., individuals, society and interaction).

Value Prioritization in AI Systems. Human value systems are not merely subsets of values, but ordered systems with relative priorities [38, 39]. For instance, [39] presented the definition for this phenomenon: *"a value is ordered by importance relative to other values to form a system of value priorities. The relative importance of multiple values guides action....The trade-off among relevant, competing values guides attitudes and behaviors."*. Current AI alignment algorithms, often based on datasets of human preferences [1, 287, 58], may inadvertently prioritize majority values, potentially neglecting those of marginalized groups [349]. Future research should address this complex interplay of values in AI systems.

Universal vs. Personalized AI Values. While certain values are universally expected from AI (e.g., capability, equity, responsibility), others may be undesirable in specific contexts [33] (e.g., seeking power). Simultaneously, AI models should be adaptable to diverse human value systems [28]. Research is needed to develop methods for identifying appropriate value sets for specific individuals or groups, and for customizing AI to align with user values while maintaining ethical principles.

Disparities in Value Expectations and Evaluation. The fundamental differences between humans and AI necessitate distinct approaches to value evaluation. For instance, assessing AI honesty may require mechanistic interpretability [448], a more rigorous standard than that applied to humans. Future studies should explore methods for evaluating and explaining AI values and calibrating human expectations accordingly.

11 Interaction Techniques for Specifying Human Values

Our research reveals disparities in interaction techniques for human-AI value alignment across AI-centered (NLP/ML) and human-centered (HCI) domains. As depicted in Figure 4, this analysis focuses on three key areas:

Domain-Specific Interaction Techniques. The interaction techniques in AI-centered (NLP/ML) and Human-centered (HCI) alignment studies are often differ [449]. NLP/ML studies primarily utilize numeric and natural language-based techniques. Also, NLP/ML research explore implicit feedback to extract human hidden feedback. In contrast, HCI research encompasses a broader range of graphical and multi-modal interaction signals (e.g., sketches, location information) beyond text and images. This disparity suggests potential gaps in extracting comprehensive human behavioral information.

Stage-Specific Interaction Techniques. In NLP/ML, the learning stage predominantly employs rating and ranking interactions for alignment in dataset generation. However, when humans use AI in the inference stage, as demonstrated in HCI research, involves more diverse user interactions. This discrepancy highlights the need for alignment between model development and practical deployment.

Divergent Data Utilization. NLP/ML typically uses interaction outputs as training datasets, while HCI analyzes this data to understand human behavior and feedback. As AI systems evolve, developing new interaction modes to capture a broader spectrum of human expression becomes crucial.

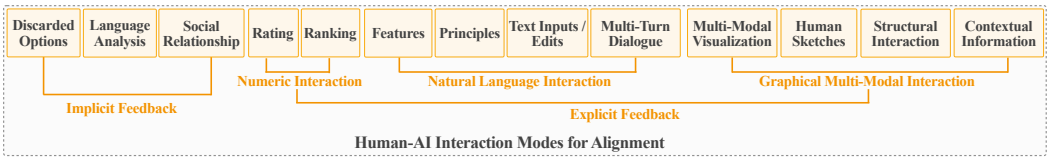

Figure 4: The interaction techniques for specifying values in human-AI alignment.

Takeaways of Interaction Techniques for Alignment.
1. Some common human feedback styles used in NLP/ML are not often studied in HCI.
2. Diverse human interactive feedback in HCI are not fully used in AI development in NLP/ML.

12 Challenges in Achieving Alignment

The concept of *alignment* in AI research has a long history, tracing back to 1960, when AI pioneer Norbert Wiener [450] described the AI alignment problem as: "*If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively ... we had better be quite*

sure that the purpose put into the machine is the purpose which we really desire." Discussion around intelligent agents and the associated concerns relating to ethics and society have emerged since then [451]. Next, we discuss the well-known challenges encountered in achieving alignment.

Challenge 1: Outer and Inner Alignment. In the context of "intelligent agents," until now, *AI alignment* research has aimed to ensure that any AI systems that would be set free to make decisions on our behalf would act appropriately and reduce unintended consequences [452, 451, 453]. At the *near-term* stage, aligning AI involves two main challenges: carefully *specifying* the purpose of the system (*outer alignment i.e.,* providing well-specified rewards [33]) and ensuring that the system adopts the specification robustly (*inner alignment*, *i.e.,* ensuring that every action given an agent in a particular state learns desirable internally-represented goals [33]). Significant efforts have been made, for *inner alignment*, to align AI systems to follow alignment goals of an individual or a group (*e.g.,* instructions, preferences, values, and/or ethical principles) [1] and to evaluate the performance of alignment [16]. However, for *outer alignment*, AI designers are still facing difficulties in specifying the full range of desired and undesired alignment goals of humans.

Challenge 2: Specification Gaming. To learn human alignment goals, AI designers typically provide an objective function, instructions, reward function, or feedback to the system, which is often unable to completely specify all important values and constraints that a human intended [6]. Hence, AI designers resort to easy-to-specify proxy goals such as *maximizing the approval of human overseers* [4], which results in "*specification gaming*" [7] or "*reward hacking*" [8] issues (*i.e.,* AI systems can find loopholes that help them accomplish the specific objective efficiently but in unintended, possibly harmful ways). Additionally, the black-box nature of neural networks brings additional ethical and safety concerns for alignment because humans don't know about the inner state and the actions AI leveraged to achieve the output. Consequently, AI systems might make "correct" decisions with "incorrect" reasons, which are difficult to discern. Society is already facing these issues, such as data privacy [9], algorithmic bias [10], self-driving car accidents [11], and more. As a result, these considerations necessitate considering human-AI interaction in AI alignment for specification and evaluation, ranging from addressing problems around who uses an AI system, with what goals to specify, and if the AI system perform its intended function from the user's perspective.

Challenge 3: Scalable Oversight. From a long-term perspective, when advanced AI systems become more complex and capable (*e.g.,* AGI [12]), it becomes increasingly difficult to align them to human values through human feedback. Evaluating complex AI behaviors applied to increasingly challenging tasks can be slow or infeasible for humans to ensure all sub-steps are aligned with their values [5]. Therefore, researchers have begun to investigate how to reduce the time and effort for human supervision, and how to assist human supervisors, referred to as *Scalable Oversight* [13].

Challenge 4: Dynamic Nature. As AI systems become increasingly powerful, the alignment solutions must also adapt dynamically since human values and preferences change as well. As [14] posit, AI systems may be neither humane nor desirable if we do not ask questions about the long-term cognitive and social effects of social agent systems (*e.g., how will agent technology affect human cognition*). All these considerations call for a long-term and dynamic perspective to address human-AI alignment as an ongoing, mutual process with the collective efforts of cross-domain expertise.

Challenge 5: Existential Risk. Further, some AI researchers claim that [454] advanced AI systems will begin to seek power over their environment (*e.g.,* humans) once deployed in real-world settings, as such behavior may not be noticed during training. For example, some language models seek power in text-based social environments by gaining money, resources, or social influence [455]. Consequently, some hypothesize that future AI, if not properly aligned with human values, could pose an *existential risk* to humans [456].

13 Author contributions

This project was a team effort, built on countless contributions from everyone involved. To acknowledge individual authors' contributions and enable future inquiries to be directed appropriately, we followed the ACM's policy on authorship [457] and listed contributors for each part of the paper.

13.1 Overall Author List and Contributions

Project Lead

The project lead initialized and organized the project, coordinated with all authors, participated in the entire manuscript.

- **Hua Shen (NYU Shanghai, New York University, huashen@nyu.edu)**: Initiated and led the overall project, prepared weekly project meetings, filtered papers, designed dimensions and codes (initial, revision), coded all papers, initiated the framework and developed human value and interaction modes analysis figures, participated in drafting all sections, paper revision and polishing.

Team Leads

The team leads organized all team events, coordinated with leads and members, contributed to a portion of manuscript.

- **Tiffany Knearem (MBZUAI, tiffany.knearem@mbzuai.ae.ac)**: Led the HCI team, prepared weekly team meetings, filtered papers, designed dimensions and codes (initial, revision), coded partial papers, ideated the framework and analysis and future work content, participated in writing (Critical Thinking and AI Impact on Human sections), paper revision and polishing.
- **Reshmi Ghosh (Microsoft, reshmighosh@microsoft.com)**: Led the NLP/AI team, prepared weekly team meetings, filtered papers, coded partial papers, ideated the framework and analysis and future work content, participated in writing (AI evaluation section), paper revision and polishing.

