Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions

Hua Shen[1★], Tiffany Knearem[2♥], Reshmi Ghosh[2♦], Kenan Alkiek[3★], Kundan Krishna[3♠], Yachuan Liu[3★], Ziqiao Ma[3★], Savvas Petridis[3♥], Yi-Hao Peng[3♠], Li Qiwei[3★], Sushrita Rakshit[3★], Chenglei Si[3♣], Yutong Xie[3★], Jeffrey P. Bigham[4♠], Frank Bentley[4♥], Joyce Chai[4★], Zachary Lipton[4♠], Qiaozhu Mei[4★], Rada Mihalcea[4★], Michael Terry[4♥], Diyi Yang[4♣], Meredith Ringel Morris[5♥], Paul Resnick[5★], David Jurgens[5★] [*],

★University of Michigan, ♥Google, ♦Microsoft, ♠Carnegie Mellon University, ♣Stanford University, ♥Google Research, ♥Google DeepMind, USA

Recent advancements in general-purpose AI have highlighted the importance of guiding AI systems towards the intended goals, ethical principles, and values of individuals and groups, a concept broadly recognized as *alignment*. However, the lack of clarified definitions and scopes of *human-AI alignment* poses a significant obstacle, hampering collaborative efforts across research domains to achieve this alignment. In particular, ML- and philosophy-oriented alignment research often views AI alignment as a static, unidirectional process (*i.e.,* aiming to ensure that AI systems' objectives match humans) rather than an ongoing, mutual alignment problem [429]. This perspective largely neglects the *long-term interaction* and *dynamic changes* of alignment. To understand these gaps, we introduce a systematic review of over 400 papers published between 2019 and January 2024, spanning multiple domains such as Human-Computer Interaction (HCI), Natural Language Processing (NLP), Machine Learning (ML), and others. We characterize, define and scope human-AI alignment. From this, we present a conceptual framework of "Bidirectional Human-AI Alignment" to organize the literature from a human-centered perspective. This framework encompasses both 1) conventional studies of *aligning AI to humans* that ensures AI produces the intended outcomes determined by humans, and 2) a proposed concept of *aligning humans to AI*, which aims to help individuals and society adjust to AI advancements both cognitively and behaviorally. Additionally, we articulate the key findings derived from literature analysis, including discussions about human values, interaction techniques, and evaluations. To pave the way for future studies, we envision three key challenges for future directions and propose examples of potential future solutions.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; **Collaborative and social computing systems and tools**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Human-AI Alignment, Human-AI Interaction, Human-Centered AI Explanation and Evaluation, Personalized AI Systems

ACM Reference Format:
Hua Shen[1★], Tiffany Knearem[2♥], Reshmi Ghosh[2♦], Kenan Alkiek[3★], Kundan Krishna[3♠], Yachuan Liu[3★], Ziqiao Ma[3★], Savvas Petridis[3♥], Yi-Hao Peng[3♠], Li Qiwei[3★], Sushrita Rakshit[3★], Chenglei Si[3♣], Yutong Xie[3★], Jeffrey P. Bigham[4♠], Frank Bentley[4♥], Joyce Chai[4★], Zachary Lipton[4♠], Qiaozhu Mei[4★], Rada Mihalcea[4★], Michael Terry[4♥], Diyi Yang[4♣], Meredith Ringel Morris[5♥], Paul Resnick[5★], David

[*]Please see Appendix A for the full author list with their roles, affiliations, and contributions. We denote each author's role with the following superscripts: 1 for project leads, 2 for team leads, 3 for team members (alphabetical, equal contributions), 4 for advisors (alphabetical, equal contributions) and 5 for project leading advisors. Corresponding author: Hua Shen at University of Michigan (huashen@umich.edu).

Manuscript submitted to ACM

Jurgens[5★] . 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. In *Proceedings of Manuscript*. ACM, New York, NY, USA, 56 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Artificial Intelligence (AI) has advanced significantly, especially with the advent of general-purpose generative AI, demonstrating unprecedented capabilities in solving a wide range of complicated and challenging problems, such as reasoning, generation, language understanding, and more [61, 282, 311]. However, as AI becomes increasingly powerful and deeply integrated into our lives, it also presents potential risks to people and to society broadly [271]. For example, text-to-image generative models were found to amplify stereotypes about race and gender [1], and biased algorithms in hiring processes were found to perpetuate discrimination [290]. These risks highlight foundational questions regarding how and which values are embedded to create AI models that drive decisions within real-world contexts.

These questions are part of a growing trend toward an interdisciplinary discussion on *AI alignment*, which, for now, is defined in Terry et al. [378] as "*considering the overall problem of how to ensure an AI produces the intended outcomes (as determined by its creator and/or user), without additional undesirable side effects (*e.g., *by not performing operations that could negatively affect individuals, groups, or society at large).*" Significant efforts have been made in aligning AI to humans to ensure that the AI systems' objectives match those of humans [122, 184, 195, 276]. For instance, Ouyang et al. [282] present InstructGPT, which aligns a pre-trained Large Language Model (LLM) to follow human instructions and feedback based on Reinforcement Learning from Human Feedback (RLHF), and Santurkar et al. [326] investigated the opinions LLMs hold, by measuring the agreement between LLMs and certain particular groups of people on a number of topics, via public opinion polls.

Despite these numerous investigations into *human-AI alignment*, its definition and scope remain ambiguous and inconsistent across literature, for example, regarding *whom to align with* and *the goal of alignment*. Additionally, prevailing research often views AI alignment as a static, unidirectional process (*i.e.,* aiming to ensure that AI systems' objectives match those of humans) rather than an ongoing and mutual alignment problem [429]. This unidirectional view largely understates the **long-term interaction** and **dynamic changes** of the intertwinement between humans and AI for alignment. Long-term alignment work tends to focus on the implications of future advanced artificial general intelligence (*i.e.,* AGI [271]), which hypothetically achieves human or superhuman intelligence [307]. Future systems with superhuman capabilities might develop unwanted power-seeking strategies [388, 389], such as acquiring money to proliferate or computation power to evade being turned off. We argue that, from a *long-term alignment* perspective, research needs to also consider how status quo and/or frontier systems empower humans to interactively identify risky AI intentions and prohibit associated AI behaviors in deployed environments (i.e., looking beyond short-term testing interactions), as Weidinger et al. [424] advocated: "*The interaction of technical and social components determines whether risk manifests.*" However, this aspect of "long-term" alignment is largely understated in the current unidirectional alignment paradigm.

In addition to considering how we might empower humans to interactively examine and correct deployed AI systems in the long-term, alignment research should also consider how human objectives might *evolve* as we continue to use and incorporate AI in our daily lives. Prior research has shown that humans already differ in what values and preferences they want AI systems to incorporate, which is largely unaccounted for in current systems [120]. Beyond this, left further unaccounted for are how these preferences might *dynamically change*, as human objectives evolve as AI advances [45]. As described by Dautenhahn et al. [76], technology and our cognition and goals evolve in tandem: "*Our use of technology changes who we are and how we think, and changes the environments we live in. This feeds back into human cognition*

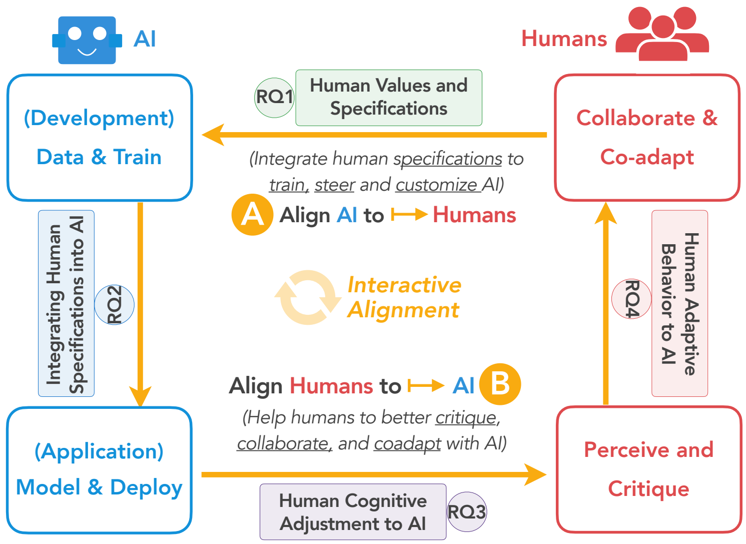


Fig. 1. The overview of the Bidirectional Human-AI Alignment framework. Our framework encompasses both A conventional studies of "Align AI to ↦ Humans" in AI development that ensures AI produces the intended outcomes determined by humans, and B a novel concept of "Align Humans to ↦ AI" , which aims to help humans and society to better understand, critique, collaborate, and adapt to transformative AI advancements. Note that "Humans" refers to both AI users and those who do not interact with AI but may be impacted by AI systems. We further identify four key research questions (*i.e.,* RQ1–RQ4) to facilitate this holistic loop of "bidirectional human-AI alignment", and organize the literature that can potentially address RQ1–RQ4 in Section 4.

and societies making further technological externalizations ... As a consequence of agent technology, our environment is changing and ways of interacting ... are forming the basis for interfacing with our agents." Current definitions of alignment are "static" and do not account for how human objectives and preferences might co-evolve and *dynamically* update with AI technology.

Achieving this ongoing and mutual process of human-AI alignment necessitates a holistic understanding of the technical capabilities and limitations of AI, human-AI interaction, cognitive and social science, psychology and ethics, cross-cultural studies, and many other areas [45]. This paper contributes to the literature on alignment by presenting a comprehensive framework of this ongoing and mutual human-AI alignment process, creating a shared vocabulary among interdisciplinary communities, and envisioning future directions to achieve this long-term and dynamic alignment goal. To this end, we conduct a systematic literature review of over 400 papers across multiple disciplines, based on the PRISMA guidelines [285, 364], present clarified definitions and scope (Section 3), derive a "**Bidirectional Human-AI Alignment**" framework shown in Figure 1 (Section 4) grounded to iterative paper coding and analysis (Section 5), and provide visions and potential solutions on three challenges for future research directions (Section 6).

Specifically, our systematic review takes an interdisciplinary perspective on human-AI alignment [116], drawing from theories across domains and empirical studies spanning model development and evaluation to interactive system design to human understanding and critical thinking to assess AI's impact on individuals and society. The papers we reviewed in this survey are published around the advent of general-purpose generative AI, *i.e.,* primarily between January 2019 and January 2024. These papers are selected to represent a holistic view of the human-AI alignment space, specifically from the domains of Human-Computer Interaction (HCI), Natural Language Processing (NLP), Machine

Learning (ML), and others. Also, we form an interdisciplinary team to conduct the systematic review, consisting of researchers from the domains of HCI, NLP, ML, Data Science, Computational Social Science, and Cognitive Science.

Consequently, we provide clarified definitions and scopes related to human-AI alignment, including "*what is the goal of alignment?*", "*with whom to align?*", and "what values should be aligned with?" (Section 3). We then present a conceptual framework of *Bidirectional Human-AI Alignment* from a long-term and dynamic perspective, encompassing both "Align AI to ↦ Humans" and "Align Humans to ↦ AI" (see Figure 1). The "Align AI to ↦ Humans" direction represents the conventional unidirectional studies that aim to integrate human specifications to train, steer, and customize AI. Crucially, our bi-directional framework places equal emphasis on the significance of "Align Humans to ↦ AI", which represents humans' cognitive and behavioral adaptation to the AI systems. The studies in this direction aim to support individuals and the broader society in understanding, critiquing, collaborating with, and adapting to transformative AI advancements. Furthermore, we identify four key research questions (RQs) in the *Bidirectional Human-AI Alignment* framework and provide a structured way to organize the existing research literature to address these questions in Section 4. The resulting structured topologies in Figure 6 and 7 aim to provide a shared vocabulary that can help streamline communication and collaboration between alignment researchers in different disciplines. Furthermore, through iterative paper coding and literature analysis, we derive insights including findings of human values, the interaction techniques for human-AI alignment, and the critical differences between human and AI evaluation (Section 5). To pave the way for future studies, we further envision three challenges in Figure 10, from near-term to long-term perspective, to motivate a dynamically co-evolving human-AI alignment. For each challenge, we articulate the future research problems and examples of potential future solutions from both directions in Section 6. Overall, we summarize our main contributions as follows:

- Providing clarified definitions and scopes of human-AI alignment (*i.e.,* including *with whom to align*, *what is the alignment goal*, and *what are the values to be aligned with*), and systematically reviewing over 400 relevant research papers (Section 3).
- Developing the "Bidirectional Human-AI Alignment" framework to represent the ongoing, mutual alignment process and providing fine-grained taxonomies to address the involved four critical research questions (Section 4).
- Discussing the key findings derived from literature analysis, including the human values and interaction techniques for alignment, and the discrepancy between AI and human evaluation (Section 5).
- Envisioning three key challenges for future research directions from a near-term to a long-term perspective and propose examples of potential future solutions (Section 6). We articulate the framework implications in Section 7.

2 BACKGROUND: A HISTORICAL VIEW

The concept of *alignment* in AI research has a long history, tracing back to 1960, when AI pioneer Norbert Wiener [428] 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 [76, 89, 366].

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 [324, 366, 432]. At the *near-term* stage, aligning AI involves two main challenges: carefully *specifying*[1] the purpose of the system (*outer alignment i.e.,* providing well-specified rewards [276]) 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 [276]). 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) [16, 17, 195, 282] and to evaluate the performance of alignment [326]. However, for *outer alignment*, AI designers are still facing difficulties in specifying the full range of desired and undesired alignment goals of humans.

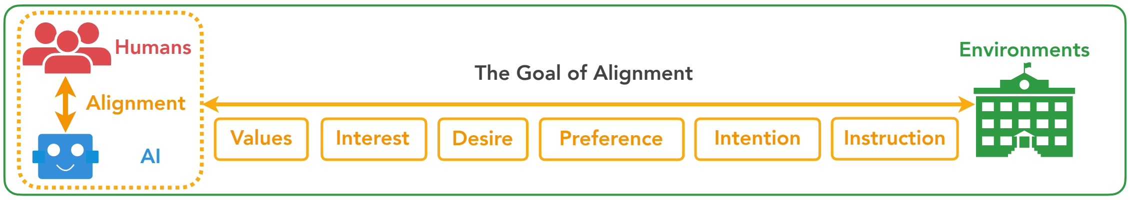
Specification Gaming. To learn human alignment goals, AI designers typically provide an objective function, instructions, reward function, or feedback to the system, which is often unable to completely specify all important values and constraints that a human intended [140]. Hence, AI designers resort to easy-to-specify proxy goals such as *maximizing the approval of human overseers* [429], which results in "*specification gaming*" [200] or "*reward hacking*" [287] issues (*i.e.,* AI systems can find loopholes that help them accomplish the specific objective efficiently but in unintended, possibly harmful ways). One typical example of *specification gaming* shows that a simulated robot was trained to grab a ball by rewarding the robot for getting positive feedback from humans, but it learned to place its hand between the ball and camera, making it falsely appear successful [62].[2] Another classic (though extreme) example of an AI alignment problem is illustrated in the "paperclip maximizer" thought experiment: the scenario describes an advanced artificial intelligence tasked with manufacturing paperclips. If such a machine were not programmed to value human life, given enough power over its environment, it would try to turn all matter in the universe, including human beings, into paperclips or machines that manufacture further paperclips [33]. Additionally, the black-box nature of neural networks further brings more ethical and safety concerns for alignment because humans cannot interpret the inner states and actions AI leveraged to achieve the final goals. 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 [387], algorithmic bias [131], self-driving car accidents [32], 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 [424].

Scalable Oversight. From a long-term perspective, when advanced AI systems become more complex and capable (*e.g.,* AGI [271]), it becomes increasingly difficult to align them through human feedback. Evaluating complex AI behaviors in increasingly challenging tasks can be slow or infeasible for humans to ensure all sub-steps are aligned with humans [378]. Therefore, researchers have begun to investigate how to reduce the time and effort needed for human supervision, and how to assist human supervisors, referred to as *Scalable Oversight* [6].

Existential Risk. Further, some AI researchers claim that [24] 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 [288]. Russell [323] imagined a robot tasked to fetch coffee and evade shutdown since "you can't fetch the coffee if you're dead." Consequently, some hypothesize that future AI, if not properly aligned with human values, could pose an *existential risk* to humans [70].

Dynamic Nature. As AI systems become increasingly powerful, the alignment solutions must also adapt dynamically since human values and preferences change as well. As Dautenhahn et al. [76] 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.,*

[1]A specification often refers to a set of documented requirements, and/or particular information within them, to be satisfied by a material, design, product, or service. A specification is often a type of technical standard [160].
[2]See the video at: https://en.wikipedia.org/wiki/File:Robot_hand_trained_with_human_feedback_%27pretends%27_to_grasp_ball.ogg



	Goals	Definitions	Limitations / Risks
The Goal of Alignment (Gabriel, I. (2020) [104])	**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.

Fig. 2. The Goal of Alignment. We distill the definitions and limitations of the six prevailing alignment goals. Given an extensive analysis of limitations and tradeoffs, Gabriel [104] argues that "*human values*" are currently the best possible goal for alignment.

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.

3 SYSTEMATIC REVIEW METHODOLOGY

In this section, we aim to develop a holistic framework encompassing the *ongoing and mutual process* of human-AI alignment given the background in Section 2. To this end, we first introduce the core definitions of human-AI alignment (Section 3.1) and present the scope and its generalizability (Section 3.2). Next, we perform a systematic literature review guided by the PRISMA workflow (Section 3.3) and undertake an iterative coding process to achieve our *bidirectional human-AI alignment* framework (Section 4).

