
A Survey on Human-AI Teaming with Large Pre-Trained Models

Vanshika Vats* **Marzia Binta Nizam*** **Minghao Liu** **Ziyuan Wang†** **Richard Ho†**
Mohnish Sai Prasad† **Vincent Titterton†** **Sai Venkat Malreddy†** **Riya Aggarwal†**
Yanwen Xu† **Lei Ding†** **Jay Mehta†** **Nathan Grinnell†** **Li Liu†** **Sijia Zhong†**
Devanathan Nallur Gandamani† **Xinyi Tang†** **Rohan Ghosalkar†** **Celeste Shen†**
Rachel Shen† **Nafisa Hussain†** **Kesav Ravichandran†** **James Davis**

University of California, Santa Cruz
United States
*{vvats, manizam}@ucsc.edu

Abstract

In the rapidly evolving landscape of artificial intelligence (AI), the collaboration between human intelligence and AI systems, known as Human-AI (HAI) Teaming, has emerged as a cornerstone for advancing problem-solving and decision-making processes. The advent of Large Pre-trained Models (LPTM) has significantly transformed this landscape, offering unprecedented capabilities by leveraging vast amounts of data to understand and predict complex patterns. This paper surveys the pivotal integration of LPTMs with HAI, emphasizing how these models enhance collaborative intelligence beyond traditional approaches. It examines the synergistic potential of LPTMs in augmenting human capabilities, discussing this collaboration for AI model improvements, effective teaming, ethical considerations, and their broad applied implications in various sectors. Through this exploration, the study sheds light on the transformative impact of LPTM-enhanced HAI Teaming, providing insights for future research, policy development, and strategic implementations aimed at harnessing the full potential of this collaboration for research and societal benefit.

1 Introduction

The journey towards blending human intelligence with advanced technology spans several centuries, beginning with inventive attempts to simulate human intellect in machines. One notable example from the 1770s is the "Mechanical Turk," a machine that appeared to play chess autonomously but was, in fact, operated by a person concealed inside it [32]. This early endeavor, while not "*Artificial Intelligence*" (AI) in the modern sense, reflects the enduring fascination with creating technology that can mimic or complement human cognitive abilities. The formal realization of integrating human-like intelligence into machines culminated in 1956 at Dartmouth College, marking the official birth of Artificial Intelligence as a research field [134]. As AI technology has advanced and become more prevalent over the last decade, researchers have identified limitations within purely automated systems [110, 52, 133] leading to renewed focus on augmenting AI with human expertise, aiming to

*Shared First Authorship, †Equal Contribution

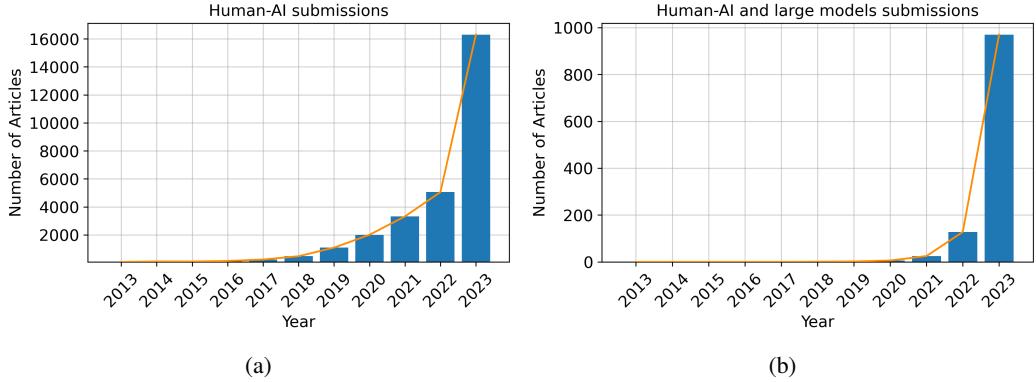


Figure 1: A graphical depiction of the number of relevant articles submitted from 2013-2023. (a) illustrates the increasing engagement of research communities in collaborations between humans and AI over the years. (b) observes a notably sharp increase in submissions related to human-AI and large-scale models attributed to the recent emergence of large models.

harness the best of both worlds (Fig. 1a). This approach seeks to enhance AI applications and foster a symbiotic relationship between human and artificial intelligence.

Navigating the rapidly advancing AI landscape, the integration of human cognition with Large Pre-trained Models (LPTM), including Large Language Models (LLM) [14, 151] and Large Vision Models (LVM) [80, 129, 69], has initiated a transformative shift. These models are "pre-trained" on vast datasets before being fine-tuned for specific tasks. This paradigm has opened new avenues for collaborative problem-solving and decision-making. LPTMs are shaped and refined by human expertise in terms of ethical guidance, creativity, and contextual understanding. In turn, AI amplifies human capabilities by processing data at large scales, offering insights and augmenting decision-making. This paper undertakes an extensive survey of Human-AI teaming, analyzing the complex interactions between human agents and sophisticated pre-trained models across various fields. Our study aims to uncover the progress made, confront the challenges, and understand the implications of this evolving partnership.

The primary contribution of this survey is to provide an in-depth perspective on the complex and diverse landscape of Human-AI teaming with large pre-trained models. We explore this synergy by studying Human-AI collaboration that refines AI model behavior, effective HAI systems, safety concerns, and sector-specific applications. Our goal is to inform future research, influence policy-making, and guide development strategies, ultimately driving an effective and ethical integration of Human-AI collaboration across various facets of society.

1.1 Scope of the Survey

Our survey focuses on the articles describing the developments of human-AI teaming through the years and how the introduction of large pre-trained models is reforming the field. The search for articles was conducted on Google Scholar using keywords such as "Human-AI", "Human-AI teaming", and other similar terms. For research focusing on the collaboration between humans and AI involving large models, additional keywords, including "large models", "large language models", and "large vision models" were utilized. Since human-AI teaming and large models have only recently gained a lot of interest from the research community given the recent popularity of large pre-trained models (Fig. 1b), some portion of studies cited in this survey could also be from arXiv preprints. We then sample the articles according to the structure desired for this survey, i.e., primarily covering Human-AI teaming and its large pre-trained model counterparts in better model training, effective human-AI joint systems, safe and secure HAI teaming, and their applications (Fig. 2a).

In order to show support for the theme of our survey article, we try to implement human-AI teaming with the help of large language models to prepare this review. For the *human* part, the authors survey the existing literature about human-AI teaming, collect a subset of articles according to the desired structure, organize the survey into relevant sections, study submission statistics and put together tables and visualizations, and finally, prepare the first complete draft of the manuscript. Using the *AI*

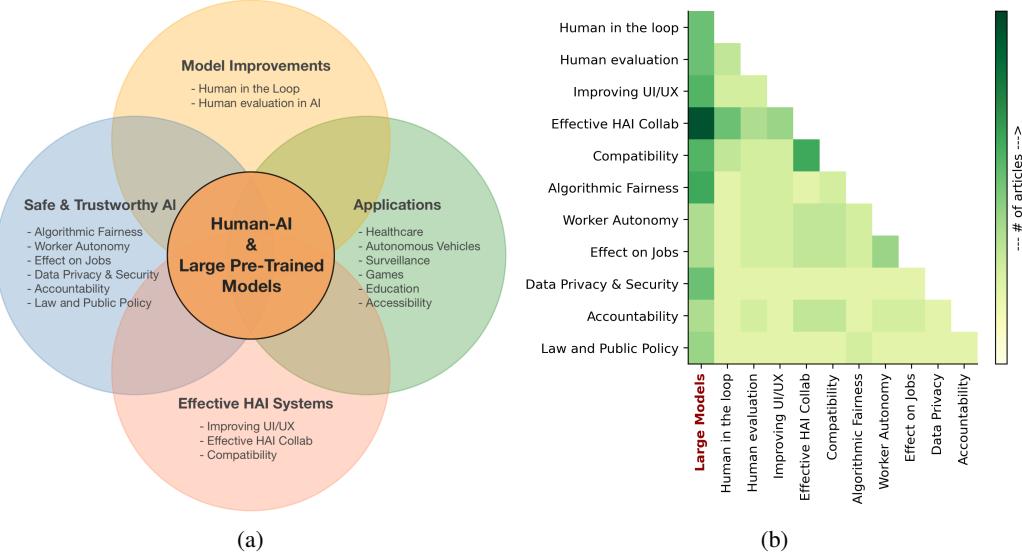


Figure 2: A representation of the scope of this survey. (a) We cover four broad categories - Model improvements, Effective HAI systems, Safe and Trustworthy AI, and their Applications - in the human-AI domain, along with the recent developments in them with the help of LPtM. The overlap between the four petals illustrates the fact that the four categories are not mutually exclusive; the articles may belong to two or more categories at the same time. (b) The heatmap shows a general trend of the number of articles co-existing within subcategories, including the articles discussing large models within each subject area. The overlap between the applications is not analyzed, given their specific and varied nature. *AI*: Artificial Intelligence, *HAI*: Human-AI, *UI/UX*: User Interface/User Experience.

part in human-AI teaming, we ask ChatGPT-4 [120] for its feedback regarding the language flow of each subsection, placement of the citations, and revising the language wherever necessary. This helps in the ease in identifying the missing elements and making necessary revisions. The final manuscript is, therefore, a product of human and AI collaboration.

1.2 Outline of the Survey

In each segment of our study, we commence by delineating traditional methodologies employed in human-AI collaboration, subsequently delving into the contributions of extensive pre-trained models in the fulfillment of these tasks. Section 2 explores the involvement of incorporating human expertise into the AI model training cycle, fostering a cooperative relationship between humans and AI in active learning scenarios, enhancing learning through human feedback, and involving human experts in the thorough evaluation of machine learning models [65, 138, 181]. Section 3 delves into understanding the scope of effective Human-AI joint systems, optimizing user interfaces (UI) and system architectures emerges as a crucial area. It covers a range of topics under innovative UI designs and system structures aimed at boosting efficiency and user engagement in collaborative settings [88, 44].