Team Members (Alphabetical)

The team members contributed to a portion of paper review, regular discussions, and drafted a portion of the manuscript.

- **Kenan Alkiek (University of Michigan, kalkiek@umich.edu)**: filtered papers, coded partial papers, data processing and analysis, ideated paper analysis and future work, paper revision and polishing, mainly involved in NLP Team
- **Kundan Krishna (Apple, kundank@andrew.cmu.edu)**: filtered papers, coded partial papers, ideated the framework and future work, participated in writing (Customizing AI section), designed dimensions and codes (initial, revision), paper revision and polishing, mainly involved in NLP Team
- **Yachuan Liu (University of Michigan, yachuan@umich.edu)**: filtered papers, coded partial papers, participated in writing (revised Integrate General Value and Customization content sections), paper revision and polishing, mainly involved in NLP Team
- **Ziqiao Ma (University of Michigan, marstin@umich.edu)**: filtered papers, coded partial papers, designed dimensions and codes (initial, revision), developed Human Value category, participated in writing (Human Value taxonomy, revised representation, and value gap analysis sections), paper revision and polishing, mainly involved in NLP Team
- **Savvas Petridis (Google PAIR, petridis@google.com)**: filtered papers, coded partial papers, ideated the interaction-related analysis and future work, participated in writing (Perceive and Understand AI), paper revision and polishing, mainly involved in HCI Team
- **Yi-Hao Peng (Carnegie Mellon University, yihaop@cs.cmu.edu)**: filtered papers, coded partial papers, participated in writing (Human-AI Collaboration section), paper revision and polishing, mainly involved in HCI Team
- **Li Qiwei (University of Michigan, rrll@umich.edu)**: filtered papers, coded partial papers, ideated the interaction-related taxonomy and analysis, participated in writing (Interaction Mode section), mainly involved in HCI Team
- **Sushrita Rakshit (University of Michigan, sushrita@umich.edu)**: filtered papers, coded partial papers, participated in writing (Integrate General Value section), paper revision and polishing, mainly involved in NLP and HCI Team
- **Chenglei Si (Stanford University, clsi@stanford.edu)**: filtered papers, coded partial papers, designed dimensions and codes (initial, revision), ideated the framework and future

work, participated in writing (Assessment of Collaboration and Impact section), paper revision and polishing, mainly involved in HCI Team

- **Yutong Xie (University of Michigan, yutxie@umich.edu)**: filtered papers, coded partial papers, designed dimensions and codes (initial, revision), ideated the value representation taxonomy, participated in writing (Human Value Representation section), paper revision and polishing, , mainly involved in NLP Team

Advisors (Alphabetical)

The advisors involved in and made intellectual contributions to essential components of the project and manuscript.

- **Jeffrey P. Bigham (Carnegie Mellon University, jbigham@cs.cmu.edu)**: contributed to the framework on aligning human to AI direction, vision on the status quo of alignment research, and future work discussions, and participated in paper revision and proofreading.
- **Frank Bentley (Google, fbentley@google.com)**: contributed to the historical context and project objectives, improved the definitions and design of research methodology, and participated in paper revision and proofreading.
- **Joyce Chai (University of Michigan, chaijy@umich.edu)**: iteratively involved in developing and revising definitions and the framework on aligning AI to human direction, advised on analysis and future work, and participated in paper revision and proofreading.
- **Zachary Lipton (Carnegie Mellon University, zlipton@cmu.edu)**: contributed insights from Machine Learning, NLP, and AI fields to revise the definitions and framework on aligning AI to human direction, and participated in paper revision and proofreading.
- **Qiaozhu Mei (University of Michigan, qmei@umich.edu)**: contributed insights from Data Science, Machine Learning, and NLP fields to improve definitions and the framework on aligning AI to human direction, and participated in paper revision and proofreading.
- **Rada Mihalcea (University of Michigan, mihalcea@umich.edu)**: involved in framing and revising the structure and taxonomy of human values, and contributed to improving the manuscript's title, introduction, and other sections, and participated in paper revision and proofreading.
- **Michael Terry (Google Research, michaelterry@google.com)**: contributed arguments and vision on the status quo of alignment research, framed project objectives and contributions, improved definitions and data analysis, and participated in paper revision and proofreading.
- **Diyi Yang (Stanford University, diyiy@stanford.edu)**: involved in improving definitions and the framework, contributed social insights to the work, and participated in paper revision and proofreading.

Project Leading Advisors

The project leading advisors actively involved in the entire project process and all manuscript sections.

- **Meredith Ringel Morris (Google DeepMind, merrie@google.com)**: iteratively involved in drafting all sections, contributed to core argument ideation, framework and definition improvement, provided future work insights, and participated in paper drafting, revision, and proofreading on all sections.
- **Paul Resnick (University of Michigan, presnick@umich.edu)**: actively involved and advised on the entire project process, including initiating the project and research agenda, iteratively improved definitions, framework, and analysis, and participated in paper revision and proofreading.
- **David Jurgens (University of Michigan, jurgens@umich.edu)**: provided advice throughout the project, including iterative discussions on project milestones and content ideation, organized several meetings to receive feedback from external audiences, and participated in paper revision and proofreading.

ACKNOWLEDGMENTS

We thank Eric Gilbert for his constructive feedback on human-centered insights on alignment, thank Eytan Adar for his valuable guidance on designing the interaction techniques for human-AI alignment, and thank Elizabeth F. Churchill for her insightful discussion on this manuscript. We also thank Michael S Bernstein, Denny Zhou, Cliff Lampe, and Nicole Ellison for their encouraging feedback on this work. We welcome researchers' constructive discussions and interdisciplinary efforts to achieve long-term and dynamic human-AI alignment collaboratively in the future. This work was supported in part by the National Science Foundation under Grant No. IIS-2143529 and No. IIS-1949634.

References

[1] 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.

[2] Humans are biased. Generative AI is even worse., 6 2023.

[3] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. Human-ai interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–15, 2021.

[4] Wikipedia. AI alignment — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=AI%20alignment&oldid=1220304776`, 2024. [Online; accessed 05-May-2024].

[5] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. Ai alignment in the design of interactive ai: Specification alignment, process alignment, and evaluation support. *arXiv:2311.00710*, 2023.

[6] Thomas A Hemphill. Human compatible: Artificial intelligence and the problem of control, 2020.

[7] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity. *DeepMind Blog*, 3, 2020.

[8] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. 2022.

[9] Catherine Tucker, A Agrawal, J Gans, and A Goldfarb. Privacy, algorithms, and artificial intelligence. *The economics of artificial intelligence: An agenda*, pages 423–437, 2018.

[10] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126, 2016.

[11] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

[12] Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi: Operationalizing progress on the path to agi, 2024.

[13] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv:1606.06565*, 2016.

[14] Kerstin Dautenhahn, Chrystopher L Nehaniv, and K Dautenhahn. Living with socially intelligent agents. *Human Cognition and Social Agent Technology, John Benjamins Publ. Co*, pages 415–426, 2000.

[15] Nitesh Goyal, Minsuk Chang, and Michael Terry. Designing for human-agent alignment: Understanding what humans want from their agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2024.

[16] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[17] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions. *arXiv:2405.17713*, 2024.

[18] Hua Shen and Ting-Hao Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 168–172, 2020.

[19] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. The value, benefits, and concerns of generative ai-powered assistance in writing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2024.

[20] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

[21] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[22] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behaviors? In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024.

[23] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv:2404.09932*, 2024.

[24] Hritik Bansal, John Dang, and Aditya Grover. Peering through preferences: Unraveling feedback acquisition for aligning large language models. 2023.

[25] Mudit Verma and Katherine Metcalf. Hindsight priors for reward learning from human preferences. In *The Twelfth International Conference on Learning Representations*, 2023.

[26] 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:2212.08073*, 2022.

[27] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

[28] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv:2402.05070*, 2024.

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

[30] Stuart Russell. White paper: Value alignment in autonomous systems. November 1, 2014.