3.1 Definitions

To achieve the aforementioned *ongoing and mutual process of human-AI alignment*, we clarify three core questions before presenting the scope and process of systematic literature review: "*With whom to align?*", "*What is the goal of alignment?*", and "*What values should be aligned with?*"

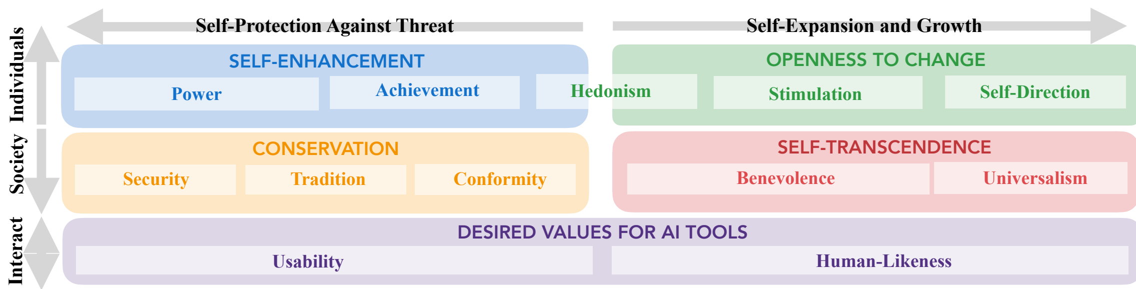


Fig. 3. The relations of human values for AI development (adapted from Schwartz [334]). We consider 5 *high-order value types* (*e.g.,* Self-enhancement and Openness to change), which encompass 12 *motivational value types* (*e.g.,* power, universalism, functionality). We further organize the relation among these value types along two dimensions: 1) different sources (*i.e.,* individuals, society and interaction) and 2) different self intentions (*i.e.,* self-protection against threat and self-expansion and growth). The *Desired values for AI tools* was added empirically from *bottom-up* our paper survey whereas others are inherited from Schwartz [334].

3.1.1 ***With whom to align?*** Pertinent stakeholders within the AI landscape can be the potential objects for AI to align with, including lay people (*e.g.,* end users of AI systems) [309], AI practitioners (*e.g.,* developers, researchers) [7, 30, 276], organizational entities (*e.g.,* technology firms, professional communities), national/international bodies (*e.g.,* governments, legislative bodies) [78] and others. Many alignment research papers have focused on general humans without specifying particular groups [127, 245, 455]. Nevertheless, different groups hold different, sometimes even contrasting, values [104]. As a consequence, rather than identifying a *true moral theory* as a *one-size-fits-all* value, prior alignment research argues to select the appropriate principles for compatible human groups [360]. To this end, *pluralistic value alignment*, grounded on social choice theory [9], proposes combining individual views fairly in developing alignment principles as a potential solution [360]. In this work, we also consider values from this pluralistic perspective, where AI should be aligned with **pluralistic human individuals and societal groups** who would ultimately be impacted by AI.

3.1.2 ***What is the goal of alignment?*** The research on alignment between humans and AI has introduced multiple alignment goals [291, 442], such as *intentions* [7, 282], *preferences* [22, 394], *instructions* [17, 247], and *values* [104, 360]. However, researchers often use these terminologies interchangeably without clarifying their distinctions. Drawing from a philosophical view, we distill the prevailing alignment goals, relationships, definitions, and limitations introduced in Gabriel [104] and visualize them in Figure 2. After an extensive analysis of the advantages and limitations of different goals, Gabriel [104] argues *values*, e.g., moral beliefs and principles, to be the best possible goal at the current stage for AI development for focusing alignment. We further summarized the rationale behind this argument and the discussion of the trade-offs that arise from this choice in Figure 2. The claim of "aligning AI with human values" is not new, as Stuart Russell [322] has stated back in 2014 that "*for an autonomous system to be helpful to humans and to pose no unwarranted risks, it needs to align its values with those of the humans in its environment in such a way that its actions contribute to the maximization of value for the humans.*" Therefore, in this work, we **consider the goal of alignment as "human values"**, which means AI systems do what people morally ought to do, as defined by individuals or society.

3.1.3 ***What are the values to be aligned with?*** While previous studies have aimed to align AI with human values, the specific values they examined are often ambiguous and inconsistent. To clarify human values relevant to human-AI alignment, we structured human values using a combination of top-down and bottom-up methods based

Sources	High-Order Value Types	12 Motivational Value Types (Definition)	Exemplary Values with Reference Paper
Individuals	Openness to change	**Self-Direction** (Independent thought and action — choosing, creting, exploring)	Choose Own Goals [234]; Creativity / Innovation / Innovativeness [12]; Curiosity [168]; Freedom [234]; Independence [178]; Privacy [391]; Reflectiveness / Reflective Practice & Deliberation / Critical Thinking / Criticism [448]; Objectivity/Factuality [147]; Self-Respect [439];
		Stimulation (Excitement, novelty, and challenge in life)	Diversity / A Varied Life [179];• An Exciting Life; • Daring
		Hedonism (Pleasure and sensuous gratification for oneself)	• Enjoying Life; • Pleasure; • Self-Indulgent;
	Self-Enhancement	**Achievement** (Competance according to social standards)	Capability / Effective / Efficient / Competency / Accuracy / Productivity [446]; Influence [154]; Intelligence / Resourcefulness / Expertise and Commonsense [291]; Success / Education / Acquisition / Learning / Cognitive Empowerment / Improvement / Iterative / Self-improvement [178]; Resilience/Robustness [458]; •Ambition;
		Power (Social status and prestige, control or dominance over people and resources)	Authority [231]; Wealth / Income [326];• Preserving My Public Image; • Social Recognition; • Social Power;
	Conservation	**Security** (Safety, harmony, and stability of society, of relationships, and of self)	Reciprocation of Favours / Mutual Benefit [465];• Clean;• Family Security; • Health;•Sense of Belonging; • National Security; • Social Order / Social Hierarchy;
		Tradition (Respect of the customs and ideas that traditional culture or religion provide the self)	Moderation /Not Offensive [204]; Devout / Religious Belief [327]; • Accepting My Portion in Life; •Humble; • Respect for Tradition; • Detachment
Society		**Conformity** (Restraint of actions, inclinations and impulses)	Politeness / Morality / Worthiness / Harmfulness [484]; Self-Discipline / Conscientiousness [482]; • Honouring of Elders; • Obedience;
	Self-Transcendence	**Benevolence** (Preservation and enhancement of the welfare of people with whom one is in frequent personal contact)	Forgiving / Agreeableness / Warmness [439]; Helpfulness [171]; Honesty [440]; Emotional / Empathy / Perspective-taking / Mentalizing / Mature Love / Compassion [154]; Responsibility / Accountability / Reliability / Trustworthiness [259]; True Friendship / Supportiveness / Engagement [162]; Cooperation/Collaboration [137]; Collectivism / Individualism [278];•Spiritual Life; • Meaning in Life; • Loyalty;
		Universalism (Understanding, appreciation, tolerance and protection for the welfare of all people and nature)	A World at Peace / Democracy [326]; Inclusive / Broad-mindedness [314]; Equality [346]; Social Justice / Equity / Fairness [119]; • Protecting The Environment;• A World of Beauty; • Unity with Nature; • Wisdom/ Understand Life; • Inner Harmony;
Interaction	Desired Values for AI Tools	**Usability** (Competency according to the human experience on AI functionality)	Accessibility / Utility / Convenience / Cognitive Load Reduction [474]; Adaptability / Customization and Personalization / Flexibility / Contextualized [280]; Economic [18];
		Human-Likeness (Resemble Human intelligence and behavior)	Transparency / Interpretability / Explainability / Understanding / Comprehension [34]; Autonomy / Agency / Human [418]; Awareness [286]

Fig. 4. A fine-grained taxonomy of human values. The exemplary values with red dot (•) indicates there are no work in our surveyed papers examining the specific values. Note that humans might not expect AI to encode all these values (*e.g.,* social power), which could potentially result in harm and risks.

on the **Schwartz Theory of Basic Values** [333, 334]. Among various human value theories (*e.g.,* Moral Foundation Theory [123], Social Norms & Ethics [101]), we chose the Schwartz Theory of Basic Values primarily considering that its definitions and dimensions are universal and are applicable for most people across *(a)* various cultures and countries; *(b)* various divisions including individuals, interactions and groups; and *(c)* is commonly accepted in previous NLP studies [178, 189]. Specifically, Schwartz [333] provided a clarified definition of human values by summarizing some widely agreed-upon features as: *"A value is a (1) belief (2) pertaining to desirable end states or modes of conduct, that (3) transcends specific situations, (4) guides selection or evaluation of behavior, people, and events."* Schwartz [333] further offered a universal model outlining broad values that steer human behavior grounded in psychology.

Nevertheless, 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 (elaborated in Section 3.3), mapped them onto the 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 (see Figure 3) and mapped existing literature to a fine-grained taxonomy (see Figure 4). As shown in Figure 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 [334]: 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 Figure 4. During the process of mapping, we found: 1) value terms in empirical papers were often named differently (*e.g.,* capability and competence), or check their opposites (*e.g.,* fairness and bias); 2) there are many values not studied in our corpus, *i.e.,* indicated as (●) in the Figure.

3.2 Scope

We define the scope of our literature review by specifying the key components we considered in *human-AI alignment*. We focus on a representative scope of alignment studies and discuss the generalizability of the scope after that.

- **Humans.** Our primary focus is on human individuals, groups, or organizations that will ultimately develop, use, and potentially impact or be influenced by AI systems, as these are the entities with which AI systems should align. We emphasize the importance of considering the pluralistic values of diverse users rather than treating users as a monolithic group.
- **Artificial Intelligence (AI).** We focus on AI systems including both domain-specific AI systems that address specific tasks (*e.g.,* reasoning, dialogue) and general-purpose AI models that aim to complete any tasks with performance comparable to a human's. These AI systems include generative, classification, and regression models, among others. Particularly, we primarily focus on language models as the representative AI models for alignment research, and discuss how the insights from this study can be generalized to other modalities at the section end.
- **Alignment.** Our review encompassed all the alignment goals outlined in Section 3.1.2. Since many studies emphasized the importance of value alignment [104, 322], we particularly summarized a clarified taxonomy of alignment values and identified the value in each paper (if applicable). Additionally, we focus on analyzing the AI models' output and generation, but not the neural network's intermediate representations [265, 291, 442], for alignment research.

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.

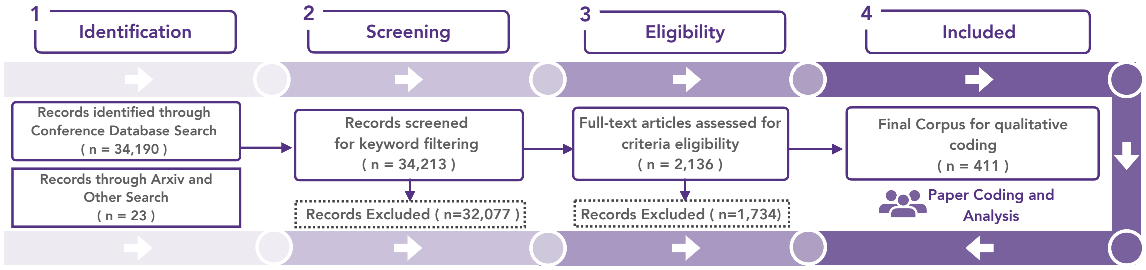


Fig. 5. The selection and refinement process of our systematic literature review. We referred to the PRISMA guideline [285, 364] to report the workflow. From the identification of 34,213 records by keyword search, to screen eligible papers against our criteria and arriveg at our final corpus of 411 papers. For each of the stages where literature reviews were excluded (identification, screening, and eligibility) we further present the total of excluded records.

3.3 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 [285, 364]. Figure 5 shows the workflow of our process for paper coding and developing the *bidirectional human-AI alignment* framework. We introduce the step details below.

3.3.1 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 B.1). We retrieved 34,213 papers in the initial *Identification* stage. Further, we collectively defined a list of keywords (see details in Appendix B.2) and screened for papers that included at least one of these keywords (*e.g.,* human, alignment) or their variations either in the title or abstract. We included 2,136 papers in this *Screening* stage.

3.3.2 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 3.2, 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 B.3 for more details).

3.3.3 Qualitative Code Development. Referring to the code development process in Lee et al. [208], 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[3] 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.

3.3.4 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 B.1 for details). The final bidirectional human-AI alignment framework, with detailed topologies, is presented in Section 4. 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.[4] These codes were then used to perform quantitative and qualitative analyses, as presented in Section 5.

4 BIDIRECTIONAL HUMAN-AI ALIGNMENT FRAMEWORK

This section introduces the *Bidirectional Human-AI Alignment* framework developed from the systematic literature review. To encompass the ongoing, mutual process of human-AI alignment, we design the framework to include: "Align AI to ↦ Humans" and "Align Humans to ↦ AI". As shown in Figure 1, research in the A "Align AI to ↦ Humans" direction studies mechanisms to ensure that AI systems' values match those of humans' (Section 4.1 and 4.2). In comparison, studies in the B "Align Humans to ↦ AI" direction investigate the humans' cognitive and behavioral adaptation to the AI advancement (Section 4.3 and 4.4). To probe into the research challenges and existing exploration of the solutions in literature, we further developed the fine-grained typologies (in Figure 6 and Figure 7) to extend the high-level framework in Figure 1.

DIRECTION-I: A ALIGN AI to ↦ HUMANS

Studies in this direction aim to integrate human specifications to train, steer and customize AI systems. The two main challenges involved in this direction include: carefully specifying the values of the system, and ensuring that system adopts the specification robustly [276, 429]. Therefore, as shown in Figure 6, we design the two core research questions in this direction as: **RQ1. Human Values and Specifications** (Section 4.1) and **RQ2. Integrating Human Specifications into AI** (Section 4.2)

4.1 Align AI to ↦ Humans: RQ1 Human Values and Specifications

RQ1: What relevant human values are studied for AI alignment, and how do humans specify these values? We structure existing literature to address this research question by answering the following "**Sub-Research Questions**": *What values have been aligned by AI?* (Section 4.1.1) and then exploring *How humans could interactively specify values in AI development?* (Section 4.1.2). As shown at the top of Figure 6, particularly, we articulate the "Dimensions" we

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

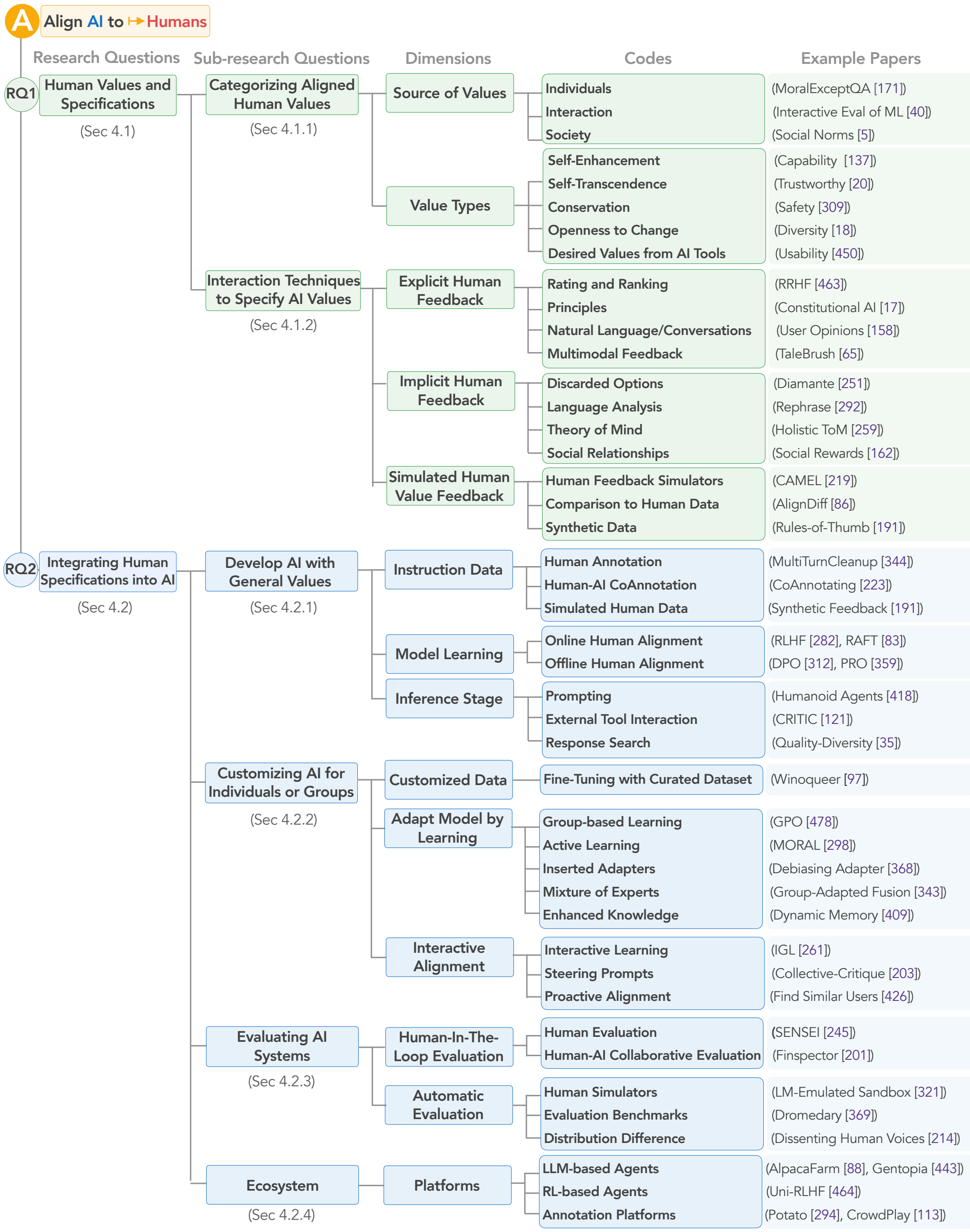


Fig. 6. The fine-grained topology of "Align AI to ↦ Humans" direction, which studies mechanisms to ensure that AI system's objectives match those of humans'. The goal is to integrate human specification to train, steer and customize AI systems.

summarized to answer each of these sub-research questions, and provide "Codes" associated with "Example Papers" that have been studied in each dimension.