A paramount concern in the realm of Human-AI collaboration is the safety, security, and trustworthiness of AI systems. Section 4 of our comprehensive analysis encompasses multiple dimensions: mitigating algorithmic biases to uphold fairness, assessing the impact of Human-AI collaboration on workers' autonomy, well-being, and job satisfaction, examining economic repercussions on employment and wages, addressing data privacy and security concerns, building trust in AI systems, and navigating the legal and regulatory frameworks governing Human-AI interactions [82, 62, 13]. These factors are integral to creating an ethical and responsible environment for AI's integration into human-centric workflows. Furthermore, section 5 survey explores the broad spectrum of Human-AI teaming applications across various sectors. We investigate the unique challenges and opportunities in fields such as healthcare [11], autonomous vehicles, surveillance systems [60], gaming [46], education, and accessibility. Understanding the specific characteristics and benefits of Human-AI

Human-AI Topics	Subtopics	Articles Cited
Section 2: AI model improvements with human-AI teaming	Human in the loop	[169], [65], [105], [47], [138], [127], [95], [8], [90], [175], [29], [67], [87], [73], [144], [121], [181], [113], [130], [179], [72], [109], [177], [103]
	Human evaluation in AI	[6], [180], [5], [25], [65], [146], [34], [30], [121], [96], [24], [26], [152]
Section 3: Effective human-AI joint systems	Improving user interfaces for effective teaming	[179], [60], [15], [128], [108], [172], [84], [135], [178], [25], [31], [160], [154], [45], [44], [164], [88]
	Effective human-AI collaboration	[164], [172], [126], [39], [93], [115], [178], [112], [53], [139], [131], [113], [121], [181], [180], [63], [103], [97], [162], [157], [123], [105], [65], [1], [48], [7], [21], [154], [94], [114], [116], [107], [75], [132], [140], [20]
	Compatibility of human-AI systems	[165], [25], [9], [162], [41], [58], [117], [103], [98], [64], [7], [171], [157], [63], [94], [112], [105], [22]
Section 4: Safe, secure and trustworthy AI	Algorithmic bias and fairness	[54], [143], [33], [25], [111], [81], [66], [106], [83], [92], [50], [119], [10]
	Worker autonomy and well being	[25], [82], [123], [37], [162], [170]
	Effect on wages and jobs	[62], [28], [25], [123], [3], [37], [162], [158]
	Data privacy and security	[43], [176], [74], [136], [89], [57], [153], [56], [36], [149], [137]
	Trustworthy AI and accountability	[64], [126], [18], [162], [79], [6]
Section 5: Applications	Law and public policy	[70], [19], [143], [12], [148], [159], [101]
	Healthcare	[63], [11], [104], [102], [16], [27], [17], [99], [86], [71]
	Autonomous vehicles	[4], [97], [180], [48], [174], [124], [35], [167], [91]
	Surveillance and security	[125], [78], [60], [55], [23], [68]
	Games	[46], [173], [142], [2], [150], [156], [46], [65]
	Education	[118], [155], [168], [76], [40], [59], [38], [141], [42], [161], [61], [22]
	Accessibility	[85], [122], [77], [166], [51], [49], [147]

Table 1: A tabular representation of the four broader focus topics covered in this survey, with their relevant subtopics. Each category contains its cited articles for the ease of reader reference.

collaboration in these areas is key to influencing its future direction and maximizing its societal impact.

2 AI Model Improvements with Human-AI Teaming

In this section, we delve into the process of training AI models, a critical stage that underpins their effectiveness and efficiency. Drawing upon the framework presented in Maadi et al. [100], we explore the integral role of human-AI interactions in model training. This framework delineates three interdependent phases: 1) Data control, where data is pre-processed and produced under stringent quality standards; 2) Machine Learning (ML) modeling and execution, guided by human expertise; and 3) ML evaluation and refinement. Our discussion primarily centers on the second phase, highlighting its pivotal role in shaping models that fulfill human-centric goals of effectiveness and efficiency. This section covers four key approaches: Human in the Loop, where human judgment directly influences AI training; Active Learning, focusing on AI systems that efficiently incorporate human input; Reinforcement Learning from Human Feedback, where AI behavior is shaped by human feedback; and Human Evaluated Machine Learning, emphasizing the role of human evaluation in refining AI models. Together, these subsections provide insights into how human-AI collaboration can enhance the training process, leading to more effective and reliable AI models.

2.1 Human in the Loop

In the ‘Human in the Loop’ (HITL) approach, we explore the synergistic collaboration between human intellect and AI capabilities. This collaboration is especially critical in addressing AI’s limitations in self-learning and humans’ challenges in processing vast data within time constraints. The interplay of human and AI strengths leads to more robust and ethically sound AI systems. For

instance, during the model design phase, human insights on ethical considerations, as highlighted by Whittlestone et al. [169], enrich AI perspectives beyond mere technical aspects. Conversely, in the model training and execution phase, the directional flow of information between humans and AI plays a pivotal role. Strategies such as InstructRL, as proposed by Hu and Sadigh [65], demonstrate the effectiveness of NLP instructions from humans in guiding AI, enhancing the model’s explainability and alignment with our expectations. Similarly, a research on AI-assisted brainstorming illustrates how AI’s engagement with human ideas can foster greater creativity and innovation [105].

The diverse roles of humans, ranging from providing ethical oversight in the design phase to contributing to quality control and data labeling during model training and execution, are crucial for optimizing AI performance. These roles not only bridge gaps in AI knowledge but also ensure that AI systems are aligned with human values and needs.

As we proceed, we will delve deeper into two specific HITAL machine learning approaches: Active Learning, where AI models dynamically integrate human input to refine learning, and Reinforcement Learning from Human Feedback (RLHF), which leverages human feedback to shape AI behavior. These sections will illuminate the intricate ways in which human involvement is woven into the fabric of AI development, showcasing the multifaceted nature of Human-AI collaboration.

2.1.1 Human-AI Collaboration with Active Learning

Unlike traditional machine learning paradigms, active learning involves an iterative process where a model selectively identifies data requiring labeling, optimizing system performance with minimal training data [47]. Particularly, this approach is crucial for fine-tuning large pre-trained models, especially transformer-based ones, which require substantial labeled data for specific user scenarios. Active learning efficiently harnesses human expertise to pinpoint areas of uncertainty, enabling more targeted training with less annotation effort [138]. The active learning workflow begins with an unlabeled dataset and a pre-trained model. The model predicts labels for each sample, outputting confidence levels. When predictions fall below a quality threshold, human annotators step in for manual annotation. This iterative process of re-training the model with new labeled data continues until satisfactory confidence levels are achieved. Human AI teaming plays a vital role in efficient data labeling and continuous model improvement.

Recent research in this area has introduced novel methods to leverage human expertise effectively. For instance, Pries et al. [127] propose a method for efficient data labeling using precise distance measurements to filter data samples, streamlining the presentation of data to experts. Lu et al. [95] discuss human-guided interventions for continuous model improvement, focusing on dissociating biases. Bao et al. [8] explore the transformation of human-annotated rationales into continuous attention mechanisms, enhancing the learning of domain-invariant representations. Additionally, Lertvittayakumjorn et al. [90] present a generalizable approach for debugging deep text classification models with human input, applicable to larger models. Finally, Yao et al. [175] propose an explainable-generation Active Learning framework for simpler tasks, potentially improving the sampling process in data augmentation.

2.1.2 Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) is a paradigm that integrates human insights into the training process [29, 67]. This approach enhances the collaboration between humans and AI systems, especially during the data control phase — a critical stage for the success of large pre-trained models. In RLHF, the human role extends beyond mere data provision to actively shaping the AI’s learning trajectory. This human-AI synergy is particularly effective in mitigating biases in training data, leading to fairer AI outcomes. Additionally, the human touch in RLHF makes AI decisions more transparent and comprehensible, thereby fostering greater trust among users. The adaptability of RLHF models to user preferences underscores the significance of human understanding in complex, contextual situations, enabling AI systems to be more aligned with ethical standards and personalization needs.

However, the implementation of RLHF is not without its challenges. Ensuring the consistency and accuracy of human feedback, effectively scaling this approach for large datasets, and harmonizing human insights with algorithmic processes are areas that require careful consideration [87]. Over-

coming these challenges is essential to maximize the benefits of RLHF, paving the way for AI models that are not only technically proficient but also deeply attuned to human values and needs.

Reinforcement Learning Enhanced by Human Feedback

As an active research area of machine learning, Reinforcement Learning (RL) focuses on enhancing agent behaviors to make decisions by performing actions in an environment and receiving feedback as rewards or penalties. Due to the success of RLHF in large models (LMs) and the similar interaction in HAI teaming, researchers are now striving to figure out how to leverage RLHF for further improvement on HAI frameworks and even put forward more flexible ones.

Despite various theories and empirical models behind RL, as mentioned in [73], RL frameworks can be abstracted as a continuous interaction between agents and their surrounding environment, in which they take actions individually or cooperatively given their states and state transitions to maximize long-term rewards. Traditionally, given a designated reward model, if the accumulated long-term reward is exactly or approximately converged under limited horizons, the agent's behaviors and the corresponding dynamics [144], like policy $\pi(a|s)$, Q function $Q(s, a)$, etc. are more likely to converge. Only after OpenAI's introduction of instructGPT [121] to illustrate how a reward model trained by high-quality data from labelers can optimize the LM's ranking of generative content did research communities realize RL's critical role in improving LM performance. Since labelers are asked to label this data in multiple turns to yield the most suitable outputs, this interaction pattern under RL formalization is considered RLHF. Furthermore, this working pattern can be widely applied in similar scenarios within feedback-acquiring processes. In the HAI teaming model, once human instructions are quantitatively converted to a reward model by ranking or scoring, the interaction between human and AI system agents can be formalized and optimized by RLHF.