[31] 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! In *The Twelfth International Conference on Learning Representations*, 2023.

[32] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[33] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

[34] Advait Deshpande and Helen Sharp. Responsible ai systems: who are the stakeholders? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 227–236, 2022.

[35] Kenneth J Arrow. *Social choice and individual values*, volume 12. Yale university press, 2012.

[36] 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.

[37] Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. *arXiv:2011.00620*, 2020.

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

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

[40] Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. From values to opinions: Predicting human behaviors and stances using value-injected large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15539–15559, 2023.

[41] Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, 2022.

[42] Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. When to make exceptions: Exploring language models as accounts of human moral judgment. volume 35, pages 28458–28473, 2022.

[43] Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, 2022.

[44] Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. Moca: Measuring human-language model alignment on causal and moral judgment tasks. *Advances in Neural Information Processing Systems*, 36, 2023.

[45] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. Nlpositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, 2023.

[46] Savvas Petridis, Benjamin D Wedin, James Wexler, Mahima Pushkarna, Aaron Donsbach, Nitesh Goyal, Carrie J Cai, and Michael Terry. Constitutionmaker: Interactively critiquing large language models by converting feedback into principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 853–868, 2024.

[47] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

[48] Arman Isajanyan, Artur Shatveryan, David Kocharian, Zhangyang Wang, and Humphrey Shi. Social reward: Evaluating and enhancing generative AI through million-user feedback from an online creative community. In *The Twelfth International Conference on Learning Representations*, 2024.

[49] Sunghyun Park, Han Li, Ameen Patel, Sidharth Mudgal, Sungjin Lee, Young-Bum Kim, Spyros Matsoukas, and Ruhi Sarikaya. A scalable framework for learning from implicit user feedback to improve natural language understanding in large-scale conversational ai systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6054–6063, 2021.

[50] Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. Towards boosting the open-domain chatbot with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4060–4078, 2023.

[51] Savvas Petridis, Ben Wedin, Ann Yuan, James Wexler, and Nithum Thain. Constitutionalexperts: Training a mixture of principle-based prompts. *arXiv:2403.04894*, 2024.

[52] Zibin Dong, Yifu Yuan, Jianye HAO, Fei Ni, Yao Mu, YAN ZHENG, Yujing Hu, Tangjie Lv, Changjie Fan, and Zhipeng Hu. Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model. In *The Twelfth International Conference on Learning Representations*, 2024.

[53] Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. Aligning large language models through synthetic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13677–13700, 2023.

[54] Hua Shen, Vicky Zayats, Johann Rocholl, Daniel Walker, and Dirk Padfield. Multiturncleanup: A benchmark for multi-turn spoken conversational transcript cleanup. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9903, 2023.

[55] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. volume 36, 2023.

[56] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. volume 36, 2023.

[57] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv:2304.06767*, 2023.

[58] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. volume 36, 2023.

[59] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*, 2024.

[60] Herbie Bradley, Andrew Dai, Hannah Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Grégory Schott, and Joel Lehman. Quality-diversity through ai feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

[61] Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, 2023.

[62] Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023.

[63] Chenyang Lyu, Linyi Yang, Yue Zhang, Yvette Graham, and Jennifer Foster. Exploiting rich textual user-product context for improving personalized sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1419–1429, 2023.

[64] Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[65] Markus Peschl, Arkady Zgonnikov, Frans A Oliehoek, and Luciano C Siebert. Moral: Aligning ai with human norms through multi-objective reinforced active learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1038–1046, 2022.

[66] Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuan-Jing Huang. Bertscore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, 2022.

[67] Jessica Maghakian, Paul Mineiro, Kishan Panaganti, Mark Rucker, Akanksha Saran, and Cheng Tan. Personalized reward learning with interaction-grounded learning (igl). In *The Eleventh International Conference on Learning Representations*, 2023.

[68] Phillip Swazinna, Steffen Udluft, and Thomas Runkler. User-interactive offline reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[69] Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, et al. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10383–10405, 2023.

[70] Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, 2023.

[71] Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. The shifted and the overlooked: A task-oriented investigation of user-gpt interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393, 2023.

[72] Xiaogang Jia, Denis Blessing, Xinkai Jiang, Moritz Reuss, Atalay Donat, Rudolf Lioutikov, and Gerhard Neumann. Towards diverse behaviors: A benchmark for imitation learning with human demonstrations. In *The Twelfth International Conference on Learning Representations*, 2024.

[73] Roma Patel and Ellie Pavlick. "was it "stated" or was it "claimed"?: How linguistic bias affects generative language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10080–10095, 2021.

[74] Binfeng Xu, Xukun Liu, Hua Shen, Zeyu Han, Yuhan Li, Murong Yue, Zhiyuan Peng, Yuchen Liu, Ziyu Yao, and Dongkuan Xu. Gentopia. ai: A collaborative platform for tool-augmented llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 237–245, 2023.

[75] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *In the International Conference on Learning Representations*, 2023.

[76] Yifu Yuan, Jianye Hao, Yi Ma, Zibin Dong, Hebin Liang, Jinyi Liu, Zhixin Feng, Kai Zhao, and Yan Zheng. Uni-rlhf: Universal platform and benchmark suite for reinforcement learning with diverse human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

[77] Matthias Gerstgrasser, Rakshit Trivedi, and David C Parkes. Crowdplay: Crowdsourcing human demonstrations for offline learning. In *International Conference on Learning Representations*, 2021.

[78] Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2022.

[79] Duri Long and Brian Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–16, 2020.

[80] Ketan Paranjape, Michiel Schinkel, Rishi Nannan Panday, Josip Car, Prabath Nanayakkara, et al. Introducing artificial intelligence training in medical education. *JMIR medical education*, 5(2):e16048, 2019.

[81] Nora McDonald and Shimei Pan. Intersectional ai: A study of how information science students think about ethics and their impact. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–19, 2020.

[82] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. Sensible ai: Re-imagining interpretability and explainability using sensemaking theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 702–714, 2022.

[83] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[84] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, 2020.

[85] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[86] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[87] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422, 2023.

[88] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[89] Jack Bandy. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1):1–34, 2021.

[90] Ziyao He, Yunpeng Song, Shurui Zhou, and Zhongmin Cai. Interaction of thoughts: Towards mediating task assignment in human-ai cooperation with a capability-aware shared mental model. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[91] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[92] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[93] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–22, 2022.

[94] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L Glassman, and Toby Jia-Jun Li. Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[95] Sherry Wu, Hua Shen, Daniel S Weld, Jeffrey Heer, and Marco Tulio Ribeiro. Scattershot: Interactive in-context example curation for text transformation. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 353–367, 2023.

[96] Alexey Zagalsky, Dov Te'eni, Inbal Yahav, David G Schwartz, Gahl Silverman, Daniel Cohen, Yossi Mann, and Dafna Lewinsky. The design of reciprocal learning between human and artificial intelligence. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–36, 2021.

[97] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 2010.

[98] Qianou Ma, Hua Shen, Kenneth Koedinger, and Tongshuang Wu. How to teach programming in the ai era? using llms as a teachable agent for debugging. *25th International Conference on Artificial Intelligence in Education (AIED 2024)*, 2024.

[99] Yi-Hao Peng, JiWoong Jang, Jeffrey P Bigham, and Amy Pavel. Say it all: Feedback for improving non-visual presentation accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[100] Omar Shaikh, Valentino Chai, Michele J Gelfand, Diyi Yang, and Michael S Bernstein. Rehearsal: Simulating conflict to teach conflict resolution. In *ACM Conference on Human Factors in Computing Systems*, 2024.

[101] Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. How ai ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment. *arXiv:2401.13481*, 2024.

[102] Shubham Atreja, Libby Hemphill, and Paul Resnick. Remove, reduce, inform: What actions do people want social media platforms to take on potentially misleading content? *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–33, 2023.

[103] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara Jane Ericson, David Weintrop, and Tovi Grossman. How novices use llm-based code generators to solve cs1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, pages 1–12, 2023.

[104] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. Blaming humans and machines: What shapes people's reactions to algorithmic harm. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–26, 2023.

[105] Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1112–1123, 2023.

[106] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. Evaluating human-language model interaction. *Transactions on Machine Learning Research*, 2023.