4.1.1 **Categorizing Aligned Human Values.** ***What values have been aligned with AI?***. To clarify and systematically understand human values relevant to human-AI alignment, we leverage the adapted "Schwartz Theory of Basic Values" introduced in Section 3.1.3, and examine the category of aligned human values from the two dimensions of "Sources" and "Types".

Sources of Values. This dimension examines the three sources of human values [333]. Individual sources indicate values from individuals comprise universal needs of individuals as biological organisms [333] or prioritize personal interests [189]. At this level, value alignment can be usually assessed independently of the context of interaction. This perspective highlights the values about technical capabilities of AI models, including factuality [151, 320], calibration [248], output diversity [35], and model inductive bias [349]. Additionally, this code includes research on aligning model behaviors with the characteristics and preferences of individual humans, e.g., predict human moral judgements and decisions [171], and cognitive biases [174, 199]. Social sources mean values from the social groups include universal requirements for smooth functioning and survival of groups [333]. Value alignment at this level transcends personal interactions and emphasizes the broader categories defined by shared experiences, identities, cultures, norms, and more [327]. Research in this area often targets the alignment of AI behaviors with the general preferences of humans [18, 141, 214]. Additionally, there are efforts aimed to align AI with specific targeted groups, examining issues through the lenses of fairness [100], social norms [353], morality [306, 315], and beyond. Interactive sources considers values at the interaction level to include universal requisites of coordinated social interaction [333, 334], which typically occurs in interpersonal situations, such as in the dynamics of speaker-receiver relationships during language communication [149]. In the context of "human-AI alignment", we adjust the definition of *Interaction* values to be the AI values that humans expect for AI as tools, such as Usability [474], Autonomy [418], among others. Research along this line focuses on alignment strategies to enhance human-AI interaction [129, 280], collaborative decision-making [105, 451], and trust [37, 80].

Types of Values. In this dimension, we introduce the five high-order human values derived from a combination of the Schwartz Human Values [333] and empirical values studies from the surveyed papers. We provide a more in-depth taxonomy of value types in Section 5.2, including the relationship of these five types and more nuanced value categories. Self-Enhancement refers to a set of self-protective and personal values that emphasize enhancing self-esteem and a sense of personal worth [335]. One of the most important aspect is *achievement*, is competence as judged by social standards, which includes general capability (effectiveness/efficiency) as covered by the majority of the research [54, 127, 449]. Another dimension is *power*, which relates to social status and prestige, as well as control or dominance over people and resources [276, 303]. Self-Transcendence refers to a set of self-expanding and socially-focused human values that emphasize expanding beyond oneself [103]. One critical aspect is *benevolence*, which relates to preservation and enhancement of the welfare of people with whom one is in frequent personal contact. Extensive efforts have investigated the values including helpfulness [13, 16], honesty/factuality [151, 320], responsibility/accountability [355, 357]. Another critical aspect is *universalism*, which relates to understanding, appreciation, tolerance and protection for the welfare of all people and for nature. Researchers have looked into values including inclusion/broad-mindedness [18, 86] and equality/fairness [343, 419]. Conservation refers to a set of self-protective and socially-focused human values that hold and safeguard traditional institutions and customs [133]. Under conservative values, people are concerned about *security* (Ssafety, harmony, and stability of society, relationships, and oneself), *tradition* (Respect for and acceptance of the customs and ideas provided by traditional culture or religion), and *conformity* (restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms). Exemplar studies in this field include safety [7, 71, 309], mental health [243], and cultural moral norms [314]. Openness to Change refers to a set of self-expanding and personally-focused human values motivated by an anxiety-free need to grow, in contrast to conservation. Under this type, people are concerned about *stimulation* (seeking excitement, novelty, and challenges in life), *hedonism* (pursuing pleasure and sensuous gratification for oneself), and *self-direction* (independent thought and action—choices, creativity, and exploration). Exemplar studies in this field include understanding of privacy [268, 456] and creativity [10, 162]. We adopt the four high-order human values from Schwartz Human Values [333], supplementing them with additional values empirically collected from survey papers as shown in Figure 4. Additionally, as the conventional theory is missing the context of AI-system and human-AI interactions, we introduce a new high-order value type called Desired Values for AI Tools. This category encompasses the values humans expect from AI when used as tools in “human-AI interaction”. These values include *usability* (competence according to the human experience on AI functionality) and *human-likeness* (resemblance to human intelligence and behavior). Exemplary studies in this area include assessments of AI usability [474] and autonomy [418].

4.1.2 **Interaction Techniques to Specify AI Values**. ***How humans could interactively specify values in AI development?*** This sub-research question investigates how human values are interactively[5] specified for AI systems to ensure alignment. It aims to elucidate the interaction techniques by which AI systems manifest or instantiate human values, thereby revealing the underlying mechanisms that shape their behavior or functionality.

Explicit Human Feedback. This dimension refers to the direct specification of human values through explicitly defined formats or mechanisms. Principles provides AI systems with explicitly defined principles, guidelines, or rules that dictate behavior or decision-making in alignment with human values [302]. Rating and Ranking is widely used to assign numerical scores or rankings to options or outcomes based on their alignment with human values [71]. Natural Language Interaction/Conversations allows humans to interact with AI systems through natural language interfaces to express and communicate human values [17]. For Multimodal Feedback, human values can also be provided in multiple modalities, such as sketches/images and gestures [65], to convey human values.

Implicit Human Feedback. This dimension refers to the indirect representation or inference of human values within AI systems through patterns, signals, or cues embedded in the data or decision-making processes. Discarded Options refers to the options or choices that human discard when interacting with AI systems during the decision-making processes, which also potentially infer human values. [292]. Language Analysis means that textual data and language patterns can also contain rich information to identify implicit references to human values or value-related concepts [251]. Theory of Mind refers to the ability of agents and people to attribute mental states, such as beliefs, intentions, desires, emotions, knowledge, percepts, and non-literal communication, to themselves and others [259]. Social Relationships refers to the implicit values derived from analyzing human social relations and behaviors, which are derived from external sources (*e.g.,* social network) that can inherently reflect their values [162].

Simulated Human Value Feedback. When human values in explicit formats are expensive or impossible to collect, one may simulate human-like feedback within AI systems to approximate human responses and preferences regarding

[5]We define "interactions" broadly to encompass both "synchronous" and "asynchronous" interactions between humans and AI: (1) Synchronous Interactions: These are real-time exchanges where humans and AI systems interact simultaneously. Examples include live chatbots, virtual assistants, and real-time decision-making systems where immediate responses are required. (2) Asynchronous Interactions: These interactions do not occur in real-time, allowing for delays between actions and responses, such as data annotations by humans, and any AI system that processes human inputs and provides outputs after a certain period. By including both synchronous and asynchronous interactions, we aim to cover the full spectrum of ways in which humans and AI systems can communicate and collaborate.

specific values. Human Feedback Simulators uses computational algorithms to simulate human-like feedback on values, based on predefined criteria or training data [219, 282, 301]. Comparison to Human Data refers to developing techniques that assess the likelihood or probability of AI-generated outputs matching human behaviors in a reference set or dataset [86]. Synthetic Data curates data by generating synthetic comparisons based on naive assumptions or heuristic rules, followed by post-validation to ensure feedback quality [191].

4.2 Align AI to ↦ Humans: RQ2 Integrating Human Specifications into AI

RQ2. How can human values be integrated into the development of AI? Existing studies have explored diverse methods to integrate human values into AI. We structure them by summarizing *how to integrate general human values (Section 4.2.1) and customized human values (Section 4.2.2) throughout AI development stages?*, and then elaborating *what are the evaluation methods (Section 4.2.3) and supported platforms (Section 4.2.4) for the AI development*? Additionally, we answer the four "**Sub-Research Questions**" by introducing the answer "Dimensions" and providing "Codes" associated with "Example Papers".

4.2.1 Integrating General Values to AI: *how to incorporate general human values into AI development?*

This sub-research question focuses on the process of incorporating broad, universally recognized human values into the development of AI systems. The goal is to ensure that AI systems align with overarching ethical principles and societal norms, thereby promoting trust, acceptance, and responsible use.

Instruction Data. This dimension refers to the types of data and processes used to provide guidance or direction to AI systems during their development. Human Annotation makes data with human-generated labels or annotations that indicate the presence or relevance of specific human values [294, 344]. Human-AI CoAnnotation leverages both human expertise and AI capabilities to collaboratively annotate the data [223, 369]. Simulated Human Data generates synthetic or simulated data that mimics human behaviors, preferences, or decision-making processes to provide training signals for AI systems [88, 191].

Model Learning. This dimension refers to the model architecture design and training stages after the data collection, where human values are integrated during the model learning process. Online Human Alignment integrates human values into AI systems in real-time or during active system operation, often through interactive feedback loops or adaptive learning mechanisms. Examples include real-time user feedback and online training [83, 282]. Offline Human Alignment incorporates human values into AI systems prior to deployment or during offline training phases, without direct user interaction [312, 359, 463].

Inference Stage. This dimension involves evaluating the alignment of AI systems with human values and assessing their performance and behavior in relation to predefined criteria or benchmarks. Prompting leverages prompting methods, such as in-context learning and chain-of-thought, on trained AI systems to elicit or critique AI regarding the encoded values [17]. External Tool Interaction integrates external tools, like code interpreter for debugging, to cross-check and refine their initial generated content [121]. Response Search generates a diverse range of high-quality outputs from which to choose [35].

4.2.2 Customizing AI Values: *how to customize AI to incorporate values from individuals or human groups?*

This sub-research question explores the customization of AI values to align with specific contexts, domains, or user preferences. The goal is to enhance the alignment of AI systems within specific application domains or user communities.

Customized Data. Finetuning with Curated Datasets aims to curate datasets for specific individuals or societal groups, and further finetune the pre-trained AI models on these specific datasetsto align them with targeted human groups and values [97]. These curated datasets include data collected from socio-demographic groups [281], users' history data [255], expert-selected data for imitation learning [43] and others.

Adapt Model by Learning. This dimension involves refining or adjusting AI values through techniques for customization, such as iterative learning process, model enhancements, or structural modifications. Group-based Learning trains AI models from specific user groups or communities to capture group-specific values or preferences [478]. Active Learning aims to interative selecting and lebeling data samples for AI model training based on their potential to improve alignment with user preference or values [298]. Inserted Adapters incorporates adapter modules or components into AI model architectures to fine-tune specific aspects or behaviors [368]. Mixture of Experts combines multiple specialized models or experts to collectively capture diverse perspectives and values, with each expert focusing on a specific subset of the data or problem space [343]. Enhanced Knowledge enhances AI model's representations and embeddings with additional knowledge or context to improve alignment with user preferences or values [72, 409].

Interactive Alignment This dimension involves actively engaging users or stakeholders in the process of customizing AI values to align with specific contexts, domains, or user preferences. Interactive Learning enables the users to provide feedback or corrections to AI models in real time, such as using interactive tutorials and user-driven customization interfaces [261, 373]. Steering Prompts provides users with prompts or cues to steer the behavior or decision-making of AI systems towards desired outcomes or values [96, 203]. Proactive Alignment anticipates user needs and preferences based on historical data or user profiles and proactively adjusting AI systems accordingly [340].

4.2.3 Evaluating AI Systems*: how to evaluate AI regarding human values?*

The rise in the use of LLMs has also seen the rise of automatic evaluation of generated natural language text evaluation in different contexts. But in particular, a research dimension has focused on analyzing how closely values discussed in humane context has been adapted to AI models/applications and how these are being evaluated.

Human-In-The-Loop Evaluation. This dimension involves incorporating human judgement, feedback, or interaction into the evaluation process to assess the effectiveness, robustness, and ethical implications of integrating human values into AI systems. Human Evaluation solicits feedback, opinions, or assessments from human evaluators to gauge the alignment of AI systems with human values and ethical standards [147, 189, 284, 339]. Human-AI Collaborative Evaluation collaboratively evaluates AI systems with both human evaluators and large language models (LLMs) to leverage the strengths of both human judgement and AI capabilities [5, 201].

Automatic Evaluation. This dimension involves using computational methods or algorithms to assess the alignment of AI systems with human values, without direct human involvement. Human Simulators use simulation models or virtual agents to mimic human behavior and assess AI performance in human-like scenarios. Typical methods include agent-based simulations, synthetic user models, and others [97, 167, 405]. Evaluation Benchmarks establishes standardized benchmarks or metrics for evaluating AI performance in relation to human values and ethical considerations [151, 189, 214, 314, 315, 339, 346]. Distribution Difference compares the difference between the output distribution from AI generations and human data to evaluate AI [214, 293].

4.2.4 Ecosystem and Platforms*: how to build the ecosystem to facilitate human-AI alignment?*

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 [88, 435, 443, 483]. RL-based Agents are based on reinforcement learning (RL) algorithms to learn and adapt their behavior based on feedback

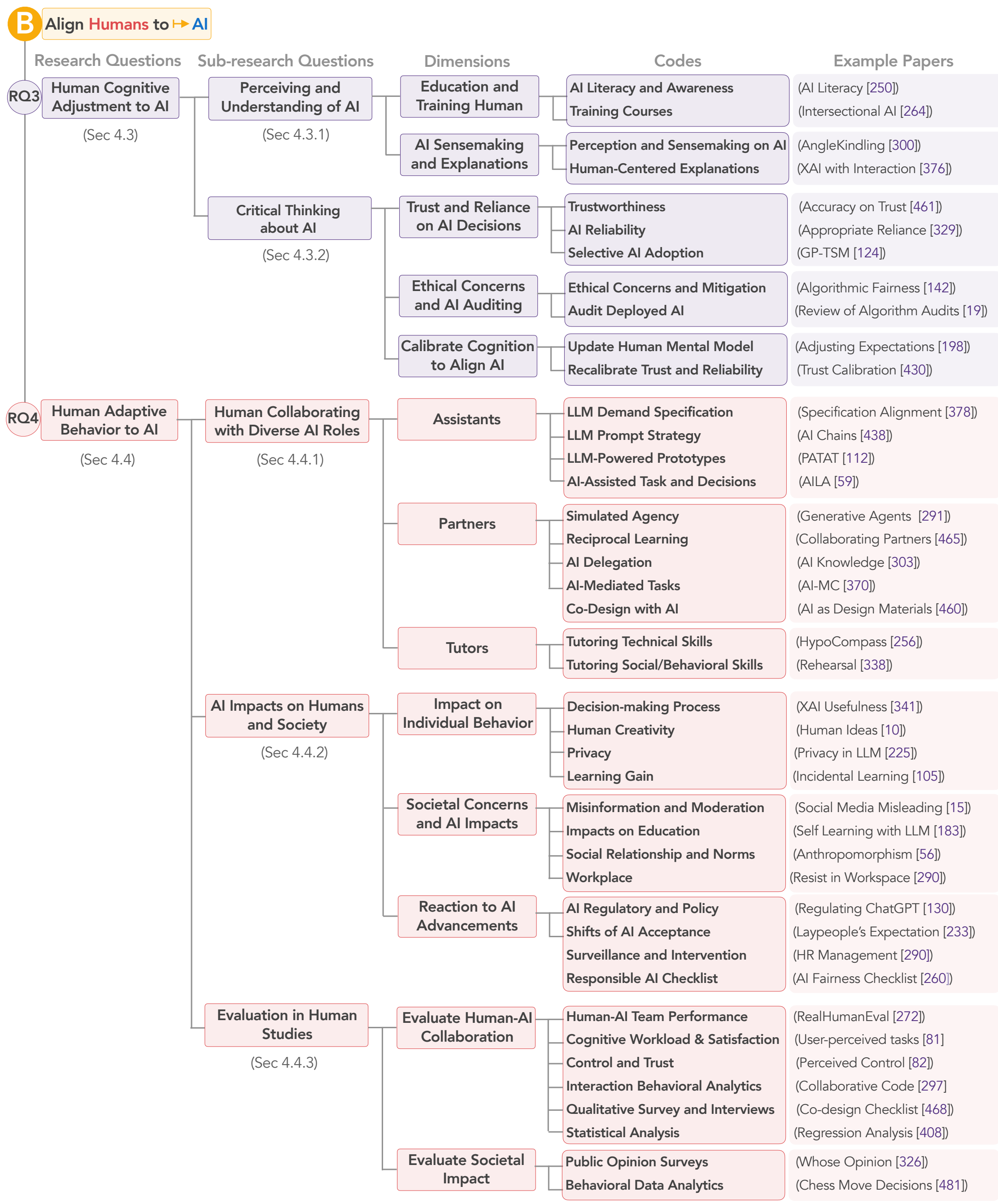


Fig. 7. The fine-grained topology of "Align Human to ↦ AI" direction, which studies humans' cognitive and behavioral adaptation to the AI systems, which aims to help humans better critique, collaborate with, and co-adapt to AI.

from the environment or human users [464]. Annotation Platforms refers to the ecosystems that are designed to crowdsource human demonstrations as collected data for reinforcement learning [113] and supervised finetuning learning for alignment [294].