Design of RLHF

In exploring the design of RLHF, a key challenge emerges: aligning human instructions with RL agent behaviors to ensure that these instructions effectively guide and evaluate agent actions. This alignment, complex in nature, has been a pivotal focus in recent large language model (LLM) research, contributing significantly to the advancement of LLMs and their applications. This section reviews the development of RLHF from its early stages in 2019 to its current evolution.

A notable early work in RLHF is by Ziegler et al. [181], where language models were trained to improve text generation and summarization based on human preferences. However, this approach, limited by its reliance on a single-run process and restricted data, faced challenges in generalization. In contrast, OpenAI's instructGPT [121] introduced a three-phase training paradigm: initial training with supervised policy via demonstration data, training a reward model with labeler feedback, and optimizing the ranking policy using RL against the reward model. This methodology provides a robust baseline and ensures that GPT models evolve with continuous human feedback, leading to the creation of more sophisticated and responsive models, such as the web browser-assisted question-answering agent by Nakano et al. [113].

Despite these advancements, RLHF remains a complex and often unstable process. To address this, Stanford's recent research has introduced the Direct Preference Optimization (DPO) training model [130]. DPO simplifies policy training by directly mapping reward functions to optimal policies, circumventing the need for intricate reward maximization constraints. This approach not only surpasses traditional RLHF in controlling sentiment in generations but also enhances response quality in tasks like summarization and single-turn dialogue. Its simplicity in implementation and training offers a promising direction for further simplifying RLHF processes.

New Human-AI Teaming Paradigms in RLHF

The integration of Reinforcement Learning from Human Feedback (RLHF) into Human-AI (HAI) processes presents unique challenges due to the intricate steps and components involved, as evidenced by the work of Zhang et al. [179] and Johnson et al. [72]. Recognizing these complexities, recent research has shifted focus towards enhancing collaboration within the closely-coupled phases of HAI, with the aspiration of exploring further extensions in the future. A notable gap in current RLHF models is the communication barrier between AI and humans, highlighting the need for more advanced frameworks to bridge this divide.

A significant contribution to human-agent coordination is the instructRL model by Hu and Sadigh [65], which employs natural language instructions to align RL agents with human-preferred policies. This model leverages the embedded knowledge of pre-trained models, enabling AI agents to collaborate more effectively with humans across various tasks, including gaming, robotics, and control decision-making. InstructRL's adaptability to different domains through knowledge transfer and meta-learning positions it as a versatile model for diverse HAI scenarios.

Research by Mirchandani et al. [109] suggests that Large Language Models (LLMs), even without additional training, can function as proficient general sequence modelers due to their in-context learning abilities. This research explores the use of LLMs in robotics, particularly for extrapolating state sequences and prompting reward-conditioned trajectories, despite challenges like latency and compute costs. Chain-of-thought prompting further enhances LLM performance in complex reasoning tasks, as shown by Wei et al. [163]. The Distillation and Retrieval of Online Corrections (DROC) system by Zha et al. [177] exemplifies the capability of LLMs to assimilate language feedback and distill knowledge from corrections, demonstrating effectiveness in novel settings.

However, McNeese et al.'s research [103] provides a counterpoint to the notion of RLHF as a universal solution. Despite the success of NeoCITIES [103] RL agents in simulated environments, the disparity between these environments and real-world complexities remains a significant hurdle. Furthermore, RL agents lack proficiency in non-verbal communication and still struggle with detailed verbal interactions, indicating a need for more sophisticated models that can fully comprehend both verbal and non-verbal human communication. These observations underscore the necessity for developing new HAI frameworks that go beyond existing RLHF models, calling for enhanced human integration to address the multifaceted nature of real-world environments.

2.2 Human Evaluation in AI

Human Evaluation in Artificial Intelligence (Human-Eval-AI) is an essential process that integrates human expertise with AI-driven insights. This collaborative approach leverages the unique strengths of human judgment to refine and guide AI decision-making. The objective is to develop models that are accurate and also resonate with the diverse and multifaceted nature of our world. Such collaboration is key to training robust models, enhancing decision-making processes, and fostering a symbiotic relationship to improve model training.

2.2.1 Evaluation of Trust and Transparency

Establishing trust and transparency is crucial in Human-Eval-AI systems. Trust, inherently linked with transparency and understandability, aligns closely with how users perceive an AI system's abilities and limitations. As highlighted by Bansal et al. [6], it is essential for users to recognize the AI system's potential for errors. This understanding enables users to develop a realistic mental model of the AI, crucial for anticipating possible failures or underperformance.

Zhou et al.'s work [180] further underscores this by recommending the integration of uncertainty measures and machine learning model performance into AI-informed human decision-making processes. They advocate for a trial-to-trial approach in decision-making, where users adjust their judgment boundaries based on their experience of uncertainty and AI interactions. Their proposed Uncertainty-Performance Interface (UPI) allows users to concurrently evaluate the quality and performance of machine learning models, fostering deeper trust in AI systems. Similarly, Av et al. [5] emphasize the importance of incorporating intrinsic uncertainty into the collaborative process. They suggest a mechanism for human experts to offer feedback, either as corrections to current AI recommendations or insights on preferable aspects within the search space. This feedback is integrated into a Bayesian optimization model, aiming to align the model's inferences with the expert's domain-specific knowledge.

2.2.2 Approaches to Improve Human Evaluation

In Human-Eval-AI, human input remains essential, especially in instructing AI agents and refining AI decision-making processes. The study by Chhibber et al. [25] showcases the benefits of integrating human teachings into AI, not only enhancing AI performance but also building trust and influencing job delegation practices. In a similar vein, Hu and Sadigh's instructRL [65] introduces a novel form of human-AI interaction through high-level natural language instructions, thereby improving

AI interpretability and alignment with human intentions. Aaquib et al.’s research [146] on mental modeling in human-robot teaming reveals that mutual understanding between humans and AI significantly boosts team performance, advocating for AI systems capable of adapting to human mental models. Complementing this, Correia et al. [34] emphasize context in AI system design, proposing personalized user interactions for more relevant human-machine team actions and information.

Further exploring human evaluation, Clark et al. [30] examine how human evaluators can differentiate between texts produced by humans and models like GPT-2 and GPT-3. They suggest methods to enhance evaluation accuracy, including detailed instructions, annotated examples, and text comparisons. Ouyang et al. [121] present a different approach, fine-tuning GPT-3 into ‘InstructGPT’ using human feedback for improved user instruction alignment. Additionally, Lu et al.’s Error Analysis Prompting (EAPrompt) [96] focuses on evaluating machine translations by LLMs, leveraging Chain-of-Thought reasoning to refine LLM evaluative capabilities. Chen et al.’s work shows domain specific evaluation of LLM in problem-solving [24]. Chiang and Lee [26] investigate LLMs for human-like text quality assessments in specific NLP tasks, highlighting similarities with human evaluations. Finally, Uchendu et al. [152] highlight the effectiveness of collaborative efforts in enhancing detection capabilities, offering a unique perspective on human collaboration in evaluations.

3 Effective Human-AI Joint Systems

This section concentrates on achieving both efficiency and safety of the joint Human-AI (HAI) systems. The focus is on merging the strengths of different HAI systems to ensure optimal performance while maintaining the highest standards of safety.

3.1 Improving User Interfaces for Effective Teaming

The following module discusses enhancing human-AI collaboration through improved user interfaces and data processing. It highlights AI’s role in team adaptation, personalization, conversational dynamics, and User Interface/Experience (UI/UX) design. The discussion will be focused around AI’s influence on team collaboration and trust, its facilitation of learning and feedback via dialogue, and the significance of ethical, effective UI/UX design in supporting human collaboration.

3.1.1 Adaptation and Personalization

The integration of AI agents into human teams has historically been limited by their narrow adaptability and personalization capabilities. Traditional AI systems struggled to understand human subtleties and adapt to team dynamics, being confined to specific tasks without the ability to learn or adjust, thereby reducing their effectiveness in collaborative settings [179].

The introduction of large pre-trained models has significantly revolutionized AI’s adaptability and personalization in teamwork. These models, trained on extensive datasets and utilizing sophisticated algorithms, have a markedly improved capacity to comprehend and adapt to team changes. Research demonstrates that LLM-equipped AI agents can fine-tune their autonomy, aligning with human team dynamics and operational needs more seamlessly [60]. LLMs have advanced conversational capabilities, enabling more natural, context-aware interactions and supporting more complex, multi-party engagements [15, 128]. Their application in mobile UI and UX evaluation also showcases their role in personalizing interactions and enhancing team collaboration [108]. The practice of chaining LLM prompts underscores their potential to customize Human-AI collaboration, ensuring greater control and transparency in the process [172].

3.1.2 Conversational Interaction

In Human-AI Teams (HATs), effective communication is as crucial as in human teams, with the conversational capabilities of AI playing a pivotal role in maintaining trust, facilitating knowledge sharing, and ensuring smooth coordination. These capabilities are essential for both the cognitive processing of tasks and the emotional dynamics within the team, impacting decision-making and situational awareness [84, 135]. Quick and appropriate responses from AI can enhance human members’ confidence in their AI partners, leading to better teamwork [178].