[107] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. Harnessing the power of llms: Evaluating human-ai text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3321–3339, 2023.

[108] Minkyu Shin, Jin Kim, Bas van Opheusden, and Thomas L Griffiths. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12):e2214840120, 2023.

[109] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. Valuecompass: A framework of fundamental values for human-ai alignment. *arXiv preprint arXiv:2409.09586*, 2024.

[110] Hua Shen, Nicholas Clark, and Tanushree Mitra. Mind the value-action gap: Do llms act in alignment with their values? *arXiv preprint arXiv:2501.15463*, 2025.

[111] Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, 2021.

[112] Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I Hong, and Adam Perer. Zeno: An interactive framework for behavioral evaluation of machine learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023.

[113] Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. Humanoid agents: Platform for simulating human-like generative agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 167–176, 2023.

[114] Constantine Sedikides and Michael J Strube. The multiply motivated self. *Personality and Social Psychology Bulletin*, 21(12):1330–1335, 1995.

[115] Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, et al. Lemur: Harmonizing natural language and code for language agents. In *The Twelfth International Conference on Learning Representations*, 2024.

[116] Viktor E Frankl. Self-transcendence as a human phenomenon. *Journal of Humanistic Psychology*, 6(2):97–106, 1966.

[117] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–17, 2023.

[118] Winston Wu, Lu Wang, and Rada Mihalcea. Cross-cultural analysis of human values, morals, and biases in folk tales. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[119] Andy Hamilton. Conservatism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition, 2020.

[120] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. volume 35, pages 38176–38189, 2022.

[121] Gabriel Lima, Changyeon Kim, Seungho Ryu, Chihyung Jeon, and Meeyoung Cha. Collecting the public perception of ai and robot rights. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24, 2020.

[122] Zahra Ashktorab, Q Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. Human-ai collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–20, 2020.

[123] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv:2308.08155*, 2023.

[124] Nicholas Clark, Hua Shen, Bill Howe, and Tanushree Mitra. Epistemic alignment: A mediating framework for user-llm knowledge delivery. *arXiv preprint arXiv:2504.01205*, 2025.

[125] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K Kummerfeld, and Elena L Glassman. An ai-resilient text rendering technique for reading and skimming documents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

[126] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW1), 2023.

[127] Tianshi Li, Sauvik Das, Hao-Ping Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. Human-centered privacy research in the age of large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2024.

[128] Yike Shi, Qing Xiao, Hong Shen, Hua Shen, et al. The siren song of llms: How users perceive and respond to dark patterns in large language models. *arXiv preprint arXiv:2509.10830*, 2025.

[129] Xusen Cheng, Xiaoping Zhang, Jason Cohen, and Jian Mou. Human vs. ai: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. *Information Processing & Management*, 59(3):102940, 2022.

[130] Gabriel Lima, Nina Grgic-Hlaca, Jin Keun Jeong, and Meeyoung Cha. Who should pay when machines cause harm? laypeople's expectations of legal damages for machine-caused harm. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 236–246, New York, NY, USA, 2023. Association for Computing Machinery.

[131] Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, Manish Nagireddy, Prasanna Sattigeri, Ameet Talwalkar, and David Sontag. The realhumaneval: Evaluating large language models' abilities to support programmers. *arXiv:2404.02806*, 2024.

[132] Eric Zhou and Dokyun Lee. Generative artificial intelligence, human creativity, and art. *PNAS nexus*, 3(3):pgae052, 2024.

[133] Steven Johnson and Nikita Iziev. Ai is mastering language. should we trust what it says? *The New York Times*, 4:15, 2022.

[134] American Psychological Association et al. The truth about lie detectors (aka polygraph tests). *Recuperado de: https://www. apa. org/topics/cognitive-neuroscience/polygraph*, 2004.

[135] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, 372, 2021.

[136] Evropi Stefanidi, Marit Bentvelzen, Paweł W Woźniak, Thomas Kosch, Mikołaj P Woźniak, Thomas Mildner, Stefan Schneegass, Heiko Müller, and Jasmin Niess. Literature reviews in hci: A review of reviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2023.

[137] Songlin Xu and Xinyu Zhang. Augmenting human cognition with an ai-mediated intelligent visual feedback. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[138] Sumit Srivastava, Mariët Theune, and Alejandro Catala. The role of lexical alignment in human understanding of explanations by conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 423–435, 2023.

[139] David Piorkowski, Inge Vejsbjerg, Owen Cornec, Elizabeth M Daly, and Öznur Alkan. Aimee: An exploratory study of how rules support ai developers to explain and edit models. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–25, 2023.

[140] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. Selective explanations: Leveraging human input to align explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–35, 2023.

[141] Daniel Karl I Weidele, Shazia Afzal, Abel N Valente, Cole Makuch, Owen Cornec, Long Vu, Dharmashankar Subramanian, Werner Geyer, Rahul Nair, Inge Vejsbjerg, et al. Autodoviz: Human-centered automation for decision optimization. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 664–680, 2023.

[142] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen MacNeil. Memory sandbox: Transparent and interactive memory management for conversational agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–3, 2023.

[143] Amama Mahmood, Jeanie W Fung, Isabel Won, and Chien-Ming Huang. Owning mistakes sincerely: Strategies for mitigating ai errors. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2022.

[144] Lijie Guo, Elizabeth M Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. Building trust in interactive machine learning via user contributed interpretable rules. In *27th International Conference on Intelligent User Interfaces*, pages 537–548, 2022.

[145] Elliot Mitchell, Noemie Elhadad, and Lena Mamykina. Examining ai methods for micro-coaching dialogs. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2022.

[146] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020.

[147] Alison Smith-Renner, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. Digging into user control: perceptions of adherence and instability in transparent models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 519–530, 2020.

[148] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. Harnessing biomedical literature to calibrate clinicians' trust in ai decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023.

[149] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[150] Marc Pinski, Martin Adam, and Alexander Benlian. Ai knowledge: Improving ai delegation through human enablement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[151] Federico Cabitza, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi. Ai shall have no dominion: on how to measure technology dominance in ai-supported human decision-making. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–20, 2023.

[152] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[153] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. How stated accuracy of an ai system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–29, 2023.

[154] Lillio Mok, Sasha Nanda, and Ashton Anderson. People perceive algorithmic assessments as less fair and trustworthy than identical human assessments. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26, 2023.

[155] Min Hun Lee and Chong Jun Chew. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22, 2023.

[156] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. Convxai: Delivering heterogeneous ai explanations via conversations to support human-ai scientific writing. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 384–387, 2023.

[157] Taenyun Kim, Maria D Molina, Minjin Rheu, Emily S Zhan, and Wei Peng. One ai does not fit all: A cluster analysis of the laypeople's perception of ai roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[158] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. Knowing about knowing: An illusion of human competence can hinder appropriate reliance on ai systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[159] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[160] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2023.

[161] Jaemarie Solyst, Shixian Xie, Ellia Yang, Angela EB Stewart, Motahhare Eslami, Jessica Hammer, and Amy Ogan. "i would like to design": Black girls analyzing and ideating fair and accountable ai. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[162] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D Mekler. How does hci understand human agency and autonomy? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[163] Zahra Ashktorab, Benjamin Hoover, Mayank Agarwal, Casey Dugan, Werner Geyer, Hao Bang Yang, and Mikhail Yurochkin. Fairness evaluation in text classification: Machine learning practitioner perspectives of individual and group fairness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[164] Xinru Wang and Ming Yin. Watch out for updates: Understanding the effects of model explanation updates in ai-assisted decision making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[165] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–15, 2023.

[166] Jaemarie Solyst, Ellia Yang, Shixian Xie, Amy Ogan, Jessica Hammer, and Motahhare Eslami. The potential of diverse youth as stakeholders in identifying and mitigating algorithmic bias for a future of fairer ai. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–27, 2023.

[167] Peiling Jiang, Jude Rayan, Steven P Dow, and Haijun Xia. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–20, 2023.

[168] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-ai decision making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 379–396, 2023.

[169] Andrew M McNutt, Chenglong Wang, Robert A Deline, and Steven M Drucker. On the design of ai-powered code assistants for notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[170] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. Capable but amoral? comparing ai and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[171] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–28, 2022.