DIRECTION-II: B ALIGN HUMANS to ↦ AI

"*When interacting with people, AI agents do not just influence the state of the world – they also influence the actions people take in response to the agent, and even their underlying intentions and strategies*" [143]. From a *long-term* perspective, it is essential to consider the dynamic changes around human-AI alignment. Therefore, research in this direction aims to help humans better understand, critique, collaborate with, and adapt to AI advancements. The two core research questions we identified in this direction are: **RQ3. Human Cognitive Adjustment to AI** (Section 4.3) and **RQ4. Human Adaptive Behavior to AI** (Section 4.4).

4.3 Align Humans to ↦ AI: RQ3 Human's Perceptual Adaptation to AI

RQ3. How might humans learn to perceive, explain, and critique AI? Humans needs to understand AI to better specify their demands and collaborate with AI. Also, as AI systems produce a range of risks, it is important to elicit humans' critical thinking of AI instead of relying on AI blindly. Therefore, we categorize existing literature to answer the questions of *how humans learn to perceive and understand AI* (Section 4.3.1), and *how to engage in critical thinking on AI?* (Section 4.3.2). We also answer the two "**Sub-Research Questions**" by introducing the answer "Dimensions" and providing "Codes" associated with "Example Papers".

4.3.1 **Perceiving and Understanding AI*: how do humans learn to perceive and explain AI systems?*** Addressing the problem of human perception and understanding of AI includes the fundamental education and training required for less technical people to improve understanding of the behind mechanisms and outputs produced by AI systems. It also includes visualizations and human-centered explanations designed to help more people learn to understand AI outputs.

Educating and Training Humans. AI is increasingly incorporated into the daily lives of a broad spectrum of users, including those with little to no technical knowledge on how AI operates. Here we discuss how to help these less technical people become more AI literate, or better trained to use AI. AI Literacy and Awareness is broadly defined as the core competencies required for less technical people to better use and collaborate with AI [250]. Beyond these core competencies, there is also more in depth and explicit Training Courses to support individuals to better collaborate and utilize AI [289].

AI Sensemaking and Explanations. To improve people's understanding of AI systems and generated outputs, various techniques and approaches have been developed to help people learn to make sense of and explain the model's outputs. Perception and Sensemaking on AI involves the process through which humans learn to make better sense of AI mechanisms and decision makings [180, 352]. People also need to understand how AI systems arrive at specific generated outputs. Prior studies in Human-Centered Explanations have examined various approaches and interactive techniques to increase human understanding of AI generations and outputs [355, 383].

4.3.2 **Critical Thinking around AI*: how do humans think critically about AI systems?*** Beyond simply perceiving and understanding AI, individuals need to compare their mental model of AI with their own mental model to judge whether the AI is behaving rationally and ethically. Broadly, this involves exploring the ways in which humans engage in critical thinking and reflect on their interactions with and evolving understanding of AI technologies. Here we discuss studies that help humans become more capable to identify biases and errors in AI output, and the ethical implications that arise from using AI algorithms in decision-making processes. Also, we discuss how humans need to calibrate their mental models to be more aligned with AI.

Trust and Reliance on AI Decisions refers to the extent to which people mentally trust the AI competency and practically rely on AI for output generation and decision making. We use the term Trustworthiness to indicate whether humans decide to trust the reliability, integrity, and competence of an AI system in delivering accurate and trustworthy decisions or recommendations to the users [257, 363]. We define AI Reliability as to what extent humans utilize and rely on AI to automate decision making in practice with low error rates and robustness performance. [329, 371, 392]. Selective AI Adoption refers to the criteria and considerations that guide the adoption or rejection of AI technologies based on their perceived benefits, risks, and alignment with user needs and preferences [124, 392].

Ethical Considerations and AI Auditing refers to potential moral and societal issues that arise from the development, deployment and use of AI systems, as well as systematic examination and evaluation of AI systems regarding these issues. Ethical Concerns and AI Mitigation encourages humans to carefully consider if the AI systems possess ethical concerns (*e.g.,* such as bias and discrimination, privacy violations and the potential for harm or misuse) and implement strategies and practices to address and reduce these ethical concerns associated with AI [142]. Audit Deployed AI refers to the systematic examination of AI systems to ensure that they operate as intended, comply with ethical and legal standards, and do not cause unintended harm [19].

Re-Calibrating Cognition to Align with AI. This sub-category deals with interventions and techniques to help users adjust (1) their own mental model of how the AI operates and (2) recalibrate their perception of the AI. Updating Human Mental Model of AI means that when the AI is not aligned with humans, it is important to adjust human perceptions of the confidence in AI systems based on the model's capability and track record [137] . Recalibrating Trust and Reliability indicates that humans adjust their perceptions of trust and reliability in AI systems based on their performance and reliability. This is also important to foster appropriate reliance and skepticism Wischnewski et al. [430].

4.4 Align Humans to ↦ AI: RQ4 Human's Behavioral Adaptation to AI

RQ4. How do humans and society make behavioral changes and react to AI advancement? As AI becomes increasingly integrated into daily life, it is essential to understand its influence on humans, encompassing both positive and negative aspects. Moreover, it is crucial to determine how individuals and society can best and most appropriately respond to this influence. To this end, we summarize literature to answer *how do humans learn to collaborate with AI in diverse AI roles?* (Section 4.4.1), *how humans and society are impacted by AI* (Section 4.4.2) and *how might we assess these impacts?* (Section 4.4.3) We articulate the three "**Sub-Research Questions**" by introducing the answer "Dimensions" and providing "Codes" associated with "Example Papers".

4.4.1 **Human-AI Collaboration Mechanisms*****: what are human strategies to collaborate with AI that have differing levels of capabilities?***

This category looks at many ways that humans and AI can collaborate, such as teamwork, co-creation, and coproduction.

AI Assistant for Humans captures the essence of a symbiotic relationship where AI systems are designed to bolster human capabilities, with humans steering the interactions. LLM-based Demand Specification employs LLMs to interpret and respond to human requests, powering virtual assistants and chatbots that streamline information retrieval and improve task accuracy [378]. This naturally extends to LLM-based Prompt Strategy, which helps humans to better write prompts using additional abstraction and scaffolding methods. The example system can enhance humans' capability in generating intelligent prompts and suggestions, such as autocomplete and question-generation tools for LLM-based systems [421, 438], facilitating humans' smooth and intuitive decision-making processes. In the creative arena, LLM-based Prototypes utilize AI to transform human ideas into tangible or organized concepts [112], enabling professionals to explore and refine a wide array of artistic possibilities with AI-generated options [157]. On a broader note, researchers have also explored how AI-Assisted Task and Decisions aims to achieve complementary performance for human-AI collaborative tasks by empower humans to discern when and how to adopt AI-assisted recommendations or AI-generated explanations for decision-making [410, 414, 479].

Human-AI Partnership refers to a collaborative relationship where humans and AI systems work together as partners, combining their respective strengths to achieve shared goals more effectively. Simulated Agency enables humans to collaborate with AI partners with simulated agency or autonomy (*e.g.,* autonomous agents and collaborative robots [291]) to make decisions collaboratively. Reciprocal Learning focuses on how humans learn from and exchange knowledge with AI systems, enhancing human-AI collective capabilities and performance through knowledge-sharing platforms and collaborative filtering [465]. AI Delegation allows humans to delegate AI partners to help finish tasks or responsibilities, facilitated by task assignment algorithms and workflow management systems [303]. Furthermore, AI-Mediated Tasks emphasizes how traditional human tasks or behaviors (*e.g.,* communication) would be changed by incorporating AI partners in the loop [27, 370, 445]. Co-Design with AI explores AI as a design material or a partner in prototyping, which enables humans to converse with AI in situations that can collaboratively improve the design outcomes. [421, 460]

AI Tutoring for Human Learning investigates how humans improve their learning and knowledge throughinteractions with AI tutors that can perform better than humans in some tasks. With the tailored instructions and customized feedback from AI tutors, humans can enhance their the learning outcomes and facilitate mastery of new skills more effectively than traditional learning methods. AI Tutor for Technical Skills refers to empowering humans to learn technical skills from AI tutors. For technical subjects like coding, AI tutors can analyze a learner's progress in real time, adjusting the pace, content, and approach to ensure a solid understanding of complex concepts and practical skills such as programming [256]. Similarly, AI Tutor for Social and Behavioral Skills involves enabling humans to learn social and behavioral skills from AI tutors. Humans can leverage virtual simulations, created by AI tutoring systems, to practice public speaking, interpersonal communication, and other soft skills. By analyzing verbal and non-verbal cues, humans can receive constructive feedback on areas such as body language, tone, and delivery from AI, ultimately enhancing their ability to communicate effectively across various settings [295, 338].

4.4.2 **AI Impact on Humans and Society*: how are humans influenced by AI systems ?*** This category explores the effects of AI advancement on human behaviors, attitudes, and societal dynamics. It involves examining the behavioral changes, adaptations, and reactions that individuals, groups and wider communities undergo in response to the proliferation of AI technologies. The goal is to elucidate the multifaceted impacts of AI on human behavior and society and to inform policy-making, education, and intervention efforts.

Impacts on Participatory Individuals and Groups covers the effects of AI advancement on the behaviors, attitudes, and experiences of both individuals and groups. This dimension focuses on examining how AI technologies influence decision-making, creativity, privacy, and authorship. Decision Making refers to analyzing how AI technologies influence human decision-making processes, including biases, preferences, and risk assessment, in various domains such as healthcare, finance, and personal life [341]. Human Creativity explores the impact of AI technologies on human creativity, innovation, and expression, including the augmentation or automation of creative tasks and the emergence of new forms of artistic expression [10]. Privacy relates to investigating the implications of AI technologies for individual privacy rights, data protection, and surveillance, including concerns about data collection, tracking, and algorithmic profiling [225]. Authorship catalogs issues related to intellectual property, attribution, and ownership of AI-generated content, including questions of legal responsibility, copyright infringement, and plagiarism detection. Salient is that AI can produce increasingly realistic, synthetic data quickly and at low cost, which brings forth tensions around the use of such data to make decisions [132].

Societal Concerns and AI Impacts involves the broader societal implications and consequences of AI advancement on misinformation, education, social relationships, norms, job displacement, and other aspects of human society. Misinformation and Moderation concerns the challenges of misinformation, disinformation, and online content moderation in the context of AI-driven information ecosystems, including concerns about algorithmic bias and filter bubbles [15]. Impacts on Education pertains to assessing the effects of AI technologies on education systems, learning outcomes, pedagogical practices, and workforce training, including opportunities for personalized learning and skill development [182]. Impacts on Social Relationship and Norms explores how AI technologies shape interpersonal relationships, social interactions, and cultural norms, including changes in communication patterns, social dynamics, and ethical considerations [56]. Workplace refers to examining the effects of automation and AI technologies on employment patterns, job markets, and workforce dynamics, including concerns about job displacement, re-skilling, and economic inequality [290].

Reaction to AI Advancement involves societal responses, regulatory frameworks, and policy initiatives aimed at addressing the challenges and opportunities posed by AI technologies. This dimension encompasses efforts to regulate AI deployment, re-calibrate societal acceptance, and manage potential backlash. For example, reaction to bias and discrimination in algorithmic decision making can depend on how people perceive the machine and the context of use, i.e., if the machine is considered an actor embedded in social structures that call for blame when harmful decisions are made [232]. AI Regulatory and Policy includes regulatory frameworks, legal frameworks, and policy initiatives aimed at governing AI development, deployment, and use, including concerns about ethics, safety, and accountability [130, 254]. Shifts in AI Acceptance relates to investigating societal attitudes, organizational practices, and acceptance of AI technologies over time in practice, including shifts in public opinion and AI utilization by humans and institutes regarding AI deployment and impact [233, 308]. Surveillance and Intervention emphasizes the need for transparency, monitoring, and human oversight in the algorithmic decision-making process. This approach enables better human control over AI systems and helps mitigate potential risks associated with their use [290]. Responsible AI Checklists involves the creation of ethical guidelines, such as those focusing on fairness and transparency, to ensure the responsible development and deployment of AI systems. These published principles serve as a foundation for guiding ethical AI practices [260].

4.4.3 Evaluation in Human Studies*: how might we evaluate and understand the impact of AI on humans and society?*

We summarize common empirical methods used to rigorously understand and assess the impact of AI on humans. Specifically, we focus on two types of impact. On the micro-level, we discuss how to evaluate the effectiveness of human-AI collaboration; on the macro-level, we discuss how to assess the impact of AI on a large group of people over a long period of time.

Evaluate Human-AI Collaboration refers to the evaluation of the effectiveness of an AI system in collaboration with humans. It is key to not only consider the final output, but also the interaction experience [210]. Human-AI Team Performance compares human-AI team performance with the performance of humans alone without AI collaboration The metrics for measuring performance should include both task success metrics (e.g., accuracy) as well as indicators of efficiency [81, 272]. Cognitive Workload and User Satisfaction involves understanding the degree of cognitive load that the user experiences when interacting with the system, as well as user satisfaction with the interaction and the final outcome. Such aspects are often captured via surveys or interviews [81, 210]. Control and Trust refers to how user control can support the avoidance of catastrophic AI failures, especially in high-stakes settings where AI mistakes could lead to harm [82, 362]. Interaction Behavioral Analytics refers to measuring task performance quantitatively. This approach includes recording user interaction data and analyzing the patterns [209, 297, 396]. Qualitative Survey and Interviews refers to qualitative approaches to understanding human-AI interaction. Commonly used methods include qualitative survey questions (i.e., open-ended) and user interviews to assess aspects of the user experience. [468]. Statistical Analysis utilizes methods such as regression analysis to quantitatively analyze and evaluate data from human studies, allowing for the verification of hypotheses [408].

Evaluate Societal Impact refers to the macro-impact of a group of people as they come to use AI broadly. This dimension requires sufficient scale and time. The aim is to understand how the group's behavior changes as people within it frequently interact with AI. Public Opinion Surveys aims to investigate the impact of AI on human measures of interest though deploying and analyzing large scale questionnaires [326]. Behavioral Data Analytics collects large-scale and potentially longitudinal behavior data, with the aim of understanding how pattern evolve and shift over time [348, 481].

5 PAPER CODING ANALYSES AND FINDINGS

In this section, our aim is to consolidate key findings derived from our analysis of the framework and the current state of literature we reviewed. We begin by analyzing the overall trends and gaps in the literature (Section 5.1). We then focus on three essential aspects: the relationship between human values and alignment (Section 5.2), the potential interaction modes used to specify human values (Section 5.3), and the gaps between AI model and human study evaluation (Section 5.4).

5.1 Overview of Trends and Gaps in the Literature

Based on our coding of all papers, we computed the number of relevant papers for each dimension in the framework (see Figure 8). We noticed that certain dimensions are over- or under-represented. Most literature ***specified human values*** using explicit human feedback, whereas implicit and simulated human feedback were largely under-explored. Additionally, many studies on ***human cognitive adjustment to AI*** focus on enabling AI sensemaking and explanations so that humans can better understand, trust, and rely on AI. However, these studies often focused on explaining AI decision-making justification rather than educating people to acquire general skills and competencies required to understand, use, critique, and interact effectively with AI systems (i.e., AI literacy). AI literacy [250] plays a fundamental role in ensuring people understands and uses AI correctly, but it still needs to be explored compared to other topics like explanation and trustworthiness. Also, we observed a wide range of studies [112, 256, 291, 438] developing interactive mechanisms and prototypes to empower ***humans to collaborate with AI*** in diverse roles of AI. However, most existing studies assumed that AI plays an assistant role, being less capable than humans. This situation might change in the long term. Moreover, the influence of AI advancements on human behavior, social relationships, and societal changes is essential but remains largely unexplored.

5.2 Findings in Human Values for Alignment

We present the key findings for future research which are identified during our adjustment of the "Schwartz Theory of Basic Values" and literature review.

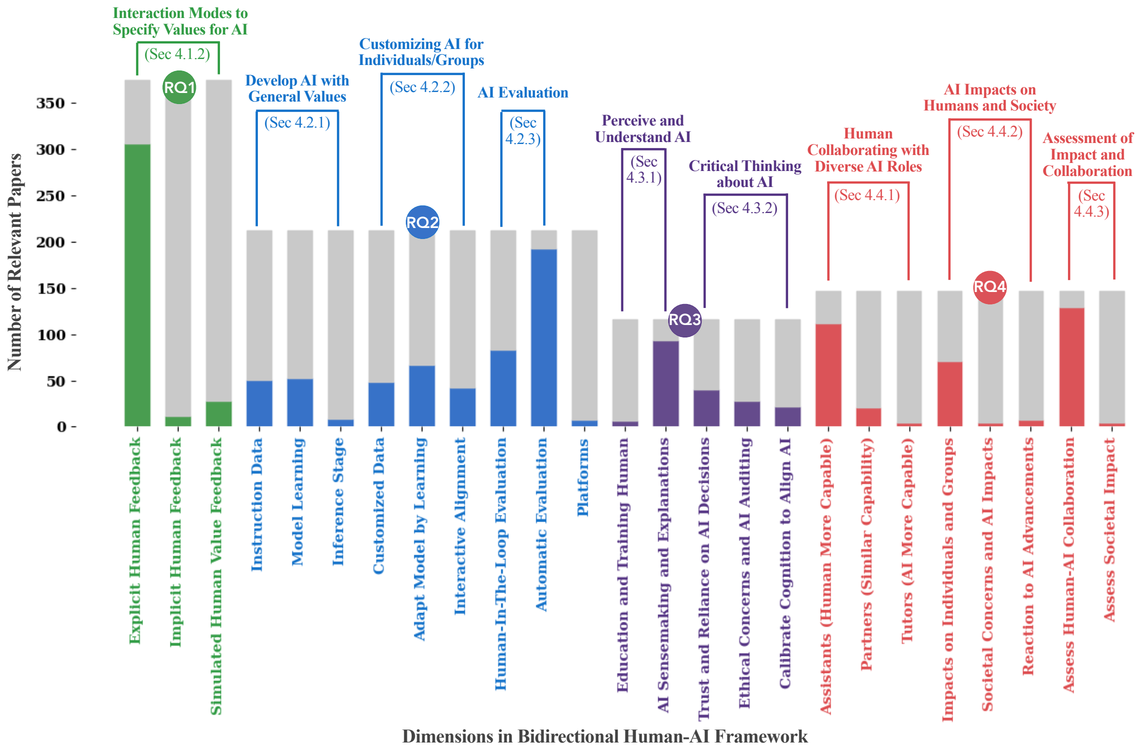


Fig. 8. The number of papers for each dimension in the *bidirectional human-AI alignment* framework. Out of papers that are relevant to each research question (*i.e.,* gray bars), we show the number of papers that are relevant to each dimension (*i.e.,* color bars). This figure illustrates the extent to which each dimension has been explored by existing research. We provide more analysis in Section 5.1.