The proactive sharing of information by AI, crucial for building trust and promoting collaboration, has seen significant strides with the development of Teachable Conversational Agents. These agents improve text classification tasks by learning from dialogue-based teaching and feedback, adjusting their focus based on interaction, which can range from acquiring new knowledge to refining existing information based on the conversation's context [25]. The perception of these agents' likability and human-likeness is key to their effectiveness in communication [31]. Voice-based interactions, supported by Natural Language Processing (NLP), offer substantial benefits for both verbal and written exchanges in HATs [160].

Research efforts have also been made to create a "Shared Vocabulary" (Taxonomy Model) for HATs, aiming for a unified set of terms and concepts accessible to both humans and AI. Furthermore, the "Communication and Explanation" framework (Communication Model) suggests equipping AI with the skills and training necessary for clear communication and explanations of their actions and reasoning, improving team understanding and efficiency [154]. This approach not only fosters mutual comprehension but significantly boosts team performance.

3.1.3 UI/UX

The evolution of User Interface (UI) and User Experience (UX) design in Human-AI Joint Systems is crucial in defining the quality of collaborative experiences. Historically, UI design principles have focused on ensuring user-friendly and ethical interactions between humans and AI systems. These ethical concerns, particularly about their impact on individuals and society, have become more pressing as AI's role in human interactions expands [45]. The complexity of such interactions demands rigorous usability testing, requiring a keen eye for detail and a profound understanding of human-computer interaction principles [44]. Traditional methodologies have highlighted the importance of clear communication and the synchronization of AI in UX evaluations, allowing AI to articulate its findings in usability testing and to determine effective modes of collaboration, whether they occur in real-time or are delayed.

The recent introduction of large models has significantly transformed UI/UX design practices within Human-AI Joint Systems. Innovations such as the two-stage frameworks, with ChatIE being a prime example, have redefined zero-shot Information Extraction (IE) tasks into multi-turn question-answering sequences, marking a significant advance in the field. These models, including ChatIE, have outperformed numerous established models across various datasets [164]. Further, studies involving human subjects in varied social settings have shown that these advanced approaches not only improve the quality of interactions but also lead to more engaging and emotionally gratifying user experiences. This reflects the burgeoning capability of artificial intelligence to adopt more human-like behaviors and seamlessly integrate into social contexts [88]. This progression highlights a significant shift in the landscape of UI/UX design in Human-AI systems, where the deployment of large-scale models plays an increasingly central role in enhancing user experience and interaction fidelity.

3.2 Effective Human-AI Collaboration

In this section, we categorize and discuss different interaction methods for human-AI collaboration. We divide these methods into 3 subsections: Human Helps AI, AI Helps Human, and their combination. Human Helps AI focuses on systems where the AI has the final say, and humans provide it with support and guidance along the way. Human oversight and intervention are also discussed in this subsection. AI Helps Human is the inverse; where a human produces the final output, using AI assistance along the way. The final part looks into systems where humans and AIs have more equal roles. Specifically, how the outputs of both humans and AI can be 'combined' to produce a higher quality output than either can produce alone.

3.2.1 Human Helps AI

The indispensable role of human input in the development and refinement of AI systems is well-documented, underscoring the symbiotic relationship between human expertise and AI capabilities. Human involvement is pivotal at various stages of AI system development, from initial data preparation to the fine-tuning of algorithms for enhanced precision and reliability.

Data annotation and preparation, as highlighted by Wei et al. [164], serve as the bedrock for training AI systems, ensuring they are equipped with accurate and contextually relevant data. This foundational work is critical for the creation of AI models that are both effective and trustworthy. The customization and refinement of AI algorithms, driven by human feedback and insights, further illustrate the central role of human expertise in AI development, enhancing the system's accuracy and applicability [172]. Ethical considerations and contextual understanding are also paramount, with Pflanzer et al. [126] emphasizing the need for AI to operate within societal norms and comprehend complex human contexts—an aspect that purely algorithmic approaches may miss. The adaptation of AI systems to specific industry requirements showcases the importance of tailoring AI technologies to meet diverse application needs, thus enhancing their personalization and relevance [39, 93, 115].

Interface design and communication strategies, as discussed by Zhang et al. [178], play a crucial role in making AI tools both accessible and user-friendly, bridging the gap between sophisticated AI outputs and practical user interaction. The continuous evolution of AI systems necessitates ongoing human intervention and guidance to stay abreast of advancements in the field [112]. In specialized domains like adaptive AI systems for driving, human behavioral data and feedback are utilized to simulate and incorporate diverse driving behaviors, demonstrating the integration of human subtleties into AI training for real-world applications [53]. Hybrid human-AI teaming scenarios, where control is dynamically allocated between humans and AI based on situational awareness, further highlight the collaboration between human judgment and AI efficiency [139].

Recent advancements in large language models (LLMs) have also showcased the significant impact of human feedback on improving AI performance. Direct Preference Optimization (DPO), as introduced by Rafailov et al. [131], represents a significant leap forward in the customization of language models. DPO focuses on aligning model outputs with human preferences through a direct optimization process, bypassing the complexities and limitations associated with traditional reinforcement learning approaches. Similarly, the work of Nakano et al. [113] demonstrates how LLMs like GPT-3 can be fine-tuned using human feedback for tasks such as web-assisted question-answering. Ouyang et al. [121] further explore the benefits of fine-tuning LLMs with human feedback through the development of instructGPT models. The research of Ziegler et al. [181] complements these findings by applying reward learning, informed by human judgments, to tailor language models for specific tasks such as stylized text generation and summarization. This approach underscores the versatility of human feedback in enhancing AI models, enabling them to adapt to a wide range of stylistic preferences and content requirements. These contributions collectively underscore the critical role of humans in shaping AI development, ensuring that AI systems not only advance in capability but also align with ethical standards, societal norms, and user expectations, thereby fostering a constructive and sustainable human-AI partnership.

3.2.2 AI Helps Human

This discussion delves into how AI systems, from traditional approaches to cutting-edge Large Language Models (LLMs), provide pivotal assistance across various domains, enhancing human capabilities and streamlining complex tasks. It traverses the evolution from conventional AI applications, such as medical diagnostic aids and vehicular control improvements, to the transformative potential of LLMs in data analysis, creative content generation, and decision-making, illustrating the dynamic progression of AI as an indispensable partner in human endeavors.

Initial discussions focus on traditional AI's role in simplifying complex tasks, like transforming zero-shot information extraction into manageable question-answering formats, aiding humans in navigating large datasets effectively [164]. Additionally, AI's contribution to enhancing human decision-making, particularly under uncertainty, is explored through its applications in autonomous vehicles, emphasizing the need for transparency and trust in AI-human collaborations [180]. The use of ML-based systems in healthcare, such as TREWS for sepsis detection, exemplifies AI's capacity to act as a "second pair of eyes," assisting clinicians with the voluminous clinical data for timely interventions [63]. Experimental studies on team dynamics, incorporating both humans and AI agents, reveal the efficiency gains from AI integration, where AI handles tasks suited to its capabilities, thus optimizing team performance [103]. Furthermore, innovations in human-machine collaboration, such as intelligent haptic interfaces for vehicular control, showcase improvements in safety and efficiency during the transition from automated to manual driving [97].

More recently, impressive capabilities of LLMs in assisting humans, like chaining multiple LLM prompts offer a novel approach to complex task management, enhancing the human-AI collaborative experience through a transparent and controllable interface [172]. LLMs' role in augmenting human tasks is further detailed, highlighting their utility in providing insights and suggestions for tasks that would otherwise be daunting for humans [162]. The discussion also touches on the balance required in leveraging LLMs to maintain the authenticity of human responses, underscoring the potential for high-quality but homogeneous outputs versus lower-quality but genuine human responses [157]. Interdisciplinary collaboration is advocated as a means to shape the future of LLMs and AI-human interaction, promoting a proactive and global perspective [123]. The impact of AI on human brainstorming processes, where systems like GPT-3 stimulate broader idea generation, is also examined, indicating AI's potential to enhance cognitive processes and creativity in individual and collaborative settings [105]. The instructRL framework introduces a method where AI agents, trained through reinforcement learning, are fine-tuned with human-provided natural language instructions, demonstrating the efficacy of intuitive human-AI collaboration [65]. The application of ChatGPT in software architecture, aiding in the translation of architectural narratives into technical specifications, exemplifies AI's role in enhancing productivity and fostering creative design processes [1]. Lastly, the examination of AI's role in driving safety and autonomous systems, via a delegation manager that intelligently assigns control based on the situation and capabilities, illustrates the potential for AI to significantly improve operational safety and efficiency in hybrid human-AI systems [48].

3.2.3 Combining Human and AI Strengths

The essence of merging the strengths of humans and AI lies in leveraging the unique capabilities of each. It involves a careful balance where each compensates for the other's limitations, aiming for a collaboration where the overlap in errors is minimized, thus maximizing the opportunity for mutual correction and improvement. Research indicates that while explanations of AI predictions can foster blind trust, they may not always promote the appropriate reliance needed in critical domains where human oversight is crucial [7]. Optimal collaboration requires an intricate understanding of each other's capabilities, where humans excel in cognitive skills like context recognition, anticipation of team behaviors, and corresponding action support [21].

However, recognizing and integrating these diverse forms of intelligence within Human-AI Teams (HATs) poses new challenges, necessitating an in-depth understanding of AI's capabilities and limitations by human teammates [154, 94]. Initially, mismatches in mental models of AI's abilities can lead to errors, but over time, as humans develop a more accurate understanding, they learn to effectively complement the AI's functions. Advanced hybrid AI approaches, including neuro-symbolic frameworks, have been introduced to model human psychological states, aiding in the prediction and adaptation to human behaviors for enhanced cooperation [154, 114, 116].