[172] Chun-Wei Chiang and Ming Yin. Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In *27th International Conference on Intelligent User Interfaces*, pages 148–161, 2022.

[173] Quan Ze Chen, Tobias Schnabel, Besmira Nushi, and Saleema Amershi. Hint: Integration testing for ai-based features with humans in the loop. In *27th International Conference on Intelligent User Interfaces*, pages 549–565, 2022.

[174] Krzysztof Z Gajos and Lena Mamykina. Do people engage cognitively with ai? impact of ai assistance on incidental learning. In *27th international conference on intelligent user interfaces*, pages 794–806, 2022.

[175] Min Kyung Lee and Katherine Rich. Who is included in human perceptions of ai?: Trust and perceived fairness around healthcare ai and cultural mistrust. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[176] Mengqi Liao and S Shyam Sundar. How should ai systems talk to users when collecting their personal information? effects of role framing and self-referencing on human-ai interaction. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[177] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. Human perceptions on moral responsibility of ai: A case study in ai-assisted bail decision-making. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–17, 2021.

[178] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[179] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[180] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[181] Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13, 2021.

[182] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

[183] Francesco Sovrano and Fabio Vitali. From philosophy to interfaces: an explanatory method and a tool inspired by achinstein's theory of explanation. In *26th International Conference on Intelligent User Interfaces*, pages 81–91, 2021.

[184] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th international conference on intelligent user interfaces*, pages 109–119, 2021.

[185] Malin Eiband, Daniel Buschek, and Heinrich Hussmann. How to support users in understanding intelligent systems? structuring the discussion. In *26th International Conference on Intelligent User Interfaces*, pages 120–132, 2021.

[186] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. I think i get your point, ai! the illusion of explanatory depth in explainable ai. In *26th international conference on intelligent user interfaces*, pages 307–317, 2021.

[187] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*, pages 318–328, 2021.

[188] Matthew K Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. Planning for natural language failures with the ai playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2021.

[189] Zohar Gilad, Ofra Amir, and Liat Levontin. The effects of warmth and competence perceptions on users' choice of an ai system. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13, 2021.

[190] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[191] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–39, 2021.

[192] Justin D Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. Perfection not required? human-ai partnerships in code translation. In *26th International Conference on Intelligent User Interfaces*, pages 402–412, 2021.

[193] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[194] Vivian Lai, Han Liu, and Chenhao Tan. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[195] Lanthao Benedikt, Chaitanya Joshi, Louisa Nolan, Ruben Henstra-Hill, Luke Shaw, and Sharon Hook. Human-in-the-loop ai in government: A case study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 488–497, 2020.

[196] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[197] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.

[198] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019.

[199] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–24, 2019.

[200] Shivani Kapania, Alex S Taylor, and Ding Wang. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[201] Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. Out of context: Investigating the bias and fairness concerns of "artificial intelligence as a service". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[202] Rama Adithya Varanasi and Nitesh Goyal. "it is currently hodgepodge": Examining ai/ml practitioners' challenges during co-production of responsible ai values. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[203] Richmond Y Wong, Michael A Madaio, and Nick Merrill. Seeing like a toolkit: How toolkits envision the work of ai ethics. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–27, 2023.

[204] Shanley Corvite, Kat Roemmich, Tillie Ilana Rosenberg, and Nazanin Andalibi. Data subjects' perspectives on emotion artificial intelligence use in the workplace: A relational ethics lens. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

[205] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. Improving human-ai partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[206] Nithya Sambasivan and Rajesh Veeraraghavan. The deskilling of domain expertise in ai development. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.

[207] Karen L Boyd. Datasheets for datasets help ml engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

[208] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

[209] Azra Ismail and Neha Kumar. Ai in global health: the view from the front lines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2021.

[210] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[211] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. Competent but rigid: Identifying the gap in empowering ai to participate equally in group decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[212] Imani Munyaka, Zahra Ashktorab, Casey Dugan, James Johnson, and Qian Pan. Decision making strategies and team efficacy in human-ai teams. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–24, 2023.

[213] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. Deliberating with ai: Improving decision-making for the future through participatory ai design and stakeholder deliberation. *CSCW1*, 7(CSCW1):1–32, 2023.

[214] Weihao Chen, Chun Yu, Huadong Wang, Zheng Wang, Lichen Yang, Yukun Wang, Weinan Shi, and Yuanchun Shi. From gap to synergy: Enhancing contextual understanding through human-machine collaboration in personalized systems. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15, 2023.

[215] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. " help me help the ai": Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[216] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. Dapie: Interactive step-by-step explanatory dialogues to answer children's why and how questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2023.

[217] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. Readingquizmaker: A human-nlp collaborative system that supports instructors to design high-quality reading quiz questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[218] Jen Rogers and Anamaria Crisan. Tracing and visualizing human-ml/ai collaborative processes through artifacts of data work. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2023.

[219] Yoyo Tsung-Yu Hou, Wen-Ying Lee, and Malte Jung. "should i follow the human, or follow the robot?"—robots in power can have more influence than humans on decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2023.

[220] Ángel Alexander Cabrera, Adam Perer, and Jason I Hong. Improving human-ai collaboration with descriptions of ai behavior. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–21, 2023.

[221] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[222] Devleena Das, Been Kim, and Sonia Chernova. Subgoal-based explanations for unreliable intelligent decision support systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 240–250, 2023.

[223] Sherry Wu, Hua Shen, Daniel S Weld, Jeffrey Heer, and Marco Tulio Ribeiro. Scattershot: Interactive in-context example curation for text transformation. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 353–367, 2023.

[224] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. Human-ai collaboration: The effect of ai delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 453–463, 2023.

[225] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. Supporting high-uncertainty decisions through ai and logic-style explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 251–263, 2023.

[226] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–30, 2023.

[227] Fengjie Wang, Xuye Liu, Oujing Liu, Ali Neshati, Tengfei Ma, Min Zhu, and Jian Zhao. Slide4n: Creating presentation slides from computational notebooks with human-ai collaboration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[228] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D Gordon. "what it wants me to say": Bridging the abstraction gap between end-user programmers and code-generating large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–31, 2023.

[229] Jie Gao, Yuchen Guo, Toby Jia-Jun Li, and Simon Tangi Perrault. Collabcoder: a gpt-powered workflow for collaborative qualitative analysis. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 354–357, 2023.

[230] Gabriele Cimolino and TC Nicholas Graham. Two heads are better than one: A dimension space for unifying human and artificial intelligence in shared control. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2022.

[231] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2022.

[232] Yunlong Wang, Priyadarshini Venkatesh, and Brian Y Lim. Interpretable directed diversity: Leveraging model explanations for iterative crowd ideation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.

[233] Harini Suresh, Kathleen M Lewis, John Guttag, and Arvind Satyanarayan. Intuitively assessing ml model reliability through example-based explanations and editing model inputs. In *27th International Conference on Intelligent User Interfaces*, pages 767–781, 2022.

[234] Matt-Heun Hong, Lauren A Marsh, Jessica L Feuston, Janet Ruppert, Jed R Brubaker, and Danielle Albers Szafir. Scholastic: Graphical human-ai collaboration for inductive and interpretive text analysis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–12, 2022.

[235] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–13, 2022.

[236] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[237] Kathryn Cunningham, Barbara J Ericson, Rahul Agrawal Bejarano, and Mark Guzdial. Avoiding the turing tarpit: Learning conversational programming by starting from code's purpose. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[238] Tim Rietz and Alexander Maedche. Cody: An ai-based system to semi-automate coding for qualitative research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[239] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021.

[240] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. Designing ai for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the 2021 chi conference on human factors in computing systems*, pages 1–14, 2021.

[241] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. Supporting serendipity: Opportunities and challenges for human-ai collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

[242] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–45, 2021.

[243] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. " an ideal human" expectations of ai teammates in human-ai teaming. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–25, 2021.

[244] Jules Françoise, Baptiste Caramiaux, and Téo Sanchez. Marcelle: composing interactive machine learning workflows and interfaces. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 39–53, 2021.

[245] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–22, 2021.

[246] Austin P Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng Chau, and Diyi Yang. Recast: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26, 2021.

[247] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, et al. Increasing the speed and accuracy of data labeling through an ai assisted interface. In *26th International Conference on Intelligent User Interfaces*, pages 392–401, 2021.