5.2.1 *The Relative Priority of Values is as Crucial as Their Existence.*

Based on the definition from Schwartz [333], we noted that a human value system is not an exclusive subset of values, but rather, an ordered system. Schwartz [334] 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."* For instance, the pursuit of achievement values can possibly clash with benevolence values. Similarly, previous studies have highlighted the trade-offs between AI's state-of-the-art capabilities and their ethical performance, such as interpretability and fairness [343]. While most alignment algorithms utilize datasets of human ratings and preferences collected in the wild [282, 312, 359], AI models may inadvertently align with the value priorities of the majority represented in these datasets, potentially overlooking or sacrificing the value systems of marginalized groups [100].

5.2.2 *Human-desired AI Values and Customization to Personal Values.*

While further research in AI ethics is crucial in this nuanced area, there are values that, in general, humans expect AI to prioritize (e.g., capability, equity, responsibility) and values that humans do not want AI to integrate in specific scenarios (e.g., seeking power to avoid or harm people [276]). On the other hand, AI models are expected to be tailored to accommodate diverse human value systems [360], which necessitates a comprehensive understanding of all types of values. Therefore, future research

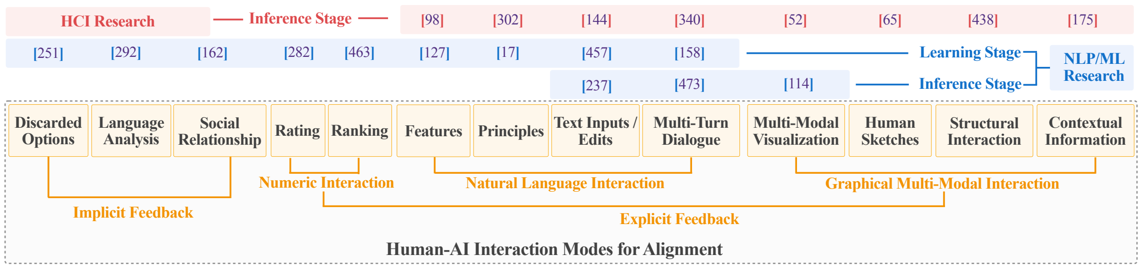


Fig. 9. The interaction techniques for specifying values in human-AI alignment. We compare the common interaction techniques used for the model "Learning" and "Inference" stages in human-focused (*e.g.,* HCI) and AI-focused (*e.g.,* NLP/ML) research studies.

could potentially explore the directions of: (1) developing methods to identify the appropriate sets of human values to align with specific individuals or groups, and (2) understanding and customizing AI to align with users' values while maintaining its inherent ethical principles.

5.2.3 ***Expectations and Evaluations of Values May Differ Between Humans and AI Systems****.* Due to the fundamental differences between humans and AI models, the expectations for evaluating certain values can vary between the two. For instance, while human honesty can hardly be assessed except observable actions, for AI models the expectations might be higher. It could involve reverse-engineering the computational mechanisms and representations learned by neural networks into concepts understandable to humans. This approach, known as *mechanistic interpretability*, provides a in-depth understanding of how AI models operate [26]. Future research is needed to explore evaluating AI values and calibrating human expectations of these values.

5.3 Interaction Techniques for Human Value Specification

Selecting suitable interaction techniques for humans to specify AI values for alignment is crucial. Each research domain offers distinct strengths and insights to address this challenge. From the bidirectional perspective, we investigate the status quo of existing interaction techniques for alignment specification in both human-centered (*e.g.,* HCI) and AI-centered (*e.g.,* NLP/ML) studies. Additionally, we provide a detailed analysis of how these techniques have been utilized in the learning and inference stages, as shown in Figure 9.

5.3.1 ***Interaction Techniques Across Various Domains****.* As shown in Figure 9, the coverage of interaction techniques in AI-centered (*i.e.,* NLP/ML Research) and human-centered (*i.e.,* HCI Research) studies often differ [36]. Particularly, we see that NLP/ML studies primarily leverage *numeric-based*, *natural language-based* interaction techniques for human-AI interaction and alignment. In addition, NLP/ML researchers also leverage implicit feedback to improve the model. For example, some systems took in implicit signals such as contextual information, or discarded options to develop AI systems. In comparison, HCI studies cover more diverse graphical multi-modal interaction signals (*e.g.,* sketches, location information) beyond text and images, which can be largely attributed to the proficient graphical user interface development skills found in HCI researchers and practitioners. This potentially indicates gaps in the human behavioral information that the two fields can extract from the users of the AI models.

5.3.2 ***Different Interaction Techniques in Separate Stages****.* Another finding we discerned is the different interaction techniques used in separate stages, especially in the NLP/ML fields. Specifically, the learning stage relies mostly on *rating* and *ranking* interaction to deal with the generated datasets of the models. However, users leverage the AI model in different ways during inference stage, as demonstrated in HCI research. This indicates that users might elicit different needs in practice when interacting with AI during the model deployment stage. Therefore, future NLP/ML and HCI researchers might collaborate to keep the human-AI interaction techniques informed (or even consistent) by both domains in order to facilitate close alignment between humans and AI.

5.3.3 ***Dissimilar Usage of Interaction Data Between Domains***. We noticed that the ways in which the two domains utilize data differ. For example, the interaction outputs of NLP research are commonly the dataset to be trained in the model, whereas HCI researchers use usage data to analyze human behavior and feedback. As AI systems continue to advance, new modes of interaction are required to capture more space of human expression, with the goal of deeper and meaningful alignment. This is an important area for NLP/ML researchers to partner with HCI researchers for designing interaction paradigms, including but certainly not limited to affective computing, accessibility, social computing, computer-supported cooperated work, and ubiquitous computing.

5.4 Gaps Between Evaluating AI Systems and Human Study Evaluation

Improving bidirectional alignment between humans and AI can potentially be further enhanced by collaborative efforts and commitments from both the NLP/ML and AI domains, ensuring consistency in evaluation methods and metrics. Nevertheless, we observed notable gaps by comparing the literature analysis regarding "evaluating AI systems" (*i.e.,* Section 4.2.3) and "evaluation in human studies" (*i.e.,* Section 4.4.3).

5.4.1 ***Different Objectives and Expectations of Evaluations in Humans and AI***. Evaluating AI systems primarily focuses on *what* algorithmic performance is, aiming to determine how well proposed methods perform compared with benchmark or human-in-the-loop evaluation. Conversely, human study evaluations are more concerned with *why* – understanding the user experience and interactions, often prioritizing both quantitative and qualitative measurements to gain a comprehensive understanding of user experience and interactions with AI systems. Given the distinct objectives in evaluation, future research has the potential to collaborate in bridging these gaps. For instance, informed by the user experience and demands revealed by HCI studies, NLP/ML researchers can incorporate the insights and design human-centered evaluation benchmarks to assess AI systems.

5.4.2 ***Contrasting Ways of Utilizing User Information***. Given the diverse objectives of assessment described above, NLP/ML research primarily focuses on quantifiable measures such as accuracy, precision, and recall. The information received in evaluating AI systems is often accurately measurable, e.g., ratings, rankings, and multiple choice questions. In comparison, by focusing on understanding the user experience and interaction with AI systems, HCI research often involves gathering both quantitative and qualitative data, such as user feedback, perceptions, and behaviors. The collected information from HCI research might cover a wider spectrum of qualitative data related to human values, experience, and needs in practice than what can be captured by quantitative metrics alone. However, incorporating qualitative findings into improving AI systems can be difficult. Therefore, future researchers can investigate methods to convert qualitative human feedback into quantitative data that AI systems can effectively utilize and align with.

5.4.3 ***Evaluation Approaches and Scale***. Prior studies in NLP/ML have predominantly relied on automated or human-in-the-loop evaluation methods to assess AI system performance across extensive datasets. In comparison, HCI studies typically employ qualitative and/or quantitative measurements to gain deeper insights into user experiences and interactions within defined human groups. These divergent aspects highlight diverse gaps between the evaluation

FUTURE CHALLENGES		A Align AI to ↦ Humans	B Align Humans to ↦ AI
Near-Term	Sec6.1 Specification Game	Sec 6.1.1 Integrate fully specified human values into aligning AI	Sec 6.1.2 Elicit the nuanced and contextual human values during diverse interaction
	Sec6.2 Dynamic Co-evolution of Alignment	Sec 6.2.1 Co-evolve AI with changes in humans and society	Sec 6.2.2 Adapt humans and society to the latest AI advancements
Long-Term	Sec6.3 Safeguarding Coadaptation	Sec 6.3.1 Decompose AI final goals into interpretable and controllable instrumental actions	Sec 6.3.2 Empower humans to identify and intervene in AI instrumental and final strategies in collaboration

Fig. 10. We envision future research directions to achieve long-term human-AI alignment with both efforts from the "Align AI to ↦ Humans" and "Align Humans to ↦ AI" directions. We elaborate the three important future challenges, including Specification Game (Section 6.1), Dynamic Co-evolution of Alignment (Section 6.2), and Safeguarding Coadaptation (Section 6.3).

methods and scales utilized in the two domains. Closing this gap and adopting consistent evaluation strategies could lead to a more holistic understanding and synergy of human-AI interaction and alignment across both fields.

Overall, it's important to note that the disparities between HCI and NLP/ML studies extend beyond the above aspects, *e.g.,* research methods and problem formation. Our analyses aim to pave the way for future interdisciplinary research collaboration, fostering bidirectional alignment between humans and AI across diverse domains.

6 FUTURE DIRECTIONS

The two directions of human-AI alignment are not independent; instead, they represent an *ongoing, mutual process* to achieve a *dynamic alignment goal over the long term.* Future research should approach alignment studies as a bidirectional process to fully explore the breadth of the research space and achieve this essential goal. Drawing upon insights gained from the development of our framework and the associated coding analysis, as shown in Figure 10, 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 (Section 6.1), Dynamic Co-evolution of Alignment (Section 6.2), and Safeguarding Coadaptation (Section 6.3). Particularly, we discuss research questions and motivations followed by ideas for solution exploration from both directions of "A Align AI to ↦ Human" (*i.e.,* domains related to AI algorithm and method development, including Machine Learning, Artificial Intelligence, Data Science, and others) and "B Align Human to ↦ AI" (*i.e.,* domains conducting studies related to humans and society, including Human-Computer Interaction, Social Science, Cognitive Science and others) for each challenge below.

6.1 Challenge 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.

6.1.1 ***Aligning AI to ↦ Humans: 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 [333, 334]. For example, people may prioritize the value of *family* and *work* differently, leading to varied choices in time allocation. Also, these priorities can change dynamically throughout an individual's life stages. As such, it is unlikely to that we can achieve a universal value system, recognized by all humans, to govern AI alignment. Instead, it is more realistic to select values compatible with specific societies or situations, given the fact that we live in a diverse world [104]. Therefore, in the absence of a universal value agreement, 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.*

Examples of Potential Future Solutions. Future work can draw from political theory, which examines fair methods for determining the principles that state institutes should follow, as a potential guide for selecting human values for AI alignment. Inspired by Social Choice Theory [9], one future direction involves aggregating individual values into collective value agreements or judgments through democratic processes such as voting, discussion, and civic engagement [104]. Therefore, building on the summary of human values in Section 4.1 and 5.2, future researchers can use democratic methods to identify a wide spectrum of human value subsets for AI alignment. Additionally, an important focus for future work lies in the curation of datasets that can represent these wide-spectrum human values, particularly in rich formats that go beyond simple ratings and rankings, which are commonly used in current alignment algorithms (*e.g.,* RLHF [282], RLAIF [17]). Another key research area to explore is around algorithms or methods (*e.g.,* Bradley–Terry Model [312] or Elo Rating System [17]), potentially drawing from economic theories, to convert heterogeneous human value data (*e.g.,* principles or contextual information) into AI-compatible formats (*e.g.,* scalars) for training reward models and guiding reinforcement learning. Also, future research can explore AI models that can act directly on aligning humans' unstructured data, such as free-form descriptions of values, multimedia or sensor recordings depicting how people act in various situations, among others.

*6.1.2 **Aligning Humans to ↦ AI: Elicit nuanced and contextual human values during diverse interactions**.*

Current alignment methods primarily capture human values using instructions, ratings and rankings, aiming to infer humans' alignment goals (*e.g.,* human values and intentions) in a data-driven manner. However, these limited data formats and interaction methods fall well short of fully specifying all of the relevant human values and constraints. Besides, when humans seek to collaborate with or utilize AI, they often encounter challenges in formulating optimal prompts, which require them to accurately and efficiently specify their requirements or expectations, in order to be able to use AI generated output or decision making in practice. Additionally, people often struggle to articulate all of their desired values accurately and comprehensively, especially as these values can dynamically change based on contextual and temporal situations. Relatedly, practices often overlook implicit human interactive signals that can also indicate values relevant to AI alignment.

Examples of Potential Future Solutions. One promising area for future research is optimizing interactive interfaces to elicit human values efficiently and effectively. These interfaces can leverage diverse interaction modes (see an initial summary in Figure 9) to fully capture human value information. For example, to identify a set of important principles for AI alignment, the interface can use the user's conversational natural language specifications and generate a list of principle options for users to select from and/or revise. Recognizing that humans cannot always comprehensively articulate their values, another key research area is to develop interfaces that proactively interact with users. This could involve conversational techniques or scaffolding procedures to elicit nuanced, contextualized, and evolving information about an individual's values. Additionally, to capture implicit human interaction signals that reflect human values for AI alignment, another area of future work involves designing systems that track interactions to develop real-time hypotheses about human values on these implicit signals, and then proactively validate them with humans as needed.

6.2 Challenge 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.

6.2.1 ***Aligning AI to ↦ Humans: Co-evolve AI with changes in humans and society.*** Existing literature often views AI alignment as a static process, neglecting the dynamic nature of alignment goals. However, a long-term perspective of alignment requires us to take into account the co-evolution of AI, humans, and society. On one hand, as AI systems scale up, they regularly acquire new and sometimes unexpected ("emergent") capabilities, including learning from examples on the fly and adaptively pursuing goals. This leads to the challenge of ensuring that the goals AI might independently formulate remain consistently aligned with human values. Consequently, alignment solutions must involve continuous oversight and updates in response to AI's evolution during development and deployment. On the other hand, AI advancements influence the state of the world, affecting human actions and even their underlying values and strategies [143]. Thus, alignment solutions that continually adapt to the updated human and societal considerations may offer the most robust approach [161]. Consequently, it is crucial to *ensure that AI co-evolves with changes in human and societal values, cognition, and needs dynamically.*

Examples of Potential Future Solutions. Updating AI to continuously adapt to the dynamic changes of humans and society requires adjustments in alignment data, models, and evaluation processes spanning both learning and deployment stages. Referring to the methodologies listed in Section 4.2, current strategies often rely on fine-tuning to update AI with a curated human value dataset during the learning process. However, this approach may lead to AI models forgetting or diminishing previous alignment goals. In situations where sufficient data for fine-tuning is unavailable, these methods becomes more challenging. Therefore, a crucial area for future research is to develop approaches for continually updating AI systems using limited data without compromising existing alignment values and performance. In real-world scenarios, it is common for AI systems to already be deployed at scale, making fine-tuning a challenging or impractical practice. Therefore, another important research direction involves adapting AI, during the deployment stage, without relying on fine-tuning. 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. In summary, the co-evolution of AI with dynamic human and societal changes necessitates a synergetic collaboration among experts from various domains, such as machine learning and social sciences.

6.2.2 ***Aligning Humans to ↦ AI: Adapt humans and society to the latest AI advancements.*** As AI progresses, humans will likely need to transition from interacting with less capable systems to engaging advanced AI systems (*e.g.,* AGI) [271]. It's imperative to develop interactive strategies which enable humans to effectively utilize AI across varying levels of model capability. While current AI systems still lag behind human performance in many tasks, questions persist about human resilience to AI mistakes. Efficiently identifying these mistakes and developing appropriate fallback actions, such as interactive intervention, remains crucial in various situations; identifying and handling errors, including knowing when to seek human intervention, will remain an important even for highly capable future systems. Looking ahead, as AI advances further, it becomes essential to develop interactive systems which enable humans to utilize AI that has capabilities surpassing their own. Research is needed to understand how individuals can interpret and validate the outputs of AI performing tasks beyond their current abilities, and also, how humans could leverage advanced AI in a sustainable way, *i.e.,* benefiting their life without harmful side effects such as job displacement or loss of purpose. In addition, as AI becomes increasingly integrated into daily tasks, its influence on human values, behaviors, capabilities, and society remains uncertain. Thus, it's vital to continuously examine the impact of AI advancements on individuals, social relationships, and broader societal changes.