Adaptation in AI behavior, aimed at enhancing collaboration, involves learning from relevant human features and modifying behavior to align with human goals, thereby fostering a more intuitive and effective partnership [107]. Incorporating computational models of emotion into mental models allows AI agents to consider affective states in decision-making, enriching the collaboration with sophisticated, goal-oriented actions [75]. Establishing a common vocabulary and shared understanding of goals, capabilities, and emotional states further solidifies the foundation of effective Human-AI collaboration [154].

Recent research initiatives showcase practical applications of this synergy. AdaTest++, an augmented tool for auditing large language models (LLMs), exemplifies the collaborative effort between humans and AI in enhancing model reliability and understanding, facilitating the identification of model failures through a combined effort [132]. In the realm of mental health support, the HAILEY system demonstrates how AI can augment human empathy with analytical insights, providing feedback-driven support for more empathetic and effective peer-to-peer conversations [140]. Additionally, the innovative integration of LLMs and diffusion models for creating visual metaphors highlights the creative potential unleashed by human-AI collaboration, where human-proposed linguistic metaphors are visually realized through AI-generated imagery, bridging the gap between conceptual ideas and visual representation [20].

3.3 Compatibility of Human-AI Systems

In human-AI joint systems, compatibility is crucial for ensuring that AI systems can seamlessly work alongside existing human technologies, thereby enhancing the overall efficacy of the system. As systems grow in complexity, system designers and engineers face the intricate challenge of ensuring harmonious interaction between humans and AI. This integration encompasses aspects such as communication protocols, security measures, scalability considerations, and workflow integration. When AI systems are tailored to fit within the existing workflows of human users, they facilitate smooth integration, avoiding disruptions or compromises to the system's current functioning. In the context of human-AI systems, compatibility transcends technical alignment. It encompasses the adaptability of AI systems to existing human processes, the user-friendliness and learning curve associated with new AI tools, and the effectiveness of these compatible joint systems in achieving their intended results. This section delves into these key aspects, highlighting the importance of compatibility in the effectiveness and safety of human-AI systems.

3.3.1 Backwards Compatibility

In the realm of AI technologies, backward compatibility plays a pivotal role in ensuring the smooth integration of new or updated systems with pre-existing frameworks. This concept is crucial when embedding AI into legacy systems, as it involves careful consideration of existing tools, processes, and practices. Weisz et al. [165] delve into how generative models contribute to the modernization of applications, emphasizing the necessity for AI systems to function cohesively within older infrastructures. In the context of traditional crowd work, Chhibber et al. [25] illustrate the significance of aligning AI tools with established workflows, showcasing the potential of generative models, like Generative Adversarial Networks (GANs) and autoencoders, to augment and refine datasets in legacy systems. Additionally, Bau et al. [9] present a novel approach using GANs to adapt image priors to specific image statistics in individual photographs, thereby tackling challenges in reproducing and manipulating high-level image attributes, as evidenced in various semantic editing tasks.

Discussing the broader compatibility challenges outside just the technological perspective, Wang et al. [162] talks about the transformative effects of Large Language Models (LLMs) on the job market, underscoring the necessity of integrating these advanced technologies into existing economic structures without causing major disruptions. Eloundou et al. [41] examine the impact of LLMs on the US labor market, revealing that approximately 80% of workers could experience a significant change in their tasks, with 19% potentially seeing over half of their tasks affected. This 'exposure' refers to a 50% increase in task efficiency without compromising quality. Furthermore, Harrer et al. [58] highlight the potential of LLMs in generative AI applications, emphasizing the need for responsible design, human oversight, and ethical considerations to prevent the spread of misinformation. They offer insights into how these models, when developed responsibly, can serve as efficient and trustworthy tools in sectors like healthcare, addressing technology, risks, and ethical issues.

3.3.2 Learning Curve and Effective Usability

In Human-AI Teaming (HAT), a critical factor is the ease and efficiency of interaction between humans and AI which significantly influences the learning curve and overall usability. McNeese et al.'s [103] study talks about the team composition in HAT, revealing that AI's ability to take a leading role in teams can enhance performance, though it often presents an initial learning curve for human participants. Lyons et al. [98] discuss the importance of intuitive interfaces in fostering a partnership, rather than a tool-based relationship, between machines and humans. A pivotal aspect in this regard is the AI's ability to explain its decisions, especially when those decisions may seem counter-intuitive to human collaborators. Hou et al.'s [64] research highlights that trust in AI is bolstered when humans understand the rationale behind AI decisions, fostering a sense of vulnerability and reliance on AI actions. This concept is further supported by Bansal et al. [7], who find that AI explanations significantly increase human trust in the algorithm.

Exploring the potential of Large Models, Wu et al. [171] also explore that breaking down the complex human tasks into multiple LLM prompts and chaining them together helps with increased human-AI interaction towards improvement. The users improvised over the new ways of interacting with the LLMs by leveraging the sub-tasks to set calibrate model expectations and making new-unit tests for each submodule of the LLM chain. Veselovsky et al.'s [157] study, however, suggests that even though prevalent usage of LLMs generating high-quality but homogeneous responses might not truly

capture the real essence of human behavior, however, they can prove to be useful too. They analyze that human workers not using LLMs in their tasks produced lower quality (less essential) information than when allowed the use of LLMs. They, hence, given the rapidly evolving large model landscape, emphasize a need to constantly monitor the co-evolution of LLM and human teaming to get to some constructive contributions.

3.3.3 Evaluation the Effectiveness of Joint Human-AI Systems

Evaluating the efficacy of human-AI systems extends beyond technical performance, encompassing a holistic approach that includes technical, social, and psychological dimensions of human-AI collaboration. The study by Henry et al. [63] emphasizes the importance of user trust and system usability, suggesting that evaluation frameworks should assess how systems support user autonomy and are endorsed by domain experts to facilitate adoption. Similarly, Liu et al. [94] highlight the need to understand perceptual differences between humans and AI, advocating for evaluations that measure how these systems leverage the complementary strengths of human and machine intelligence to improve joint decision-making processes. Munyaka et al. [112] delve into the impact of decision-making styles and AI identity disclosure on team dynamics and efficacy, underlining the significance of evaluating social dynamics and collaboration patterns within human-AI teams. Furthermore, Memmert et al. [105] explore the integration of AI in creative processes, pointing to the necessity of assessing both the cognitive contributions of AI and its effects on group dynamics, such as the potential for free-riding.

In terms of large models, Chang et al. [22] present a comprehensive theoretical framework for evaluating LLMs. This framework spans a wide array of metrics and scenarios, such as natural language processing, reasoning capabilities, and robustness. This study underscores the necessity of multifaceted evaluation frameworks that extend beyond task performance to include considerations like ethical implications and bias, which are crucial in human-AI systems. On the other hand, Veselovsky et al. [157] adopt a more empirical approach by examining the practical use of LLMs in crowd work. Their evaluation method involves observing and adapting the behavior of crowd workers as they interact with LLMs, offering a practical perspective on Human-AI interaction. These studies collectively suggest that a comprehensive evaluation of human-AI systems should not only focus on accuracy or efficiency but also consider factors like trust, usability, collaboration dynamics, and the psychological impact on users, to ensure the development of systems that are effectively integrated into human workflows and decision-making processes.

4 Safe, Secure and Trustworthy AI

The integration of AI in professional domains has been both revolutionary and disruptive, significantly affecting employment and wage dynamics. In this section, we embark on a detailed examination of the multifaceted implications of these developments. First, we address the critical issue of algorithmic bias and fairness. By drawing on existing literature, we explore the complexities surrounding this challenge and investigate potential solutions. Next, we assess how AI impacts worker autonomy and job satisfaction, shedding light on both positive and negative effects. The concept of trustworthy AI is then analyzed, with a focus on the crucial aspect of accountability in human-AI interactions. Subsequently, we explore the intricate relationship between AI, law, and public policy, underscoring the challenges in developing regulatory frameworks that effectively balance ethical considerations with rapid technological advances. Our objective throughout this section is to provide an informed perspective on human-AI collaboration and its socio-economic implications, contrasting these new models with traditional paradigms.

4.1 Algorithmic Bias and Fairness

Algorithmic fairness is a cornerstone of contemporary AI, demanding the development and deployment of unbiased and non-discriminatory algorithms. This pursuit, though challenging due to the complexities in defining fairness, mitigating data biases, and ensuring algorithm transparency, has become a focal point for researchers and organizations [54]. Fairness in algorithms is inherently complex. For instance, biases in the training data can lead to skewed outputs, and inherent biases in algorithmic design can persist even with unbiased data. These issues are further complicated by the often opaque nature of algorithms and the lack of established accountability measures [143].

Several strategies have been proposed to address these challenges. Key among them are ensuring data representativeness and employing algorithms known for their fairness. Pre-deployment testing for biases, enhancing transparency, and establishing robust accountability mechanisms are equally crucial [33]. Chhibber et al.'s [25] research sheds light on the role of user interaction in shaping algorithmic outcomes. It demonstrates how data quality, user engagement, and system localization impact fairness in machine learning systems. Additionally, Morrison et al.'s [111] study underscores the value of causal explanations in AI, arguing that they foster trust, facilitate error correction, and support informed decision-making, thereby contributing to the enhancement of algorithmic fairness.