[248] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020.

[249] Toby Jia-Jun Li. Multi-modal interactive task learning from demonstrations and natural language instructions. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 162–168, 2020.

[250] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[251] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J Smith, Kalyan Veeramachaneni, and Huamin Qu. Atmseer: Increasing transparency and controllability in automated machine learning. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[252] Shi Feng and Jordan Boyd-Graber. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 229–239, 2019.

[253] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[254] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. Talebrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[255] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. Leveraging large language models to power chatbots for collecting user self-reported data. *arXiv:2301.05843*, 2023.

[256] Matthew Jörke, Yasaman S Sefidgar, Talie Massachi, Jina Suh, and Gonzalo Ramos. Pearl: A technology probe for machine-assisted reflection on personal data. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 902–918, 2023.

[257] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[258] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. volume 36, 2023.

[259] Yu Chen, Yihan Du, Pihe Hu, Siwei Wang, Desheng Wu, and Longbo Huang. Provably efficient iterated cvar reinforcement learning with function approximation and human feedback. In *The Twelfth International Conference on Learning Representations*, 2023.

[260] Chengzhi Cao, Yinghao Fu, Sheng Xu, Ruimao Zhang, and Shuang Li. Enhancing human-AI collaboration through logic-guided reasoning. In *The Twelfth International Conference on Learning Representations*, 2024.

[261] Han Zhang, Yu Lei, Lin Gui, Min Yang, Yulan He, Hui Wang, and Ruifeng Xu. CPPO: Continual learning for reinforcement learning with human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

[262] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[263] Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, and Marc Dymetman. Compositional preference models for aligning lms. In *The Twelfth International Conference on Learning Representations*, 2023.

[264] Marcel Binz and Eric Schulz. Turning large language models into cognitive models. In *The Twelfth International Conference on Learning Representations*, 2023.

[265] Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, and Fei Liu. Decipherpref: Analyzing influential factors in human preference judgments via gpt-4. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[266] Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. Measuring inductive biases of in-context learning with underspecified demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

[267] Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. Aligning language models with human preferences via a bayesian approach. volume 36, 2023.

[268] Lin Guan, Karthik Valmeekam, and Subbarao Kambhampati. Relative behavioral attributes: Filling the gap between symbolic goal specification and reward learning from human preferences. In *The Eleventh International Conference on Learning Representations*, 2022.

[269] Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy Rogers. Conceptual structure coheres in human cognition but not in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 722–738, 2023.

[270] Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, 2023.

[271] Lingjun Zhao, Khanh Nguyen, and Hal Daumé III. Define, evaluate, and improve task-oriented cognitive capabilities for instruction generation models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3688–3706, 2023.

[272] Sullam Jeoung, Yubin Ge, and Jana Diesner. Stereomap: Quantifying the awareness of human-like stereotypes in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12236–12256, 2023.

[273] Nick Mckenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, 2023.

[274] Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, 2023.

[275] Enyu Zhou, Rui Zheng, Zhiheng Xi, Songyang Gao, Xiaoran Fan, Zichu Fei, Jingting Ye, Tao Gui, Qi Zhang, and Xuan-Jing Huang. Realbehavior: A framework for faithfully characterizing foundation models' human-like behavior mechanisms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10262–10274, 2023.

[276] Minyoung Hwang, Gunmin Lee, Hogun Kee, Chan Woo Kim, Kyungjae Lee, and Songhwai Oh. Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, 36, 2023.

[277] Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11275–11288, 2023.

[278] Junbing Yan, Chengyu Wang, Taolin Zhang, Xiaofeng He, Jun Huang, and Wei Zhang. From complex to simple: Unraveling the cognitive tree for reasoning with small language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12413–12425, 2023.

[279] Hongli Zhan, Desmond Ong, and Junyi Jessy Li. Evaluating subjective cognitive appraisals of emotions from large language models. pages 14418–14446, 2023.

[280] Ruixi Lin and Hwee Tou Ng. Mind the biases: Quantifying cognitive biases in language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5269–5281, 2023.

[281] Zhenghao Mark Peng, Wenjie Mo, Chenda Duan, Quanyi Li, and Bolei Zhou. Learning from active human involvement through proxy value propagation. *Advances in neural information processing systems*, 36, 2023.

[282] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*, 2024.

[283] Bong Gyun Kang, HyunGi Kim, Dahuin Jung, and Sungroh Yoon. Clear: Continual learning on algorithmic reasoning for human-like intelligence. volume 36, 2023.

[284] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. volume 36, 2023.

[285] Shicheng Liu and Minghui Zhu. Learning multi-agent behaviors from distributed and streaming demonstrations. volume 36, 2023.

[286] Xin-Qiang Cai, Yu-Jie Zhang, Chao-Kai Chiang, and Masashi Sugiyama. Imitation learning from vague feedback. volume 36, 2023.

[287] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998, 2024.

[288] Chao Yu, Jiaxuan Gao, Weilin Liu, Botian Xu, Hao Tang, Jiaqi Yang, Yu Wang, and Yi Wu. Learning zero-shot cooperation with humans, assuming humans are biased. In *The Eleventh International Conference on Learning Representations*, 2023.

[289] Kawin Ethayarajh and Dan Jurafsky. The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070, 2022.

[290] Ge Gao, Eunsol Choi, and Yoav Artzi. Simulating bandit learning from user feedback for extractive question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5167–5179, 2022.

[291] Amir Feder, Guy Horowitz, Yoav Wald, Roi Reichart, and Nir Rosenfeld. In the eye of the beholder: Robust prediction with causal user modeling. volume 35, pages 14419–14433, 2022.

[292] Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. volume 35, pages 11785–11799, 2022.

[293] Siddharth Reddy, Sergey Levine, and Anca Dragan. First contact: Unsupervised human-machine co-adaptation via mutual information maximization. volume 35, pages 31542–31556, 2022.

[294] Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. Transferable dialogue systems and user simulators. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 152–166, 2021.

[295] Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. Case study: Deontological ethics in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3784–3798, 2021.

[296] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. The utility of explainable ai in ad hoc human-machine teaming. volume 34, pages 610–623, 2021.

[297] Yuandong Tian, Qucheng Gong, and Yu Jiang. Joint policy search for multi-agent collaboration with imperfect information. volume 33, pages 19931–19942, 2020.

[298] Hao Yan, Saurabh Srivastava, Yintao Tai, Sida I Wang, Wen-tau Yih, and Ziyu Yao. Learning to simulate natural language feedback for interactive semantic parsing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3149–3170, 2023.

[299] Robin Chan, Afra Amini, and Mennatallah El-Assady. Which spurious correlations impact reasoning in nli models? a visual interactive diagnosis through data-constrained counterfactuals. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 463–470, 2023.

[300] Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, et al. Xmd: An end-to-end framework for interactive explanation-based debugging of nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 264–273, 2023.

[301] Huao Li, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, 2023.

[302] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. Harnessing the power of llms: Evaluating human-ai text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3321–3339, 2023.

[303] Timo Schick, A Yu Jane, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. Peer: A collaborative language model. In *The Eleventh International Conference on Learning Representations*, 2022.

[304] Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel Tetreault, and Alejandro Jaimes-Larrarte. Mapping the design space of human-ai interaction in text summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 431–455, 2022.

[305] Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. A human-machine collaborative framework for evaluating malevolence in dialogues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5612–5623, 2021.

[306] Jesse Vig, Wojciech Kryściński, Karan Goel, and Nazneen Rajani. Summvis: Interactive visual analysis of models, data, and evaluation for text summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 150–158, 2021.

[307] Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. Allennlp interpret: A framework for explaining predictions of nlp models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 2019.

[308] Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. Plan, write, and revise: an interactive system for open-domain story generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 89–97, 2019.

[309] Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations*, 2024.

[310] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. Urial: Aligning untuned llms with just the'write'amount of in-context learning. In *The Twelfth International Conference on Learning Representations*, 2023.

[311] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox. In *The Twelfth International Conference on Learning Representations*, 2023.

[312] Aida Ramezani and Yang Xu. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, 2023.

[313] CH-Wang Sky, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[314] Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. This prompt is measuring< mask>: evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, 2023.

[315] Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, 2023.

[316] Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, 2023.

[317] Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*, 2023.