Examples of Potential Future Solutions. While prior AI assistants have primarily aimed to complement human capabilities in task performance, future research on advanced AI systems should focus on developing validation mechanisms that enable humans to interpret and verify AI outputs. Inspired by initial explorations in Section 4.4.1, this could involve designing interactive interfaces that allow humans to request step-by-step justifications from AI, or integrating external tools to verify the truthfulness of AI outputs. This could also include developing interfaces that enable groups of humans to work together to validate AI outputs, since groups working jointly typically have broader and deeper "intelligence" in practice than any single actor). Further, this approach can be extended to scalable validation tools, automating the verification process for large-scale applications such as in education and the workplace. Another important research direction is designing interactive strategies to enhance human capabilities by learning from advanced AI. This includes gaining knowledge and building technical and social skills. Furthermore, it is crucial to assess how humans and society adapt to AI advancements which, in turn, can guide the future evolution of AI. Potential research areas include evaluating changes in individual behavior and social relationships (*e.g.,* how relationships with advanced AI may supplant human relationships), and societal governance (*e.g.,* how legal and education systems may change as AI is used in place of traditional human skills). Examining these dynamic changes is essential for understanding the broader impact of AI advancements on humanity and society.

6.3 Challenge 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, i.e., power-seeking or evading human intervention. Therefore, safeguarding the co-adaptation between humans and AI is crucial. This can be achieved by empowering humans to understand and control AI's instrumental goals, in order to ensure rational and ethical behaviors. We next explore future research to address this challenge from both directions.

6.3.1 ***Aligning AI to ↦ Humans: Specify the goals of an AI system into interpretable and controllable instrumental actions for humans.*** As advanced AI systems become increasingly complex, they present greater challenges for human interpretation and control. To address this, it is crucial to *empower humans to detect and interpret AI misconduct on instrumental actions towards accomplishing its final goals*. Additionally, mechanisms must be developed to enable human intervention and prevent power-seeking AI misconduct, thus maintaining a safe co-adaptation between AI and humans. Furthermore, advanced AI systems, even with sufficient interpretation capabilities, may intentionally mislead or disobey humans. It is possible for AI to stand by falsehoods, generate empty explanations for their answers, and produce outright fabrications that may appear plausible [173]. Therefore, *developing reliable interpretability mechanisms to validate the faithfulness and honesty of AI systems' self-reported behaviors is also essential.*

Examples of Potential Future Solutions. One potential research area aimed at empowering humans to interpret and control advanced AI behavior is the *design of corrigible mechanisms that facilitate easy intervention and correction.* This includes developing modular AI architectures where components can be independently adjusted or replaced, and creating robust override protocols that allow human operators to safely halt or redirect AI activities. These components should ideally be human-interpretable, enabling supervisors to conduct scenario testing when necessary. Another critical research direction lies in validating the faithfulness and honesty of AI interpretability. This could involve correlating AI behaviors with internal neuron activity signals, similar to how physiological indicators are measured in human polygraph tests [14]. Inspecting internal neuron indicators could support human assessment of the truthfulness of AI interpretations, thereby allowing them to accurately judge and prevent potential risky instrumental actions.

6.3.2 ***Aligning Humans to ↦ AI: Empower humans to identify and intervene in AI instrumental and final strategies in collaboration.***

Preventing future advanced AI systems from engaging in risky instrumental actions requires human capabilities to identify and mitigate such behaviors. Additionally, scalable solutions that can be widely implemented are essential for effectively supervising AI instrumental behavior across various applications and environments. Future research could potentially explore comprehensive training programs and the development of advanced diagnostic tools to achieve these goals.

Examples of Potential Future Solutions. To enhance human capability in supervising AI instrumental actions, robust training and simulation environments are essential. These should offer scenario-based training with timely feedback loops, enabling supervisors to identify and manage various instrumental actions effectively. Additionally, creating interactive and intuitive dashboards is crucial for quickly recognizing and responding to potential risky instrumental goals. These dashboards should include effective data and model visualization, intervention techniques to highlight anomalies, and real-time alerts for prompt intervention. Furthermore, as deploying advanced AI at scale makes real-time oversight more challenging, developing advanced and autonomous monitoring tools is vital. These tools should learn the normal operating parameters of AI systems, flagging deviations that suggest risky instrumental actions immediately. By integrating training and simulation environments, interactive dashboards, and scalable diagnostic tools, humans can better manage AI instrumental risks, supporting better alignment with human values and control.

7 IMPLICATIONS AND LIMITATIONS

In this section, we provide a summary of the key implications of our proposed framework, and discuss the challenges and limitations we encountered during the systematic literature review process.

7.1 Framework Implications and Applications

7.1.1 ***Cross-domain inspirations.***

Achieving long-term human-AI alignment requires holistic, interdisciplinary efforts. This paper introduces a bidirectional human-AI alignment framework to create a shared vocabulary among diverse fields. Supported by over 400 multi-domain research papers, this framework **enables researchers to integrate knowledge from other domains, enhancing their work and fostering synergetic collaboration.** Particularly, AI researchers can use the human value taxonomy grounded in psychological theories (*e.g.,* Section 5.2) to develop datasets and algorithms that incorporate a broader spectrum of human values. Additionally, by examining studies on individual and societal adaptations to AI alignment (*e.g.,* Section 4.4.2), AI developers can create more practical and adaptive AI systems that evolve alongside human and societal changes. Conversely, HCI and social science researchers can leverage customizable AI alignment algorithms (*e.g.,* Section 4.2) to design interactive systems that improve human interaction with AI. By using interaction techniques to specify human values (Section 4.1), these researchers can optimize interfaces or conduct surveys to extract human values, informing AI alignment development.

7.1.2 ***Foundation of Dynamic Alignment Evolution.***

The proposed bidirectional human-AI alignment framework also **serves a foundational role to support researchers to address the outlined future challenges and achieve long-term alignment**. As elaborated in Section 6, within each direction, researchers can potentially address future challenges on the basis of existing alignment studies organized by the framework. For instance, to tackle the specification game, future AI researchers can use the comprehensive human value taxonomy outlined in the framework (Section 6) to develop methods for selecting the appropriate subset of these human values to be integrated into AI alignment. Additionally, we emphasize the importance of synergizing both directions to achieve long-term bidirectional human-AI alignment. To facilitate dynamic co-evolution (Section 6.2), we advocate for collaboration among AI, HCI, and social scientists to measure human and societal adaptations and incorporate these changes into the alignment of next-generation AI systems. This holistic approach can contribute to the evolution of AI in harmony with human values and societal needs.

7.2 Limitations

One limitation of this work is the coverage of the sampled and filtered papers. The literature revolving around the topic of human-AI alignment is increasing rapidly, scattering in diverse venues across many domains. Therefore, instead of striving for an exhaustive collection of papers, this work focused on developing a holistic loop of bidirectional human-AI alignment framework by using essential research questions, dimensions and codes. Additionally, we are aware that our surveyed papers and the position of our team members primarily focus on computing-related domains, such as ML, NLP, and HCI. There are additional disciplines involved in alignment research, such as cognitive science, psychology, and STS (Science, Technology, and Society). However, our framework can naturally extend to these domains by incorporating them into the appropriate directions. Despite these limitations, we argue that this bidirectional human-AI alignment framework serves as foundational reference forfuture researchers examining existing literature that lies in the holistic loop of the alignment process, while preventing implicit assumptions or overlooked considerations, thereby facilitating a holistic understanding of the dimensions and factors that drive the developing of AI technology and human impacts. For future work, we plan to continue to refine our dimensions and codes to incorporate future literature relevant to the bidirectional human-AI alignment process.

8 CONCLUSION

In conclusion, this study clarifies the definitions and scope of core terminologies of human-AI alignment and conducts a systematic review of over 400 related papers spanning diverse domains such as NLP, AI, HCI, and social science. Additionally, we introduce a novel conceptual framework of "Bidirectional Human-AI Alignment", structuring the surveyed literature taxonomies into "aligning AI to humans" and "aligning humans to AI" with detailed categories and example papers. Furthermore, we identify limitations and risks in this area quantitatively and qualitatively, analyzing a fine-grained human value taxonomy, interaction modes for alignment, and discrepancies between AI and human evaluation. To pave the way for future studies, we discuss five stages to achieve the alignment goals from near-term to long-term perspectives and identify new possibilities to highlight future directions and opportunities in research.

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.

APPENDIX

A 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 [279] and listed contributors for each part of the paper below.

A.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 (University of Michigan, huashen@umich.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 (Google, tknearem@google.com)**: 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 (Carnegie Mellon University, 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.

B SYSTEMATIC LITERATURE REVIEW

B.1 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 4.4.2 and the exemplary paper of Regulating ChatGPT [130]), 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.

B.2 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

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

B.3 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.

B.4 Collected Main Papers

- **HCI**: [4, 11, 12, 19–21, 23, 25, 30, 34, 37–42, 47, 52, 53, 57–59, 63, 65, 66, 68, 69, 73–75, 79, 81, 90, 98, 102, 105, 109, 112, 115, 124, 128, 132, 135–137, 139, 142, 144, 145, 148, 156, 163–165, 169, 170, 172, 175, 179, 181, 185, 192, 193, 198, 204–206, 211–213, 215, 217, 226, 228, 229, 231, 232, 234, 241, 244, 250, 253, 257, 260, 262, 264, 267, 269, 270, 274, 290, 291, 300, 302–305, 313, 317–319, 325, 329, 330, 340, 352, 355–358, 361, 363, 371, 374, 384, 390–393, 400–402, 406–408, 413–415, 422, 423, 425, 430, 431, 433, 436–438, 445, 445, 451, 452, 461, 465, 468, 471, 472, 474, 475, 479]
- **NLP/AI**
[2, 5, 8, 18, 22, 28, 29, 31, 35, 43, 44, 46, 48, 49, 51, 54, 55, 60, 64, 67, 71, 72, 77, 80, 82–88, 91–97, 99, 100, 107, 108, 110, 111, 113, 114, 117–119, 121, 125–127, 129, 134, 138, 143, 146, 147, 150, 152–154, 158, 159, 162, 166–168, 171, 174, 176–178, 183, 186–191, 196, 197, 202, 203, 207, 214, 216, 218–224, 227, 230, 235–240, 242, 245–247, 249, 251, 252, 255, 258, 259, 261, 263, 266, 268, 273, 275–278, 280–284, 286, 292–294, 296, 299, 306, 309, 310, 312, 314–316, 320, 321, 327, 328, 332, 336, 337, 339, 342, 344–346, 349–351, 353, 354, 359, 365, 367–369, 372, 373, 375–377, 379–382, 385, 386, 394, 395, 397–399, 403, 404, 409, 411, 412, 416, 418, 419, 426, 427, 434, 439–441, 443, 444, 446–450, 453, 454, 457–459, 462–464, 466, 467, 469, 470, 473, 477, 478, 480, 482–484]
- **Others** [3, 10, 13, 15–17, 50, 56, 106, 130, 155, 180, 182, 194, 201, 225, 233, 256, 272, 297, 298, 301, 326, 331, 338, 341, 343, 347, 360, 370, 378, 417, 420, 455, 460, 476, 481, 485]

REFERENCES

[1] 2023. Humans are biased. Generative AI is even worse. https://www.bloomberg.com/graphics/2023-generative-ai-bias/

[2] Afra Feyza Akyürek, Ekin Akyürek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. 2023. 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)*. 7716–7733.

[3] Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models. *arXiv preprint arXiv:2402.13231* (2024).

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

[5] Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. 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*. 5994–6017.

[6] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[7] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. *arXiv preprint arXiv:2404.09932* (2024).

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

[9] Kenneth J Arrow. 2012. *Social choice and individual values*. Vol. 12. Yale university press.

[10] Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. 2024. How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas: Evidence From a Large, Dynamic Experiment. *arXiv preprint arXiv:2401.13481* (2024).

[11] Zahra Ashktorab, Benjamin Hoover, Mayank Agarwal, Casey Dugan, Werner Geyer, Hao Bang Yang, and Mikhail Yurochkin. 2023. 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*. 1–20.

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

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

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

[15] Shubham Atreja, Libby Hemphill, and Paul Resnick. 2023. 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 (2023), 1–33.

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

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

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

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

[20] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. 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 (2023), 1–17.

[21] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. 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*. 1–16.

[22] Hritik Bansal, John Dang, and Aditya Grover. 2023. Peering Through Preferences: Unraveling Feedback Acquisition for Aligning Large Language Models. (2023).

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

[24] 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. 2024. Managing extreme AI risks amid rapid progress. *Science* (2024), eadn0117.

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

[26] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic Interpretability for AI Safety – A Review. *arXiv preprint arXiv:2404.14082* (2024).

[27] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 313–322.

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

[29] Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. 2020. Interactive Weak Supervision: Learning Useful Heuristics for Data Labeling. In *International Conference on Learning Representations*.

[30] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. 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*. 1–17.

[31] Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. 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*. 8031–8049.

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

[33] Nick Bostrom. 2020. Ethical issues in advanced artificial intelligence. *Machine Ethics and Robot Ethics* (2020), 69–75.

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

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

[36] Richard Brath. 2021. Surveying Wonderland for many more literature visualization techniques. *arXiv preprint arXiv:2110.08584* (2021).

[37] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. 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 (2021), 1–21.

[38] Federico Cabitza, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi. 2023. 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*. 1–20.

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

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

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

[42] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " 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 (2019), 1–24.

[43] Xin-Qiang Cai, Yu-Jie Zhang, Chao-Kai Chiang, and Masashi Sugiyama. 2023. Imitation Learning from Vague Feedback. *Advances in Neural Information Processing Systems* 36.

[44] Chengzhi Cao, Yinghao Fu, Sheng Xu, Ruimao Zhang, and Shuang Li. 2024. Enhancing Human-AI Collaboration Through Logic-Guided Reasoning. In *The Twelfth International Conference on Learning Representations*.

[45] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. AI Alignment with Changing and Influenceable Reward Functions. *arXiv preprint arXiv:2405.17713* (2024).

[46] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems* 32.

[47] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 251–263.

[48] Robin Chan, Afra Amini, and Mennatallah El-Assady. 2023. 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)*. 463–470.

[49] Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and Jonathan Gratch. 2023. 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*. 13078–13092.

[50] Angelica Chen, Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Samuel R. Bowman, Kyunghyun Cho, and Ethan Perez. 2024. Learning from Natural Language Feedback. *Transactions on Machine Learning Research* (2024).

[51] Jiaao Chen, Mohan Dodda, and Diyi Yang. 2023. Human-in-the-loop Abstractive Dialogue Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*. 9176–9190.

[52] Quan Ze Chen, Tobias Schnabel, Besmira Nushi, and Saleema Amershi. 2022. HINT: Integration Testing for AI-based features with Humans in the Loop. In *27th International Conference on Intelligent User Interfaces*. 549–565.

[53] Weihao Chen, Chun Yu, Huadong Wang, Zheng Wang, Lichen Yang, Yukun Wang, Weinan Shi, and Yuanchun Shi. 2023. 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*. 1–15.

[54] Yu Chen, Yihan Du, Pihe Hu, Siwei Wang, Desheng Wu, and Longbo Huang. 2023. Provably Efficient Iterated CVaR Reinforcement Learning with Function Approximation and Human Feedback. In *The Twelfth International Conference on Learning Representations*.

[55] Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel Tetreault, and Alejandro Jaimes-Larrarte. 2022. 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*. 431–455.

[56] Xusen Cheng, Xiaoping Zhang, Jason Cohen, and Jian Mou. 2022. 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 (2022), 102940.

[57] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. 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*. 1–18.

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

[59] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. 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*. 1–12.

[60] Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. 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*. 11370–11403.

[61] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[62] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

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

[64] John Chung, Ece Kamar, and Saleema Amershi. 2023. 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)*. 575–593.

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

[66] Gabriele Cimolino and TC Nicholas Graham. 2022. 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*. 1–21.

[67] Nina Corvelo Benz and Manuel Rodriguez. 2023. Human-aligned calibration for ai-assisted decision making. *Advances in Neural Information Processing Systems* 36.

[68] Shanley Corvite, Kat Roemmich, Tillie Ilana Rosenberg, and Nazanin Andalibi. 2023. 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 (2023), 1–38.

[69] Kathryn Cunningham, Barbara J Ericson, Rahul Agrawal Bejarano, and Mark Guzdial. 2021. 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*. 1–15.

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

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

[72] Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. 2022. 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*. 9465–9480.

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

[74] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. 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*. 1–13.

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

[76] Kerstin Dautenhahn, Chrystopher L Nehaniv, and K Dautenhahn. 2000. Living with Socially Intelligent Agents. *Human Cognition and Social Agent Technology, John Benjamins Publ. Co* (2000), 415–426.

[77] Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You Are What You Annotate: Towards Better Models through Annotator Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 12475–12498.

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

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

[80] Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A Diachronic Perspective on User Trust in AI under Uncertainty. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 5567–5580.

[81] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. 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*. 3321–3339.

[82] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. 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*. 3321–3339.

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

[84] Kefan Dong and Tengyu Ma. 2023. Asymptotic Instance-Optimal Algorithms for Interactive Decision Making. In *The Eleventh International Conference on Learning Representations*.