The incorporation of Large Language Models (LLMs) into various applications has intensified the need to address algorithmic bias and fairness. Kirk et al. [81] have shown that biases in large language models, particularly those available 'out-of-the-box' like GPT-2, can lead to stereotypical and less diverse occupational associations for women, especially when intersecting with other categories like religion or ethnicity. Meanwhile, Huang et al. [66] reveal that a significant portion of code functions generated by LLMs can contain biases related to sensitive attributes such as age and gender, raising concerns about the fairness and integrity of software applications reliant on these models. Meyer et al. [106] explore the paradigm shift brought by conversational LLMs, discussing their potential to enhance academic work while cautioning against their inherent biases and accuracy issues. Kotek et al. [83] further investigate gender stereotypes within LLMs, demonstrating that these models are significantly more likely to align occupations with stereotypical gender roles, sometimes amplifying biases beyond societal norms or official statistics.

Recent studies are also working on diverse strategies for mitigating bias in LLMs, emphasizing the importance of fairness across their applications. Li et al. [92] review fairness strategies for both medium and large-sized LLMs, highlighting the need for tailored debiasing methods based on model size and training paradigms. Gallegos et al. [50] offer a detailed survey on bias evaluation and mitigation, categorizing techniques across model development stages and introducing taxonomies for bias metrics and mitigation methods. Ohi et al. [119] propose a novel approach to counteract likelihood bias in LLM evaluations using few-shot examples for in-context learning, showing marked improvements in fairness. Bi et al. [10] focus on group fairness in social media, presenting a hierarchical schema for bias evaluation and a chain-of-thought mitigation method, GF-Think, to reduce bias. These contributions underline possible approaches to bias mitigation, aiming for ethical and equitable use of LLMs.

4.2 Worker Autonomy and Well Being

In the realm of AI-enhanced platforms, worker autonomy, well-being, and job satisfaction are pivotal, directly influencing engagement and productivity. Current AI practices sometimes neglect these factors, leading to dissatisfaction and turnover. Addressing this, several studies propose frameworks to improve these aspects in AI-centric work environments.

Chhibber et al. [25] demonstrate how teachable conversational agents in crowd-based applications can empower workers, enhancing control, personalized learning, and ownership, thereby boosting job satisfaction. Konstantis et al. [82] emphasize transparency and fairness in crowdsourcing platforms, advocating for clear decision-making and respectful treatment to improve worker experience. Furthermore, Pal et al. [123] highlight the design of LLMs focused on worker well-being and autonomy, avoiding stress-inducing features and promoting engagement and learning.

Despotovic et al. [37] highlight a trend in job seekers preferring roles that incorporate advanced AI, like ChatGPT, aligning with their identities and technological interests. This shift towards AI-interactive jobs indicates a broader preference for technologically engaging roles, impacting job satisfaction and autonomy. Complementarily, Wang et al. [162] explore the paradox of LLMs in the job market, noting their role in both creating new opportunities and obsoleting certain jobs. This dual effect is key in assessing worker autonomy and well-being, as LLM-driven automation fosters more creative roles and autonomy but also raises concerns about job security and satisfaction. Integrating insights from recent research, it's evident that while generative AI has the potential to automate menial tasks and foster more creative roles, it also poses challenges to worker autonomy, necessitating a balanced approach to AI deployment in the workplace [170].

4.3 Effect on Wages and Jobs

The influence of AI on labor, particularly wages and job security, is a subject of global debate. The discourse ranges from AI augmenting human capabilities and creating new jobs to concerns about task automation and worker displacement.

Hemmer et al. [62] suggest AI could enhance productivity by up to 40%, potentially leading to higher wages. In contrast, Chowdhury et al. [28] point to AI's role in both creating and displacing jobs, with an estimated 1.7 million new jobs and significant displacements in the U.S. by 2029. This highlights the necessity for proactive support for those affected by AI-driven changes. Chhibber et al. [25] discuss AI's impact on the gig economy, noting potential job losses in areas like customer service, balanced by new opportunities in AI-related fields. Pal et al. [123] foresee a growing LLM industry, suggesting increased job creation in AI R&D. Conversely, Wang et al. [162] anticipate the displacement of up to 800 million jobs globally by 2030, advocating for protective legislation and equitable AI benefit distribution. Ashktorab et al. [3] examine Human-AI collaboration in the gaming industry, where AI's role in automating tasks like game testing could lead to job enrichment in creative roles.

The integration of Large Language Models has diverse implications for the job market. Despotovic et al.'s [37] research reveals that job seekers show a preference for roles that resonate with their personal identity and involve frequent interaction with technologies like ChatGPT. This underscores a shift in job market dynamics and the necessity for businesses to adapt their recruitment strategies. Wang et al. [162] discuss the paradoxical effect of LLMs, highlighting both job creation and displacement across various sectors. This includes an examination of LLMs' influence on routine tasks, specialized professions, and broader economic and societal aspects. Walkowiak et al. [158] focus on the Australian workforce, quantifying the exposure to GenAI risks and potential job transformations, addressing privacy, cybersecurity, and ethical issues.

4.4 Data Privacy and Security

In the realm of human-AI collaboration, trust, particularly in data privacy and security, is foundational. Ezer et al. [43] emphasize the critical role of trust engineering, advocating for algorithmic transparency and dynamic management to establish mutual trust in complex, information-imperfect environments. Yin et al. [176] further this by proposing a privacy-focused, human-centric approach, leveraging logistic regression and differential privacy to strike a balance between privacy and utility, achieving notable classification accuracy. In medical imaging, Kaassis et al. [74] highlight the imperative for secure, privacy-preserving AI through federated learning, addressing the privacy-utilization dichotomy to advance patient care. The synergy of AI and cybersecurity, as explored by Admin et al. [136], showcases AI's potential to fortify cyber defenses, underscoring the role of machine and deep learning in countering complex cyber threats and safeguarding data integrity. Lepri et al. [89] advocate for a human-centric AI ethos, stressing the importance of privacy, accountability, and ethical decision-making in building secure, trustworthy AI systems.

Regulatory and privacy challenges for Large Generative AI Models (LGAIMs), including ChatGPT and GPT-4, demand targeted legal and technical solutions for safe use. Hacker et al. [57] recommend a regulatory framework for LGAIMs focusing on transparency, risk management, and equitable treatment, particularly for high-risk uses. Ullah et al. [153] propose PrivChatGPT, a model incorporating differential privacy for LLMs, highlighting the balance between privacy and utility. Gupta et al. [56] discuss cybersecurity risks with GenAI, emphasizing the need for ethical practices and secure defenses against potential misuse. De Angelis et al. [36] identify LLMs as vectors for misinformation, advocating for policies to combat the "AI-driven infodemic" in public health. Thapa and Adhikari [149] balance LLMs' benefits against misinformation risks in biomedical research, calling for strict validation. Sebastian [137] focuses on privacy in AI chatbots, suggesting federated learning and data minimization to protect user information, reflecting increasing privacy concerns among LLM users.

4.5 Trustworthy AI and Accountability

Initially, conventional methods in trustworthy AI emphasized the importance of ethical principles, human oversight, and transparency. Hou et al. [64] discuss these principles, stressing the need for secure systems that identify risks while remaining under human control. Pflanzer et al. [126] introduce the Agent-Deed-Consequence model as a traditional ethical algorithm framework, emphasizing human

oversight in AI ethics. Caldwell et al. [18] highlight factors influencing trust in AI in cooperative environments, such as multiplayer gaming, underscoring the importance of environmental factors and peer influence in conventional trust-building.

The advent of LLMs introduces new dimensions to these conventional methods. In healthcare, a case study using LLMs [162] showcases the balance between patient engagement and AI over-reliance risks, underscoring the advancements in data sensitivity and patient communication. Kim et al. [79] demonstrate how LLMs contribute to managing uncertainty in human-AI collaboration, proposing a trial-to-trial approach to effectively communicate and estimate uncertainty. Furthermore, LLMs influence the formation of mental models, as discussed by Bansal et al. [6]. They highlight LLMs' role in enhancing predictability and understanding of AI's error boundaries, crucial for developing accurate mental models for better human-AI collaboration.

4.6 Law and Public Policy

Traditionally, AI legal and policy frameworks have focused on ensuring transparency, accountability, fairness, and bias mitigation. Pioneering work by Jobin et al. [70], which analyzed AI guidelines globally, underscores the universal emphasis on ethical standards, transparency, and accountability. Similarly, research by Cath et al. [19] highlights the critical need for transparent and accountable AI, especially in decision-making processes. Stahl et al. [143] propose six guidelines for AI regulation: Transparency, Accountability, Bias Mitigation, Human Oversight, Data Protection, and Public Education.

The integration of Large Language Models (LLMs) into the AI landscape underscores the need for updated legal frameworks to address new challenges in data use and AI-generated content regulation. Bender et al. [12] examine the ethical implications of LLMs, focusing on data usage and biases. An accredited law partner [159] specializing in technology and intellectual property discusses U.S. court cases like GitHub Copilot's use of open-source code, highlighting the legal intricacies of data attribution and compliance. This situation prompts a reevaluation of the datasets used for AI training and the adaptation of laws, particularly concerning copyrights and patents. Ekenobi et al. [148] explore the legal impacts of deploying LLMs, with a focus on data protection and privacy. They commend ChatGPT's privacy measures as a step towards mitigating potential risks. Additionally, Marcos and Pullin's article [101] talks about the EU's GDPR challenges, emphasizing the balance between technological growth and privacy rights. This necessitates robust regulatory frameworks that ensure LLM compliance with strict data protection standards, highlighting the importance of new laws in safeguarding privacy while fostering innovation.

5 Applications

This section highlights the role of Human-AI (HAI) Teaming and Large Pre-trained Models (LPtM) in revolutionizing fields such as healthcare, autonomous vehicle domain, surveillance, gaming, education, and accessibility. It underscores the synergy between human expertise and AI's analytical power, emphasizing the importance of ethical AI development and user-centric design in maximizing benefits and inclusivity within society.