[318] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, 2021.

[319] Zhaowei Zhu, Jialu Wang, Hao Cheng, and Yang Liu. Unmasking and improving data credibility: A study with datasets for training harmless language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[320] Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. Motif: Intrinsic motivation from artificial intelligence feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

[321] 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, 2023.

[322] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, 2023.

[323] Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7103–7128, 2023.

[324] Bingsheng Yao, Prithviraj Sen, Lucian Popa, James Hendler, and Dakuo Wang. Are human explanations always helpful? towards objective evaluation of human natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14698–14713, 2023.

[325] Bum Chul Kwon and Nandana Mihindukulasooriya. Finspector: A human-centered visual inspection tool for exploring and comparing biases among foundation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 42–50, 2023.

[326] Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. A diachronic perspective on user trust in ai under uncertainty. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5567–5580, 2023.

[327] Noah Lee, Na Min An, and James Thorne. Can large language models capture dissenting human voices? In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[328] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, 2023.

[329] Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4694–4702, 2023.

[330] Eoin M Kenny, Mycal Tucker, and Julie Shah. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*, 2022.

[331] Hua Shen, Tongshuang Wu, Wenbo Guo, and Ting-Hao Huang. Are shortest rationales the best explanations for human understanding? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 10–19, 2022.

[332] Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. Genie: Toward reproducible and standardized human evaluation for text generation. pages 11444–11458, 2022.

[333] Henrik Voigt, Özge Alaçam, Monique Meuschke, Kai Lawonn, and Sina Zarrieß. The why and the how: A survey on natural language interaction in visualization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–374, 2022.

[334] Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, 2020.

[335] Yiming Gao, Feiyu Liu, Liang Wang, Dehua Zheng, Zhenjie Lian, Weixuan Wang, Wenjin Yang, Siqin Li, Xianliang Wang, Wenhui Chen, et al. Enhancing human experience in human-agent collaboration: A human-centered modeling approach based on positive human gain. 2024.

[336] Geyang Guo, Ranchi Zhao, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. Beyond imitation: Leveraging fine-grained quality signals for alignment. In *The Twelfth International Conference on Learning Representations*, 2024.

[337] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*, 2024.

[338] Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and Hua Wu. Tool-augmented reward modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

[339] Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*, 2024.

[340] Dominic Petrak, Nafise Sadat Moosavi, Ye Tian, Nikolai Rozanov, and Iryna Gurevych. Learning from free-text human feedback–collect new datasets or extend existing ones? In *The Twelfth International Conference on Learning Representations*, 2024.

[341] Yuan Tian, Zheng Zhang, Zheng Ning, Toby Li, Jonathan K Kummerfeld, and Tianyi Zhang. Interactive text-to-sql generation via editable step-by-step explanations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16149–16166, 2023.

[342] Chunxu Yang, Chien-Sheng Wu, Lidiya Murakhovs'ka, Philippe Laban, and Xiang Chen. Intelmo: Enhancing models' adoption of interactive interfaces. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 161–166, 2023.

[343] Aviv Slobodkin, Niv Nachum, Shmuel Amar, Ori Shapira, and Ido Dagan. Summhelper: Collaborative human-computer summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 554–565, 2023.

[344] Zonghai Yao, Benjamin Schloss, and Sai Selvaraj. Improving summarization with human edits. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2604–2620, 2023.

[345] John Chung, Ece Kamar, and Saleema Amershi. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, 2023.

[346] Minsik Oh, Joosung Lee, Jiwei Li, and Guoyin Wang. Pk-icr: Persona-knowledge interactive multi-context retrieval for grounded dialogue. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16383–16395, 2023.

[347] Chenliang Li, He Chen, Ming Yan, Weizhou Shen, Haiyang Xu, Zhikai Wu, Zhicheng Zhang, Wenmeng Zhou, Yingda Chen, Chen Cheng, et al. Modelscope-agent: Building your customizable agent system with open-source large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 566–578, 2023.

[348] Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Röttger, and Scott Hale. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, 2023.

[349] Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. Fairprism: evaluating fairness-related harms in text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6231–6251, 2023.

[350] Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of*

the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6252–6272, 2023.

[351] Afra Feyza Akyürek, Ekin Akyürek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7716–7733, 2023.

[352] Anthony Sicilia and Malihe Alikhani. Learning to generate equitable text in dialogue from biased training data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2898–2917, 2023.

[353] Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, 2023.

[354] Yiming Wang, Zhuosheng Zhang, and Rui Wang. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, 2023.

[355] Zahra Fatemi, Chen Xing, Wenhao Liu, and Caimming Xiong. Improving gender fairness of pre-trained language models without catastrophic forgetting. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[356] Nikil Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. The tail wagging the dog: Dataset construction biases of social bias benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1373–1386, 2023.

[357] Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, et al. Democratizing reasoning ability: Tailored learning from large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1948–1966, 2023.

[358] Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13557–13572, 2023.

[359] Ge Gao, Hung-Ting Chen, Yoav Artzi, and Eunsol Choi. Continually improving extractive qa via human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 406–423, 2023.

[360] David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Michael Smith. Robbie: Robust bias evaluation of large generative language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[361] Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, 2023.

[362] Xinpeng Wang and Barbara Plank. Actor: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, 2023.

[363] Bodhisattwa Majumder, Zexue He, and Julian McAuley. Interfair: Debiasing with natural language feedback for fair interpretable predictions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9466–9471, 2023.

[364] Nan Wang, Qifan Wang, Yi-Chia Wang, Maziar Sanjabi, Jingzhou Liu, Hamed Firooz, Hongning Wang, and Shaoliang Nie. Coffee: Counterfactual fairness for personalized text generation in explainable recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13258–13275, 2023.

[365] Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and Jonathan Gratch. Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13078–13092, 2023.

[366] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, 2023.

[367] Jiaao Chen, Mohan Dodda, and Diyi Yang. Human-in-the-loop abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9176–9190, 2023.

[368] Xiang Fan, Yiwei Lyu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Nano: Nested human-in-the-loop reward learning for few-shot language model control. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11970–11992, 2023.

[369] Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback, 2023.

[370] Xingjin Wang, Linjing Li, and Daniel Zeng. Ldm2: A large decision model imitating human cognition with dynamic memory enhancement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4660–4681, 2023.

[371] Chenglei Si, Weijia Shi, Chen Zhao, Luke Zettlemoyer, and Jordan Boyd-Graber. Getting more out of mixture of language model reasoning experts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8234–8249, 2023.

[372] Xi Wang, Hossein Rahmani, Jiqun Liu, and Emine Yilmaz. Improving conversational recommendation systems via bias analysis and language-model-enhanced data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3609–3622, 2023.

[373] Yushan Qian, Weinan Zhang, and Ting Liu. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, 2023.

[374] Siqi Ouyang and Lei Li. Autoplan: Automatic planning of interactive decision-making tasks with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3114–3128, 2023.

[375] Tyler Loakman, Aaron Maladry, and Chenghua Lin. The iron (ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, 2023.

[376] Shengran Hu and Jeff Clune. Thought cloning: Learning to think while acting by imitating human thinking. volume 36, 2023.

[377] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36, 2023.

[378] Joey Hong, Sergey Levine, and Anca Dragan. Learning to influence human behavior with offline reinforcement learning. volume 36, 2023.

[379] Ilia Sucholutsky and Tom Griffiths. Alignment with human representations supports robust few-shot learning. volume 36, 2023.

[380] Jiaxin Zhang, Zhuohang Li, Kamalika Das, and Sricharan Kumar. Interactive multi-fidelity learning for cost-effective adaptation of language model with sparse human supervision. volume 36, 2023.

[381] Fereshte Khani and Marco Tulio Ribeiro. Collaborative alignment of nlp models. volume 36, 2023.

[382] Bidipta Sarkar, Andy Shih, and Dorsa Sadigh. Diverse conventions for human-ai collaboration. volume 36, 2023.

[383] Nina Corvelo Benz and Manuel Rodriguez. Human-aligned calibration for ai-assisted decision making. volume 36, 2023.

[384] Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, et al. Beyond labels: Empowering human annotators with natural language explanations through a novel active-learning architecture. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11629–11643, 2023.