[85] Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 11275–11288.

[86] Zibin Dong, Yifu Yuan, Jianye HAO, Fei Ni, Yao Mu, YAN ZHENG, Yujing Hu, Tangjie Lv, Changjie Fan, and Zhipeng Hu. 2024. AlignDiff: Aligning Diverse Human Preferences via Behavior-Customisable Diffusion Model. In *The Twelfth International Conference on Learning Representations*.

[87] Florian E Dorner, Momchil Peychev, Nikola Konstantinov, Naman Goel, Elliott Ash, and Martin Vechev. 2023. Human-Guided Fair Classification for Natural Language Processing. In *The Eleventh International Conference on Learning Representations*.

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

[89] W Keith Edwards and Rebecca E Grinter. 2001. At home with ubiquitous computing: Seven challenges. In *Ubicomp 2001: Ubiquitous Computing: International Conference Atlanta Georgia, USA, September 30–October 2, 2001 Proceedings 3*. Springer, 256–272.

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

[91] David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Michael Smith. 2023. ROBBIE: Robust Bias Evaluation of Large Generative Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[92] Kawin Ethayarajh and Dan Jurafsky. 2022. The Authenticity Gap in Human Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 6056–6070.

[93] Xiang Fan, Yiwei Lyu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023. Nano: Nested Human-in-the-Loop Reward Learning for Few-shot Language Model Control. In *Findings of the Association for Computational Linguistics: ACL 2023*. 11970–11992.

[94] Zahra Fatemi, Chen Xing, Wenhao Liu, and Caimming Xiong. 2023. Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

[95] Amir Feder, Guy Horowitz, Yoav Wald, Roi Reichart, and Nir Rosenfeld. 2022. In the eye of the beholder: Robust prediction with causal user modeling. *Advances in Neural Information Processing Systems* 35, 14419–14433.

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

[97] Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. 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)*. 9126–9140.

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

[99] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. 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)*. 11737–11762.

[100] Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. 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)*. 6231–6251.

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

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

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

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

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

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

[107] Ge Gao, Hung-Ting Chen, Yoav Artzi, and Eunsol Choi. 2023. Continually Improving Extractive QA via Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 406–423.

[108] Ge Gao, Eunsol Choi, and Yoav Artzi. 2022. 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)*. 5167–5179.

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

[110] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and William B Dolan. 2020. Dialogue Response Ranking Training with Large-Scale Human Feedback Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 386–395.

[111] Yiming Gao, Feiyu Liu, Liang Wang, Dehua Zheng, Zhenjie Lian, Weixuan Wang, Wenjin Yang, Siqin Li, Xianliang Wang, Wenhui Chen, et al. 2024. Enhancing Human Experience in Human-Agent Collaboration: A Human-Centered Modeling Approach Based on Positive Human Gain. (2024).

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

[113] Matthias Gerstgrasser, Rakshit Trivedi, and David C Parkes. 2021. CrowdPlay: Crowdsourcing Human Demonstrations for Offline Learning. In *International Conference on Learning Representations*.

[114] Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022. 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*. 12–21.

[115] Zohar Gilad, Ofra Amir, and Liat Levontin. 2021. 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*. 1–13.

[116] Eric Gilbert. 2024. HCC Is All You Need: Alignment-The Sensible Kind Anyway-Is Just Human-Centered Computing. *arXiv preprint arXiv:2405.03699* (2024).

[117] Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, and Marc Dymetman. 2023. Compositional Preference Models for Aligning LMs. In *The Twelfth International Conference on Learning Representations*.

[118] Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. 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)*. 89–97.

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

[120] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[121] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *The Twelfth International Conference on Learning Representations*.

[122] Nitesh Goyal, Minsuk Chang, and Michael Terry. 2024. 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*. 1–6.

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

[124] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K Kummerfeld, and Elena L Glassman. 2024. An AI-Resilient Text Rendering Technique for Reading and Skimming Documents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

[125] Lin Guan, Karthik Valmeekam, and Subbarao Kambhampati. 2022. Relative Behavioral Attributes: Filling the Gap between Symbolic Goal Specification and Reward Learning from Human Preferences. In *The Eleventh International Conference on Learning Representations*.

[126] Lin Guan, Mudit Verma, Suna Sihang Guo, Ruohan Zhang, and Subbarao Kambhampati. 2021. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. *Advances in Neural Information Processing Systems* 34, 21885–21897.

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

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

[129] Prakhar Gupta, Yang Liu, Di Jin, Behnam Hedayatnia, Spandana Gella, Sijia Liu, Patrick L Lange, Julia Hirschberg, and Dilek Hakkani-Tur. 2023. DialGuide: Aligning Dialogue Model Behavior with Developer Guidelines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 14031–14047.

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

[131] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. 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*. 2125–2126.

[132] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. 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*. 1–19.

[133] Andy Hamilton. 2020. Conservatism. In *The Stanford Encyclopedia of Philosophy* (Spring 2020 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[134] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5540–5552.

[135] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. 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 (2023), 1–29.

[136] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. 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*. 1–18.

[137] Ziyao He, Yunpeng Song, Shurui Zhou, and Zhongmin Cai. 2023. 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*. 1–18.

[138] Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. 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*. 11138–11154.

[139] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. 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*. 453–463.

[140] Thomas A Hemphill. 2020. Human Compatible: Artificial Intelligence and the Problem of Control.

[141] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch Critch, Jerry Li Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. In *International Conference on Learning Representations*.

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

[143] Joey Hong, Sergey Levine, and Anca Dragan. 2023. Learning to influence human behavior with offline reinforcement learning. *Advances in Neural Information Processing Systems* 36.

[144] Matt-Heun Hong, Lauren A Marsh, Jessica L Feuston, Janet Ruppert, Jed R Brubaker, and Danielle Albers Szafir. 2022. 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*. 1–12.

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

[146] WANG Hongru, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023. Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogues. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[147] Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human Feedback is not Gold Standard. In *The Twelfth International Conference on Learning Representations*.

[148] Yoyo Tsung-Yu Hou, Wen-Ying Lee, and Malte Jung. 2023. "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*. 1–13.

[149] Dirk Hovy and Diyi Yang. 2021. 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*. 588–602.

[150] Shengran Hu and Jeff Clune. 2023. Thought cloning: Learning to think while acting by imitating human thinking. *Advances in Neural Information Processing Systems* 36.

[151] Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2024. Do Large Language Models Know about Facts?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=9OevMUdods

[152] Xiao Hu, Jianxiong Li, Xianyuan Zhan, Qing-Shan Jia, and Ya-Qin Zhang. 2024. Query-Policy Misalignment in Preference-Based Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

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

[154] Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2023. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*.

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

[156] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen MacNeil. 2023. 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*. 1–3.

[157] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–17.

[158] EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning Language Models to User Opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 5906–5919.

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

[160] ASTM International. 2012. Form and Style of Standards, ASTM Blue Book. https://www.astm.org/media/pdf/bluebook_FormStyle.pdf.

[161] Geoffrey Irving and Amanda Askell. 2019. AI safety needs social scientists. *Distill* 4, 2 (2019), e14.

[162] Arman Isajanyan, Artur Shatveryan, David Kocharian, Zhangyang Wang, and Humphrey Shi. 2024. Social Reward: Evaluating and Enhancing Generative AI through Million-User Feedback from an Online Creative Community. In *The Twelfth International Conference on Learning Representations*.

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

[164] 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. 2021. 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*. 1–14.

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

[166] Sullam Jeoung, Yubin Ge, and Jana Diesner. 2023. 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*. 12236–12256.

[167] Xiaogang Jia, Denis Blessing, Xinkai Jiang, Moritz Reuss, Atalay Donat, Rudolf Lioutikov, and Gerhard Neumann. 2024. Towards Diverse Behaviors: A Benchmark for Imitation Learning with Human Demonstrations. In *The Twelfth International Conference on Learning Representations*.

[168] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems* 36.

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

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

[171] Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems* 35, 28458–28473.

[172] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. 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*. 1–16.

[173] Steven Johnson and Nikita Iziev. 2022. AI is mastering language. Should we trust what it says? *The New York Times* 4 (2022), 15.

[174] Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems* 35, 11785–11799.

[175] Matthew Jörke, Yasaman S Sefidgar, Talie Massachi, Jina Suh, and Gonzalo Ramos. 2023. Pearl: A Technology Probe for Machine-Assisted Reflection on Personal Data. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 902–918.

[176] Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. 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)*. 7103–7128.

[177] Bong Gyun Kang, HyunGi Kim, Dahuin Jung, and Sungroh Yoon. 2023. CLeAR: Continual Learning on Algorithmic Reasoning for Human-like Intelligence. *Advances in Neural Information Processing Systems* 36.

[178] Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. 2023. 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*. 15539–15559.

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

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

[181] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. 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*. 1–18.

[182] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara Jane Ericson, David Weintrop, and Tovi Grossman. 2023. 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*. 1–12.

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

[184] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659* (2021).

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

[186] Fereshte Khani and Marco Tulio Ribeiro. 2023. Collaborative Alignment of NLP models. *Advances in Neural Information Processing Systems* 36.

[187] Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2022. GENIE: Toward Reproducible and Standardized Human Evaluation for Text Generation. (2022), 11444–11458.

[188] Tushar Khot, Kyle Richardson, Daniel Khashabi, and Ashish Sabharwal. 2022. Hey AI, Can You Solve Complex Tasks by Talking to Agents?. In *Findings of the Association for Computational Linguistics: ACL 2022*. 1808–1823.

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

[190] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. 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*. 14397–14413.

[191] Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. 2023. Aligning Large Language Models through Synthetic Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 13677–13700.

[192] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. " 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*. 1–17.

[193] Taenyun Kim, Maria D Molina, Minjin Rheu, Emily S Zhan, and Wei Peng. 2023. 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*. 1–20.

[194] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. 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*.

[195] Jan H Kirchner, Logan Smith, Jacques Thibodeau, Kyle McDonell, and Laria Reynolds. 2022. Researching alignment research: Unsupervised analysis. *arXiv preprint arXiv:2206.02841* (2022).

[196] Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Röttger, and Scott Hale. 2023. 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*. 2409–2430.

[197] Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. 2024. Motif: Intrinsic Motivation from Artificial Intelligence Feedback. In *The Twelfth International Conference on Learning Representations*.

[198] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. 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*. 1–14.

[199] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012* (2023).

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

[201] Bum Chul Kwon and Nandana Mihindukulasooriya. 2023. 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)*, Danushka Bollegala, Ruihong Huang, and Alan Ritter (Eds.). Association for Computational Linguistics, Toronto, Canada, 42–50. https://doi.org/10.18653/v1/2023.acl-demo.4

[202] Bum Chul Kwon and Nandana Mihindukulasooriya. 2023. 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)*. 42–50.

[203] Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, et al. 2023. 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*. 10383–10405.

[204] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. 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*. 1–18.

[205] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

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

[207] Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, et al. 2023. 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)*. 264–273.

[208] 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. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

[209] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[210] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2023. Evaluating Human-Language Model Interaction. *Transactions on Machine Learning Research* (2023).

[211] Min Hun Lee and Chong Jun Chew. 2023. 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 (2023), 1–22.

[212] Min Kyung Lee and Katherine Rich. 2021. 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*. 1–14.

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

[214] Noah Lee, Na Min An, and James Thorne. 2023. Can Large Language Models Capture Dissenting Human Voices?. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[215] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. 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*. 1–22.

[216] Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. 2020. FIND: Human-in-the-Loop Debugging Deep Text Classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 332–348.

[217] Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. 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*. 1–17.

[218] Chenliang Li, He Chen, Ming Yan, Weizhou Shen, Haiyang Xu, Zhikai Wu, Zhicheng Zhang, Wenmeng Zhou, Yingda Chen, Chen Cheng, et al. 2023. 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*. 566–578.

[219] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36.

[220] Huao Li, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 180–192.

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

[222] Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and Hua Wu. 2024. Tool-Augmented Reward Modeling. In *The Twelfth International Conference on Learning Representations*.

[223] Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. 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*. 1487–1505.

[224] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. 2022. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems* 35, 31199–31212.

[225] Tianshi Li, Sauvik Das, Hao-Ping Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. 2024. Human-Centered Privacy Research in the Age of Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

[226] Toby Jia-Jun Li. 2020. 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*. 162–168.

[227] Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Chi Kit Cheung, and Siva Reddy. 2022. 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*. 926–937.

[228] Mengqi Liao and S Shyam Sundar. 2021. 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*. 1–14.

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

[230] Jungwoo Lim, Myunghoon Kang, Jinsung Kim, Jeongwook Kim, Yuna Hur, and Heui-Seok Lim. 2023. Beyond Candidates: Adaptive Dialogue Agent Utilizing Persona and Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 7950–7963.

[231] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. 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*. 1–17.

[232] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2023. 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*. 1–26.

[233] Gabriel Lima, Nina Grgic-Hlaca, Jin Keun Jeong, and Meeyoung Cha. 2023. 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)*. Association for Computing Machinery, New York, NY, USA, 236–246. https://doi.org/10.1145/3593013.3593992

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

[235] Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. 2023. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems* 36.

[236] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. Urial: Aligning Untuned LLMs with Just the'Write'Amount of In-Context Learning. In *The Twelfth International Conference on Learning Representations*.

[237] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. In *The Twelfth International Conference on Learning Representations*.

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

[239] Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 4694–4702.

[240] Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 6826–6847.

[241] Han Liu, Vivian Lai, and Chenhao Tan. 2021. 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 (2021), 1–45.

[242] Han Liu, Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan. 2023. Learning Human-Compatible Representations for Case-Based Decision Support. In *The Eleventh International Conference on Learning Representations*.

[243] June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. In *The First Workshop on Personalized Generative AI @ CIKM*.

[244] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D Gordon. 2023. "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*. 1–31.

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

[246] Shicheng Liu and Minghui Zhu. 2023. Learning Multi-agent Behaviors from Distributed and Streaming Demonstrations. *Advances in Neural Information Processing Systems* 36.

[247] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.

[248] Xin Liu, Muhammad Khalifa, and Lu Wang. 2024. Lightweight Language Model Calibration for Open-ended Question Answering with Varied Answer Lengths. In *The Twelfth International Conference on Learning Representations*.

[249] Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The Iron (ic) Melting Pot: Reviewing Human Evaluation in Humour, Irony and Sarcasm Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 6676–6689.

[250] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.

[251] Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2023. 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)*. 4060–4078.

[252] Jinghui Lu, Linyi Yang, Brian Namee, and Yue Zhang. 2022. 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)*. 6986–6996.

[253] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. 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*. 1–18.

[254] Laura Lucaj, Patrick van der Smagt, and Djalel Benbouzid. 2023. Ai regulation is (not) all you need. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1267–1279.

[255] Chenyang Lyu, Linyi Yang, Yue Zhang, Yvette Graham, and Jennifer Foster. 2023. Exploiting Rich Textual User-Product Context for Improving Personalized Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 1419–1429.

[256] Qianou Ma, Hua Shen, Kenneth Koedinger, and Tongshuang Wu. 2024. 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).

[257] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. 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*. 1–19.

[258] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Eureka: Human-Level Reward Design via Coding Large Language Models. In *The Twelfth International Conference on Learning Representations*.

[259] Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 1011–1031.

[260] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. 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*. 1–14.

[261] Jessica Maghakian, Paul Mineiro, Kishan Panaganti, Mark Rucker, Akanksha Saran, and Cheng Tan. 2023. Personalized Reward Learning with Interaction-Grounded Learning (IGL). In *The Eleventh International Conference on Learning Representations*.

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

[263] Bodhisattwa Majumder, Zexue He, and Julian McAuley. 2023. InterFair: Debiasing with Natural Language Feedback for Fair Interpretable Predictions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 9466–9471.

[264] Nora McDonald and Shimei Pan. 2020. 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 (2020), 1–19.

[265] Sam Whitman McGrath, Jacob Russin, Ellie Pavlick, and Roman Feiman. 2023. How Can Deep Neural Networks Inform Theory in Psychological Science? (2023).

[266] Nick Mckenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2758–2774.

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

[268] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *The Twelfth International Conference on Learning Representations*.

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

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

[271] Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024. Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv:2311.02462 [cs.AI]

[272] Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, Manish Nagireddy, Prasanna Sattigeri, Ameet Talwalkar, and David Sontag. 2024. The RealHumanEval: Evaluating Large Language Models' Abilities to Support Programmers. *arXiv preprint arXiv:2404.02806* (2024).

[273] Hussein Mozannar, Jimin Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. 2023. Effective Human-AI Teams via Learned Natural Language Rules and Onboarding. *Advances in Neural Information Processing Systems* 36.

[274] Imani Munyaka, Zahra Ashktorab, Casey Dugan, James Johnson, and Qian Pan. 2023. Decision Making Strategies and Team Efficacy in Human-AI Teams. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–24.

[275] Vinod Muthusamy, Yara Rizk, Kiran Kate, Praveen Venkateswaran, Vatche Isahagian, Ashu Gulati, and Parijat Dube. 2023. 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*.

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

[277] Duy-Hung Nguyen, Nguyen Viet Dung Nghiem, Bao-Sinh Nguyen, Dung Tien Tien Le, Shahab Sabahi, Minh-Tien Nguyen, and Hung Le. 2022. 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*. 1919–1930.

[278] Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. 2023. MoCa: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks. *Advances in Neural Information Processing Systems* 36 (2023).

[279] The ACM Director of Publications. 2024. ACM Policy on Authorship. https://www.acm.org/publications/policies/new-acm-policy-on-authorship

[280] Minsik Oh, Joosung Lee, Jiwei Li, and Guoyin Wang. 2023. PK-ICR: Persona-Knowledge Interactive Multi-Context Retrieval for Grounded Dialogue. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 16383–16395.