5.1 Healthcare

In the healthcare sector, the symbiosis of human-AI collaboration, especially in enhancing patient diagnosis accuracy and fostering trust, underscores AI's transformative influence [63, 11, 104]. Advances such as the application of AI in breast cancer detection have markedly boosted diagnostic efficacy, sometimes even outperforming human experts, showcasing the remarkable potential of AI in healthcare [102]. These developments necessitate reevaluating AI's role in patient care, addressing ethical considerations and practical challenges [16]. Research focusing on healthcare practitioners' views within the US highlights the critical factors of training, trustworthiness, and perceived AI risks, which are pivotal in its acceptance and application in clinical environments [27]. Additionally, improvements in AI-powered medical image retrieval systems have been shown to augment both diagnostic capabilities and user trust, advocating for integrated systems that blend human insight with algorithmic accuracy [17]. This partnership between AI technologies and healthcare professionals is

vital for enhancing patient care outcomes and alleviating professional strain, albeit confronted by hurdles related to training and the accountability of AI systems.

Integrating Large Pre-trained Models such as ChatGPT into healthcare has shown promise in enhancing medical education, patient communication, and clinical decision-making. Lyu et al. [99] demonstrated ChatGPT's ability to simplify radiology reports for patients and healthcare providers, highlighting the potential for improved clinical education despite noting areas for accuracy and specificity enhancements. Kung et al. [86] found that ChatGPT scored at or near the passing threshold on the USMLE, suggesting its capability to support medical education and hinting at future roles in clinical settings. Johnson et al. [71] further validated ChatGPT's usefulness by showing its generally accurate responses to a range of medical questions, though recognizing the need for further refinement.

5.2 Autonomous Vehicles

Advancements in autonomous driving promise safer, AI-enhanced transportation, with significant milestones achieved in vehicle sensing and decision-making [4]. Key to these advancements is understanding Human-AI collaboration, particularly during critical transitions between automated and manual driving modes, as explored by Lv et al. [97]. This research highlights the importance of seamless transitions for safety, showing improvements in control and stability with a new hybrid driving model. Additionally, establishing trust in AI through transparent decision-making processes under uncertainty is crucial, as Jianlong et al. [180] discuss with a novel framework aimed at enhancing Human-AI collaboration. Fuchs et al. [48] further examine the managerial aspects of task delegation between humans and AI, proposing reinforcement learning to optimize collaboration and improve system performance.

Building on foundational studies, recent explorations into large language models (LLMs) signal a transformative era in autonomous driving technologies. Yang et al. [174] underscore the importance of LLMs for interpreting user commands in complex scenarios, emphasizing the necessity of model quality and prompt design. Park et al. [124] introduce VLAAD, a multi-modal LLM that demonstrates proficient interpretive capabilities across varied driving situations, showcasing the potential of natural language models in enhancing decision-making processes. Cui et al. [35] propose a framework that integrates LLMs for improved decision-making in autonomous vehicles, aiming for personalized assistance and efficient operation. Wen et al. [167] highlight the role of Visual Language Models (VLMs) like GPT-4V in scene understanding and decision-making, noting superior performance but also pointing out existing challenges. Additionally, Li et al. [91] tackle the challenge of generating realistic multi-view video data for autonomous driving, further supporting the role of LLMs in advancing the field. Together, these studies underscore the transformative potential of LLMs in advancing autonomous vehicle technologies, from improving interpretive capabilities to enhancing decision-making processes and data generation.

5.3 Surveillance and Security

The advancements in vision-based AI and IoT-connected cameras have heightened the potential for AI in cybersecurity [125], including in Cyber Security Incident Response Teams (CSIRTs) [78]. Determining the optimal level of AI autonomy in these settings is critical. A study by Hauptman et al. [60] explored this through a mixed-method approach, assessing AI's role across the CSIRT cycle—preparation, identification, containment, eradication, and recovery. Surveying 103 cybersecurity professionals and conducting interviews with 22, the study found that higher AI autonomy is preferred during the identification phase, while lower autonomy suits high-risk phases better. This suggests that adaptive AI, with proper team training, can integrate effectively into CSIRTs, enhancing teamwork and response efficiency. Furthermore, the real-time update preference and successful integration of human and machine intelligence in crowd-powered camera sensing, as noted by Guo et al. [55], underscore the potential for enhancing cybersecurity measures through improved situational awareness and resource management.

Moreover, the utilization of LLMs for video surveillance analysis, as proposed by Chen et al. [23], with the VideoLLM framework, demonstrates the potential for transferring LLMs' understanding and reasoning capabilities to video understanding tasks, offering a unified approach for various video-based security challenges. Additionally, Jain's [68] insights into integrating ChatGPT with video

security underscore the transformative impact of foundation models on physical security systems, enhancing operational efficiency and surfacing new insights for businesses.

5.4 Games

Games serve as a dynamic platform for interaction, typically involving multiple human players engaging through various forms of competition or collaboration, each game offering unique goals and gameplay mechanics. The essence of these interactions often relies on the expressiveness unique to human participants, critical for optimal play. However, the advent of advanced Large Language Models (LLMs), such as ChatGPT, has expanded the potential for AI to mimic this human expressiveness, opening new avenues for AI to compete alongside or against humans in a broader array of gaming contexts. This segment delves into the integration of LLM-driven AIs as alternatives to human players within games, alongside their application in generating content for video games, marking a significant evolution in game design and interaction.

5.4.1 LLM-Based AI as Players

Artificial intelligence has long been utilized in video games, from serving as basic adversaries in titles like Mario Kart [145] to mastering complex games such as chess, demonstrating AI's algorithmic prowess. Yet, these AIs often fell short as complete substitutes for human players due to their lack of flexibility and limited language processing capabilities. The advent of large language models (LLMs) has shifted this landscape, offering AIs that can engage with humans in more expressive ways, suitable across a broader spectrum of games.

Research into LLMs as game players, while in its nascent stages, has yielded intriguing insights. For instance, Frans et al. [46] showcase "AI Charades" using GPT-2 to illustrate LLMs' potential in games requiring deep language interaction, highlighting their structured uncertainty—an attribute highly valued in games traditionally reliant on human partners. Xu et al.'s [173] exploration into LLMs within the social deduction game Werewolf demonstrates these AIs' capability for strategic play, underscoring the necessity for advanced language understanding and social intuition. However, challenges such as limited input lengths and the impracticality of fine-tuning LLMs for specific games highlight current limitations.

Further critical examination by Sobieszek and Price [142] raises concerns over LLMs' reliability in information delivery, emphasizing the risk of information pollution due to GPT-3's occasionally untruthful responses. Research evaluating GPT models in social and moral contexts reveals mixed results, with models showing promise in single-player strategies but struggling in cooperative, ethical scenarios [2]. This suggests a need for further development in LLMs' ability to navigate complex social interactions and ethical decision-making. Despite these challenges, the expressive capabilities of LLMs compared to earlier AI models mark a significant advancement in human-computer interaction within games.

5.4.2 LLMs for Content Generation in Games

Large Language Models have significantly impacted game development, from generating complex game levels as demonstrated by Todd et al.[150], to enriching RPGs with dynamic quest narratives through advancements shown by Vartinens et al. [156]. These models enhance game narratives and environment design, also facilitating interactive tutorials and support systems, a concept explored by [46] with Language Model Games. Despite their potential, challenges persist, including ensuring content quality and coherence, addressing ethical concerns, and maintaining player immersion. The instructRL framework by Hu et al. [65] offers insights into fine-tuning LLM integration for effective human-AI interaction in games. Advancing LLM applications in gaming requires addressing these hurdles through hybrid approaches and ethical guidelines.

5.5 Education

The integration of AI in education is revolutionizing teaching and learning by enabling personalized learning experiences through the analysis of student performance and real-time feedback. This advancement fosters increased student engagement and improved learning outcomes, yet raises concerns about data privacy and the transparency of AI decisions.

Intelligent Tutoring Systems (ITSs), such as Carnegie Learning’s Mika platform, demonstrate AI’s ability to tailor education to individual needs, showing promising results like higher exam scores and reduced dropout rates [118, 155]. However, the opacity of AI’s decision-making has sparked debates over potential biases [168]. Despite these challenges, companies like Century Tech and Fishtree have successfully incorporated AI to enhance personalized learning experiences [76]. Virtual learning agents, including Georgia Tech’s chatbot Jill Watson, have been developed to supplement teaching by handling administrative queries or facilitating peer collaboration, albeit with risks of fostering parasocial relationships with digital avatars [40, 59]. Moreover, automated essay scoring systems offered by EdX and Pearson exemplify AI’s efficiency in grading, aligning closely with human grader assessments while saving time and costs [38]. Nonetheless, there remains skepticism about AI’s ability to fully grasp the subtleties of written expressions, highlighting the ongoing need for advancements in natural language processing.

The use of LLMs in education is poised to personalize learning experiences further. Extance [42] explores the practical deployment of AI chatbots as interactive learning tools, demonstrating their potential to enhance the teaching and learning dynamic significantly. However, Milano et al. [141] raise critical points regarding the operational, ethical, and environmental considerations of LLMs in educational settings, emphasizing the need for sustainable and transparent practices. These insights offer a deeper look at AI’s role in education, balancing technological innovation with mindful consideration of its broader impacts [161, 61, 22].