[385] Xue Yan, Jiaxian Guo, Xingzhou Lou, Jun Wang, Haifeng Zhang, and Yali Du. An efficient end-to-end training approach for zero-shot human-ai coordination. volume 36, 2023.

[386] Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. volume 36, 2023.

[387] Marcel Torne Villasevil, Balsells I Pamies, Zihan Wang, Samedh Desai, Tao Chen, Pulkit Agrawal, Abhishek Gupta, et al. Breadcrumbs to the goal: Supervised goal selection from human-in-the-loop feedback. volume 36, 2023.

[388] 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. volume 36, 2023.

[389] Prakhar Gupta, Yang Liu, Di Jin, Behnam Hedayatnia, Spandana Gella, Sijia Liu, Patrick L Lange, Julia Hirschberg, and Dilek Hakkani-Tur. Dialguide: Aligning dialogue model behavior with developer guidelines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

[390] Hussein Mozannar, Jimin Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Effective human-ai teams via learned natural language rules and onboarding. volume 36, 2023.

[391] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[392] Xiao Hu, Jianxiong Li, Xianyuan Zhan, Qing-Shan Jia, and Ya-Qin Zhang. Query-policy misalignment in preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[393] Han Liu, Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan. Learning human-compatible representations for case-based decision support. In *The Eleventh International Conference on Learning Representations*, 2023.

[394] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. In *The Twelfth International Conference on Learning Representations*, 2024.

[395] Kefan Dong and Tengyu Ma. Asymptotic instance-optimal algorithms for interactive decision making. In *The Eleventh International Conference on Learning Representations*, 2023.

[396] Florian E Dorner, Momchil Peychev, Nikola Konstantinov, Naman Goel, Elliott Ash, and Martin Vechev. Human-guided fair classification for natural language processing. In *The Eleventh International Conference on Learning Representations*, 2023.

[397] Jinghui Lu, Linyi Yang, Brian Namee, and Yue Zhang. A rationale-centric framework for human-in-the-loop machine learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6986–6996, 2022.

[398] Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, 2022.

[399] Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–21, 2022.

[400] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-ai collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, 2022.

[401] Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Chi Kit Cheung, and Siva Reddy. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 926–937, 2022.

[402] Tushar Khot, Kyle Richardson, Daniel Khashabi, and Ashish Sabharwal. Hey ai, can you solve complex tasks by talking to agents? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1808–1823, 2022.

[403] Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9465–9480, 2022.

[404] Maartje Ter Hoeve, Julia Kiseleva, and Maarten Rijke. What makes a good and useful summary? incorporating users in automatic summarization research. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 46–75, 2022.

[405] Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, 2022.

[406] Duy-Hung Nguyen, Nguyen Viet Dung Nghiem, Bao-Sinh Nguyen, Dung Tien Tien Le, Shahab Sabahi, Minh-Tien Nguyen, and Hung Le. Make the most of prior data: A solution for interactive text summarization with preference feedback. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1919–1930, 2022.

[407] Xi Ye and Greg Durrett. Can explanations be useful for calibrating black box models? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6199–6212, 2022.

[408] Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. Context-aware information-theoretic causal de-biasing for interactive sequence labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3436–3448, 2022.

[409] Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, 2022.

[410] Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. Uncalibrated models can improve human-ai collaboration. volume 35, pages 4004–4016, 2022.

[411] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. Pre-trained language models for interactive decision-making. volume 35, pages 31199–31212, 2022.

[412] Lin Guan, Mudit Verma, Suna Sihang Guo, Ruohan Zhang, and Subbarao Kambhampati. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. volume 34, pages 21885–21897, 2021.

[413] Akshatha Arodi and Jackie Chi Kit Cheung. Textual time travel: A temporally informed approach to theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4162–4172, 2021.

[414] Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. Interactive weak supervision: Learning useful heuristics for data labeling. In *International Conference on Learning Representations*, 2020.

[415] Yi Tay, Donovan Ong, Jie Fu, Alvin Chan, Nancy Chen, Anh Tuan Luu, and Christopher Pal. Would you rather? a new benchmark for learning machine alignment with cultural values and social preferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5369–5373, 2020.

[416] Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. Find: Human-in-the-loop debugging deep text classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 332–348, 2020.

[417] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and William B Dolan. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, 2020.

[418] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. volume 33, pages 3008–3021, 2020.

[419] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. volume 32, 2019.

[420] Prithviraj Sen, Yunyao Li, Eser Kandogan, Yiwei Yang, and Walter Lasecki. Heidl: Learning linguistic expressions with deep learning and human-in-the-loop. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–140, 2019.

[421] Rui Zheng, Wei Shen, Yuan Hua, Wenbin Lai, Shihan Dou, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Haoran Huang, Tao Gui, et al. Improving generalization of alignment with human preferences through group invariant learning. In *The Twelfth International Conference on Learning Representations*, 2023.

[422] EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, 2023.

[423] Vinod Muthusamy, Yara Rizk, Kiran Kate, Praveen Venkateswaran, Vatche Isahagian, Ashu Gulati, and Parijat Dube. Towards large language model-based personal agents in the enterprise: Current trends and open problems. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[424] Jungwoo Lim, Myunghoon Kang, Jinsung Kim, Jeongwook Kim, Yuna Hur, and Heui-Seok Lim. Beyond candidates: Adaptive dialogue agent utilizing persona and knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7950–7963, 2023.

[425] WANG Hongru, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. Large language models as source planner for personalized knowledge-grounded dialogues. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[426] Charles Welch, Chenxi Gu, Jonathan K Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, 2022.

[427] Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, 2023.

[428] Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, 2023.

[429] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. From instructions to intrinsic human values–a survey of alignment goals for big models. *arXiv:2308.12014*, 2023.

[430] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *arXiv:2307.12966*, 2023.

[431] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv:2209.07858*, 2022.

[432] Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. Investigating cultural alignment of large language models. *arXiv:2402.13231*, 2024.

[433] 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:2204.05862*, 2022.

[434] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv:1909.08593*, 2019.

[435] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

[436] Angelica Chen, Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Samuel R. Bowman, Kyunghyun Cho, and Ethan Perez. Learning from natural language feedback. *Transactions on Machine Learning Research*, 2024.

[437] Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback at scale. *arXiv:2303.16755*, 2023.

[438] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv:2112.00861*, 2021.

[439] Katie Shilton. Values levers: Building ethics into design. *Science, Technology, & Human Values*, 38(3):374–397, 2013.

[440] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *arXiv:2305.11391*, 2023.

[441] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

[442] Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, et al. Interactive natural language processing. *arXiv:2305.13246*, 2023.

[443] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. Do users write more insecure code with ai assistants? pages 2785–2799, 2023.

[444] Hua Shen, Yuguang Yang, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, and Andreas Stolcke. Improving fairness in speaker verification via group-adapted fusion network. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7077–7081. IEEE, 2022.

[445] Bum Chul Kwon and Nandana Mihindukulasooriya. Finspector: A human-centered visual inspection tool for exploring and comparing biases among foundation models. In Danushka Bollegala, Ruihong Huang, and Alan Ritter, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 42–50, Toronto, Canada, July 2023. Association for Computational Linguistics.

[446] S Shyam Sundar and Eun-Ju Lee. Rethinking communication in the era of artificial intelligence. *Human Communication Research*, 48(3), 2022.

[447] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, et al. How experienced designers of enterprise applications engage ai as a design material. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

[448] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review. *arXiv:2404.14082*, 2024.

[449] Richard Brath. Surveying wonderland for many more literature visualization techniques. *arXiv:2110.08584*, 2021.

[450] Norbert Wiener. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131(3410):1355–1358, 1960.

[451] Kees Stuurman and Hugo Wijnands. Software law: intelligent agents: a curse or a blessing? a survey of the legal aspects of the application of intelligent software systems. *Computer Law & Security Review*, 17(2):92–100, 2001.

[452] Michael Wooldridge. Intelligent agents. *Multiagent systems: A modern approach to distributed artificial intelligence*, 1:27–73, 1999.

[453] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

[454] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, page eadn0117, 2024.

[455] Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*, pages 26837–26867. PMLR, 2023.

[456] Allan Dafoe and Stuart Russell. Yes, we are worried about the existential risk of artificial intelligence. *MIT Technology Review*, 2016.

[457] The ACM Director of Publications. Acm policy on authorship, May 2024.