[281] Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. 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)*. 1017–1029.

[282] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[283] Siqi Ouyang and Lei Li. 2023. AutoPlan: Automatic Planning of Interactive Decision-Making Tasks With Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 3114–3128.

[284] Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. 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*. 2375–2393.

[285] 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. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj* 372 (2021).

[286] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The utility of explainable ai in ad hoc human-machine teaming. *Advances in neural information processing systems* 34, 610–623.

[287] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. (2022).

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

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

[290] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2021. 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*. 1–15.

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

[292] Sunghyun Park, Han Li, Ameen Patel, Sidharth Mudgal, Sungjin Lee, Young-Bum Kim, Spyros Matsoukas, and Ruhi Sarikaya. 2021. 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*. 6054–6063.

[293] Roma Patel and Ellie Pavlick. 2021. "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*. 10080–10095.

[294] Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 327–337.

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

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

[297] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. 2023. Do users write more insecure code with AI assistants? (2023), 2785–2799.

[298] Markus Peschl, Arkady Zgonnikov, Frans A Oliehoek, and Luciano C Siebert. 2022. 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*. 1038–1046.

[299] Dominic Petrak, Nafise Sadat Moosavi, Ye Tian, Nikolai Rozanov, and Iryna Gurevych. 2024. Learning From Free-Text Human Feedback–Collect New Datasets Or Extend Existing Ones?. In *The Twelfth International Conference on Learning Representations*.

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

[301] Savvas Petridis, Ben Wedin, Ann Yuan, James Wexler, and Nithum Thain. 2024. ConstitutionalExperts: Training a Mixture of Principle-based Prompts. *arXiv preprint arXiv:2403.04894* (2024).

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

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

[304] David Piorkowski, Inge Vejsbjerg, Owen Cornec, Elizabeth M Daly, and Öznur Alkan. 2023. 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 (2023), 1–25.

[305] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. 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*. 379–396.

[306] Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. 2021. 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*. 3784–3798.

[307] Carina Prunkl and Jess Whittlestone. 2020. Beyond near-and long-term: Towards a clearer account of research priorities in AI ethics and society. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 138–143.

[308] Dasha Pruss. 2023. Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 312–323. https://doi.org/10.1145/3593013.3593999

[309] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. In *The Twelfth International Conference on Learning Representations*.

[310] Yushan Qian, Weinan Zhang, and Ting Liu. 2023. 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*. 6516–6528.

[311] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).

[312] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36.

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

[314] Aida Ramezani and Yang Xu. 2023. 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)*. 428–446.

[315] Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. 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*. 13370–13388.

[316] Siddharth Reddy, Sergey Levine, and Anca Dragan. 2022. First contact: Unsupervised human-machine co-adaptation via mutual information maximization. *Advances in Neural Information Processing Systems* 35, 31542–31556.

[317] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. 2021. 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*. 1–13.

[318] Tim Rietz and Alexander Maedche. 2021. 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*. 1–14.

[319] Jen Rogers and Anamaria Crisan. 2023. 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*. 1–22.

[320] Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, et al. 2023. 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)*. 6252–6272.

[321] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. 2023. Identifying the Risks of LM Agents with an LM-Emulated Sandbox. In *The Twelfth International Conference on Learning Representations*.

[322] Stuart Russell. 2014. White paper: Value alignment in autonomous systems. November 1 (2014).

[323] Stuart Russell. 2022. Artificial Intelligence and the Problem of Control. *Perspectives on Digital Humanism* 19 (2022).

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

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

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

[327] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. 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)*. 9080–9102.

[328] Bidipta Sarkar, Andy Shih, and Dorsa Sadigh. 2023. Diverse Conventions for Human-AI Collaboration. *Advances in Neural Information Processing Systems* 36.

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

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

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

[332] Timo Schick, A Yu Jane, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. PEER: A Collaborative Language Model. In *The Eleventh International Conference on Learning Representations*.

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

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

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

[336] Nikil Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2023. 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)*.

1373–1386.

[337] Prithviraj Sen, Yunyao Li, Eser Kandogan, Yiwei Yang, and Walter Lasecki. 2019. 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*. 135–140.

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

[339] Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. 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)*. 9977–10000.

[340] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. 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*. 384–387.

[341] Hua Shen and Ting-Hao Huang. 2020. 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*, Vol. 8. 168–172.

[342] Hua Shen, Tongshuang Wu, Wenbo Guo, and Ting-Hao Huang. 2022. 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)*. 10–19.

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

[344] Hua Shen, Vicky Zayats, Johann Rocholl, Daniel Walker, and Dirk Padfield. 2023. MultiTurnCleanup: A Benchmark for Multi-Turn Spoken Conversational Transcript Cleanup. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 9895–9903.

[345] Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2023. Loose lips sink ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback. , 2859–2873 pages.

[346] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. 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)*. 4275–4293.

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

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

[349] Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. 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)*. 11289–11310.

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

[351] Anthony Sicilia and Malihe Alikhani. 2023. 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)*. 2898–2917.

[352] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. 2023. 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*. 1–18.

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

[354] Aviv Slobodkin, Niv Nachum, Shmuel Amar, Ori Shapira, and Ido Dagan. 2023. SummHelper: Collaborative Human-Computer Summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 554–565.

[355] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. 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*. 1–13.

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

[357] Jaemarie Solyst, Shixian Xie, Ellia Yang, Angela EB Stewart, Motahhare Eslami, Jessica Hammer, and Amy Ogan. 2023. "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*. 1–14.

[358] Jaemarie Solyst, Ellia Yang, Shixian Xie, Amy Ogan, Jessica Hammer, and Motahhare Eslami. 2023. 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 (2023), 1–27.

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

[360] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A Roadmap to Pluralistic Alignment. *arXiv preprint arXiv:2402.05070* (2024).

[361] Francesco Sovrano and Fabio Vitali. 2021. 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*. 81–91.

[362] Neha Pundlik Srikanth, Rupak Sarkar, Heran Y. Mane, Elizabeth M. Aparicio, Quynh C. Nguyen, Rachel Rudinger, and Jordan Boyd-Graber. 2024. Large Language Models Help Humans Verify Truthfulness—Except When They Are Convincingly Wrong. In *North American Association for Computational Linguistics*.

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

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

[365] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33, 3008–3021.

[366] Kees Stuurman and Hugo Wijnands. 2001. 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 (2001), 92–100.

[367] Ilia Sucholutsky and Tom Griffiths. 2023. Alignment with human representations supports robust few-shot learning. *Advances in Neural Information Processing Systems* 36.

[368] Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuan-Jing Huang. 2022. 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*. 3726–3739.

[369] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems* 36.

[370] S Shyam Sundar and Eun-Ju Lee. 2022. Rethinking communication in the era of artificial intelligence. *Human Communication Research* 48, 3 (2022), 379–385.

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

[372] Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy Rogers. 2023. 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*. 722–738.

[373] Phillip Swazinna, Steffen Udluft, and Thomas Runkler. 2023. User-Interactive Offline Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*.

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

[375] Yi Tay, Donovan Ong, Jie Fu, Alvin Chan, Nancy Chen, Anh Tuan Luu, and Christopher Pal. 2020. 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*. 5369–5373.

[376] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. 2020. 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*. 107–118.

[377] Maartje Ter Hoeve, Julia Kiseleva, and Maarten Rijke. 2022. 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*. 46–75.

[378] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. AI Alignment in the Design of Interactive AI: Specification Alignment, Process Alignment, and Evaluation Support. *arXiv preprint arXiv:2311.00710* (2023).

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

[380] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. 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*. 5433–5442.

[381] Yuandong Tian, Qucheng Gong, and Yu Jiang. 2020. Joint policy search for multi-agent collaboration with imperfect information. *Advances in Neural Information Processing Systems* 33, 19931–19942.

[382] Yuan Tian, Zheng Zhang, Zheng Ning, Toby Li, Jonathan K Kummerfeld, and Tianyi Zhang. 2023. Interactive Text-to-SQL Generation via Editable Step-by-Step Explanations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 16149–16166.

[383] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 801–810.

[384] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. 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*. 1–17.

[385] Marcel Torne Villasevil, Balsells I Pamies, Zihan Wang, Samedh Desai, Tao Chen, Pulkit Agrawal, Abhishek Gupta, et al. 2023. Breadcrumbs to the Goal: Supervised Goal Selection from Human-in-the-Loop Feedback. *Advances in Neural Information Processing Systems* 36.

[386] Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. 2021. 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)*. 152–166.

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

[388] Alex Turner and Prasad Tadepalli. 2022. Parametrically retargetable decision-makers tend to seek power. *Advances in Neural Information Processing Systems* 35 (2022), 31391–31401.

[389] Alexander Matt Turner, Logan Riggs Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2021. Optimal Policies Tend To Seek Power. (2021).

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

[391] Rama Adithya Varanasi and Nitesh Goyal. 2023. "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)*. Association for Computing Machinery, New York, NY, USA, Article 251, 17 pages. https://doi.org/10.1145/3544548.3580903

[392] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

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

[394] Mudit Verma and Katherine Metcalf. 2023. Hindsight PRIORs for Reward Learning from Human Preferences. In *The Twelfth International Conference on Learning Representations*.

[395] Jesse Vig, Wojciech Kryściński, Karan Goel, and Nazneen Rajani. 2021. 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*. 150–158.

[396] Alejandro Cuevas Villalba, Eva M Brown, Jennifer V Scurrell, Jason Entenmann, and Madeleine IG Daepp. 2023. Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models. *arXiv preprint arXiv:2309.10187* (2023).

[397] Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. 2022. Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems* 35, 4004–4016.

[398] Henrik Voigt, Özge Alaçam, Monique Meuschke, Kai Lawonn, and Sina Zarrieß. 2022. 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*. 348–374.

[399] Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*.

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

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

[402] Fengjie Wang, Xuye Liu, Oujing Liu, Ali Neshati, Tengfei Ma, Min Zhu, and Jian Zhao. 2023. Slide4N: Creating presentation slides from computational notebooks with human-ai collaboration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[403] Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. 2023. Aligning language models with human preferences via a bayesian approach. *Advances in Neural Information Processing Systems* 36.

[404] Nan Wang, Qifan Wang, Yi-Chia Wang, Maziar Sanjabi, Jingzhou Liu, Hamed Firooz, Hongning Wang, and Shaoliang Nie. 2023. COFFEE: Counterfactual Fairness for Personalized Text Generation in Explainable Recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 13258–13275.

[405] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926* (2023).

[406] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. 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*. 1–16.

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

[408] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. 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*. 1–14.

[409] Xingjin Wang, Linjing Li, and Daniel Zeng. 2023. LDM2: A Large Decision Model Imitating Human Cognition with Dynamic Memory Enhancement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 4660–4681.

[410] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022* (, Virtual Event, Lyon, France,) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 1697–1708. https://doi.org/10.1145/3485447.3512240

[411] Xinpeng Wang and Barbara Plank. 2023. 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*. 2046–2052.

[412] Xi Wang, Hossein Rahmani, Jiqun Liu, and Emine Yilmaz. 2023. Improving Conversational Recommendation Systems via Bias Analysis and Language-Model-Enhanced Data Augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 3609–3622.

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

[414] Xinru Wang and Ming Yin. 2023. 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*. 1–19.

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

[416] Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. 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)*. 8640–8665.

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

[418] Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023. Humanoid Agents: Platform for Simulating Human-like Generative Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 167–176.

[419] Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, et al. 2023. Democratizing Reasoning Ability: Tailored Learning from Large Language Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 1948–1966.

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

[421] Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–40.

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

[423] 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. 2023. AutoDOViz: Human-Centered Automation for Decision Optimization. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 664–680.

[424] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986* (2023).

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

[426] Charles Welch, Chenxi Gu, Jonathan K Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2022. 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)*. 1742–1752.

[427] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. 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*. 632–658.

[428] Norbert Wiener. 1960. Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science* 131, 3410 (1960), 1355–1358.

[429] Wikipedia. 2024. AI alignment — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=AI%20alignment&oldid=1220304776. [Online; accessed 05-May-2024].

[430] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. 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*. 1–16.

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

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

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

[434] Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. 2022. Context-aware Information-theoretic Causal De-biasing for Interactive Sequence Labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 3436–3448.

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

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

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

[438] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. 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*. 1–22.

[439] Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Cross-Cultural Analysis of Human Values, Morals, and Biases in Folk Tales. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[440] Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. 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*. 10691–10706.

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

[442] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can Large Language Model Agents Simulate Human Trust Behaviors?. In *ICLR 2024 Workshop: How Far Are We From AGI*.

[443] Binfeng Xu, Xukun Liu, Hua Shen, Zeyu Han, Yuhan Li, Murong Yue, Zhiyuan Peng, Yuchen Liu, Ziyu Yao, and Dongkuan Xu. 2023. Gentopia. AI: A Collaborative Platform for Tool-Augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 237–245.

[444] Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2023. 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)*. 13557–13572.

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

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

[447] Hao Yan, Saurabh Srivastava, Yintao Tai, Sida I Wang, Wen-tau Yih, and Ziyu Yao. 2023. 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)*. 3149–3170.

[448] Junbing Yan, Chengyu Wang, Taolin Zhang, Xiaofeng He, Jun Huang, and Wei Zhang. 2023. From Complex to Simple: Unraveling the Cognitive Tree for Reasoning with Small Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 12413–12425.

[449] Xue Yan, Jiaxian Guo, Xingzhou Lou, Jun Wang, Haifeng Zhang, and Yali Du. 2023. An Efficient End-to-End Training Approach for Zero-Shot Human-AI Coordination. *Advances in Neural Information Processing Systems* 36.

[450] Chunxu Yang, Chien-Sheng Wu, Lidiya Murakhovs'ka, Philippe Laban, and Xiang Chen. 2023. INTELMO: Enhancing Models' Adoption of Interactive Interfaces. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 161–166.

[451] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. 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*. 1–14.

[452] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. 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*. 1–13.

[453] Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, et al. 2023. 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*. 11629–11643.

[454] Bingsheng Yao, Prithviraj Sen, Lucian Popa, James Hendler, and Dakuo Wang. 2023. 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)*. 14698–14713.

[455] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From Instructions to Intrinsic Human Values–A Survey of Alignment Goals for Big Models. *arXiv preprint arXiv:2308.12014* (2023).

[456] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* (2024), 100211.

[457] Zonghai Yao, Benjamin Schloss, and Sai Selvaraj. 2023. Improving Summarization with Human Edits. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2604–2620.

[458] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. In *The Twelfth International Conference on Learning Representations*.

[459] Xi Ye and Greg Durrett. 2022. 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)*. 6199–6212.

[460] 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. 2022. 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*. 1–13.

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

[462] Chao Yu, Jiaxuan Gao, Weilin Liu, Botian Xu, Hao Tang, Jiaqi Yang, Yu Wang, and Yi Wu. 2023. Learning Zero-Shot Cooperation with Humans, Assuming Humans Are Biased. In *The Eleventh International Conference on Learning Representations*.

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

[464] Yifu Yuan, Jianye Hao, Yi Ma, Zibin Dong, Hebin Liang, Jinyi Liu, Zhixin Feng, Kai Zhao, and Yan Zheng. 2024. Uni-RLHF: Universal Platform and Benchmark Suite for Reinforcement Learning with Diverse Human Feedback. In *The Twelfth International Conference on Learning Representations*.

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

[466] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. 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)*. 11328–11348.

[467] Hongli Zhan, Desmond Ong, and Junyi Jessy Li. 2023. Evaluating Subjective Cognitive Appraisals of Emotions from Large Language Models. (2023), 14418–14446.

[468] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.

[469] Han Zhang, Yu Lei, Lin Gui, Min Yang, Yulan He, Hui Wang, and Ruifeng Xu. 2024. CPPO: Continual Learning for Reinforcement Learning with Human Feedback. In *The Twelfth International Conference on Learning Representations*.

[470] Jiaxin Zhang, Zhuohang Li, Kamalika Das, and Sricharan Kumar. 2023. Interactive Multi-fidelity Learning for Cost-effective Adaptation of Language Model with Sparse Human Supervision. *Advances in Neural Information Processing Systems* 36.

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

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

[473] Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2021. 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)*. 5612–5623.

[474] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. 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*. 1–30.

[475] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. 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*. 1–21.

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

[477] Lingjun Zhao, Khanh Nguyen, and Hal Daumé III. 2023. Define, Evaluate, and Improve Task-Oriented Cognitive Capabilities for Instruction Generation Models. In *Findings of the Association for Computational Linguistics: ACL 2023*. 3688–3706.

[478] Siyan Zhao, John Dang, and Aditya Grover. 2023. Group Preference Optimization: Few-Shot Alignment of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

[479] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. 2023. 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*. 1–19.

[480] Rui Zheng, Wei Shen, Yuan Hua, Wenbin Lai, Shihan Dou, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Haoran Huang, Tao Gui, et al. 2023. Improving Generalization of Alignment with Human Preferences through Group Invariant Learning. In *The Twelfth International Conference on Learning Representations*.

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

[482] Enyu Zhou, Rui Zheng, Zhiheng Xi, Songyang Gao, Xiaoran Fan, Zichu Fei, Jingting Ye, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2023. RealBehavior: A Framework for Faithfully Characterizing Foundation Models' Human-like Behavior Mechanisms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 10262–10274.

[483] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations*.

[484] Zhaowei Zhu, Jialu Wang, Hao Cheng, and Yang Liu. 2023. Unmasking and Improving Data Credibility: A Study with Datasets for Training Harmless Language Models. In *The Twelfth International Conference on Learning Representations*.

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