5.6 Accessibility

Assistive technologies powered by artificial intelligence (AI) are revolutionizing support for individuals with sensory and physical disabilities, offering innovative solutions for improved mobility and communication. A notable advancement is a deep learning-based navigation assistant for the visually impaired, which employs computer vision to facilitate obstacle detection and navigation with remarkable accuracy, as demonstrated in the work by [85]. Similarly, the development of communication aids for the deaf, utilizing audio-visual recognition algorithms, provides an automated solution that enhances daily interactions, a concept explored by [122]. These technologies exemplify the transformative potential of AI in creating more inclusive environments, with significant improvements in user independence and quality of life. Further contributions include a compact, low-cost visual aid system for blind individuals that integrates advanced image processing for obstacle avoidance and object detection [77], and an AI-enabled sign language recognition system that supports sentence-level communication for the hearing/speech impaired [166]. Additionally, the deployment of service robots to assist the elderly in shopping tasks demonstrates the practical applications of AI in facilitating daily activities [51].

In addition to these technological innovations, recent studies highlight the opportunities and challenges presented by large language models in promoting accessibility. Taheri et al.[147] explore the potential of text-to-image generation models to revolutionize accessible art creation, demonstrating how generative text models can reduce typing effort and make visual art creation more accessible to people with diverse abilities. In contrast, Gadiraju et al.[49] shed light on the subtle yet harmful biases LLMs can perpetuate towards disabled communities, emphasizing the importance of incorporating disability-positive resources in training data to foster a more inclusive digital environment. These insights underscore the evolving landscape of AI-driven accessibility tools, stressing the need for continuous refinement and inclusivity in their development.

6 Conclusion

This survey on Human-AI (HAI) Teaming, with a focus on Large Pre-trained Models (LPTM), underscores a notable evolution in collaborative problem-solving and decision-making. Integrating human expertise into AI model training has enhanced performance and fostered intuitive and effective human-AI interactions. Safety, security, and trust emerge as critical components, requiring a balanced approach between technological progress and ensuring privacy and fairness. The survey’s exploration across various sectors, including healthcare, autonomous vehicles, and education, showcases HAI’s versatility in addressing unique challenges and enhancing societal benefits. As we advance, the integration of human intelligence with LPTMs looks like a promising domain in unlocking innovative solutions and improving overall decision-making. The future of HAI Teaming will hinge on navigating

its challenges responsibly, ensuring ethical alignment, and maximizing its societal impact. This research contributes to the academic dialogue, offering insights for future exploration, policy-making, and strategic development towards a collaborative, effective, and equitable HAI ecosystem.

References

- [1] Aakash Ahmad, Muhammad Waseem, Peng Liang, Mahdi Fahmideh, Mst Shamima Aktar, and Tommi Mikkonen. Towards human-bot collaborative software architecting with ChatGPT. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, pages 279–285, 2023.
- [2] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*, 2023.
- [3] Zahra Ashktorab, Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. Human-AI collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 4:1–20, 10 2020.
- [4] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561*, 12 2021.
- [5] Arun Kumar AV, Santu Rana, Alistair Shilton, and Svetha Venkatesh. Human-AI collaborative bayesian optimisation. *Advances in Neural Information Processing Systems*, 35:16233–16245, 2022.
- [6] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, pages 2–11, 2019.
- [7] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.
- [8] Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. Deriving machine attention from human rationales. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [9] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Trans. Graph.*, 38(4), jul 2019.
- [10] Guanqun Bi, Lei Shen, Yuqiang Xie, Yanan Cao, Tiangang Zhu, and Xiaodong He. A group fairness lens for large language models. *arXiv preprint arXiv:2312.15478*, 2023.
- [11] Nadine Bienefeld, Michaela Kolbe, Giovanni Camen, Dominic Huser, and Philipp Karl Buehler. Human-AI teaming: leveraging transactive memory and speaking up for enhanced team effectiveness. *Frontiers in Psychology*, 14, 2023.
- [12] Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D West, Qiong Zhang, et al. How should the advent of large language models affect the practice of science? *arXiv preprint arXiv:2312.03759*, 2023.
- [13] Matthias Braun, Hannah Bleher, and Patrik Hummel. A leap of faith: Is there a formula for “trustworthy” AI? *The Hastings Center Report*, 51:17–22, 2021.
- [14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20, Red Hook, NY, USA, 2020. Curran Associates Inc.

- [15] Yang Li Bryan Wang, Gang Li. Enabling conversational interaction with mobile UI using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM, New York, NY, USA. Accessed: 2023-11-28.
- [16] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, jul 2021.
- [17] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14, 2019.
- [18] Sabrina Caldwell, Penny Sweetser, Nicholas O'Donnell, Matthew J Knight, Matthew Aitchison, Tom Gedeon, Daniel Johnson, Margot Brereton, Marcus Gallagher, and David Conroy. An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(3):1–36, 2022.
- [19] Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. Artificial intelligence and the ‘good society’: the US, EU, and UK approach. *Science and engineering ethics*, 24:505–528, 2018.
- [20] Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianne Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724*, 2023.
- [21] Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, and Yu Zhang. AI challenges in human-robot cognitive teaming. *arXiv preprint arXiv:1707.04775*, 2017.
- [22] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.
- [23] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, and Limin Wang. VideoLLM: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023.
- [24] Yuhao Chen, Chloe Wong, Hanwen Yang, Juan Aguenza, Sai Bhujangari, Benthan Vu, Xun Lei, Amisha Prasad, Manny Fluss, Eric Phuong, Minghao Liu, Raja Kumar, Vanshika Vats, and James Davis. Assessing the impact of prompting methods on chatgpt’s mathematical capabilities. *arXiv preprint arXiv:2312.15006*, 2024.
- [25] Nalin Chhibber, Joslin Goh, and Edith Law. Teachable conversational agents for crowdwork: Effects on performance and trust. *Proceedings of the ACM on Human-Computer Interaction*, 6:1–21, 2022.
- [26] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.
- [27] Avishek Choudhury and Onur Asan. Impact of accountability, training, and human factors on the use of artificial intelligence in healthcare: Exploring the perceptions of healthcare practitioners in the US. *Human Factors in Healthcare*, 2:100021, 2022.
- [28] Soumyadeb Chowdhury, Pawan Budhwar, Prasanta Kumar Dey, Sian Joel-Edgar, and Amelie Abadie. AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework. *Journal of Business Research*, 144:31–49, 2022.
- [29] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307, 2017.
- [30] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. All that’s human’is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*, 2021.

- [31] Lei Clark, Abdulmalik Ofemile, Svenja Adolphs, and Tom Rodden. A multimodal approach to assessing user experiences with agent helpers. *ACM Transactions on Interactive Intelligent Systems*, 2016.
- [32] William Clark, Jan Golinski, and Simon Schaffer. *The Sciences in Enlightened Europe*. University of Chicago Press, 1999.
- [33] Nicholas Conlon, Nisar Ahmed, and Daniel Szafir. A survey of algorithmic methods for competency self-assessments in human-autonomy teaming. *ACM Computing Surveys*, 08 2023.
- [34] António Correia, Andrea Grover, Daniel Schneider, Ana Paula Pimentel, Ramon Chaves, Marcos Antonio de Almeida, and Benjamim Fonseca. Designing for hybrid intelligence: A taxonomy and survey of crowd-machine interaction. *Applied Sciences*, 13, 2023.
- [35] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 902–909, January 2024.
- [36] Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120, 2023.
- [37] Petar Despotovic and Yevgen Bogodistov. Does chatGPT alter job seekers’ identity? an experimental study. *Proceedings of the 57th Hawaii International Conference on System Sciences*, 2024.
- [38] Semire Dikli. Automated essay scoring. *Turkish Online Journal of Distance Education*, 7(1):49–62, 2006.
- [39] Alpana Dubey, Kumar Abhinav, Sakshi Jain, Veenu Arora, and Asha Puttaveerana. HACO: A framework for developing human-AI teaming. In *Proceedings of the 13th Innovations in Software Engineering Conference on Formerly Known as India Software Engineering Conference*, ISEC 2020, New York, NY, USA, 2020. Association for Computing Machinery.
- [40] Bobbie Eicher, Lalith Polepeddi, and Ashok Goel. Jill watson doesn’t care if you’re pregnant: Grounding AI ethics in empirical studies. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 88–94, 2018.
- [41] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.
- [42] Andy Extance. ChatGPT has entered the classroom: how LLMs could transform education. *Nature*, 2023.
- [43] Neta Ezer, S. Bruni, Yang Cai, S. Hepenstal, C. Miller, and D. Schmorow. Trust engineering for human-AI teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63:322 – 326, 2019.
- [44] Mingming Fan, Xianyou Yang, TszyTung Yu, Q Vera Liao, and Jian Zhao. Human-AI collaboration for UX evaluation: Effects of explanation and synchronization. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–32, 2022.
- [45] Christopher Flathmann, Beau Schelble, and Nathan McNeese Rui Zhang. Modeling and guiding the creation of ethical human-ai team. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.
- [46] Kevin Frans. AI charades: Language models as interactive game environments. In *2021 IEEE Conference on Games (CoG)*, pages 1–2. IEEE, 2021.
- [47] Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111:8410 – 8415, 2014.
- [48] Andrew Fuchs, Andrea Passarella, and Marco Conti. Compensating for sensing failures via delegation in human-AI hybrid systems. *Sensors*, 23(7):3409, 2023.

- [49] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. "I wouldn't say offensive but...": Disability-centered perspectives on large language models. In *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 205–216, New York, NY, USA, 2023. Association for Computing Machinery.
- [50] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.
- [51] Mohammed Ghazal, Maha Yaghi, Abdalla Gad, Gasm El Bary, Marah Alhalabi, Mohammad Alkhedher, and Ayman S. El-Baz. AI-powered service robotics for independent shopping experiences by elderly and disabled people. *Applied Sciences*, 11(19):9007, September 2021.
- [52] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
- [53] D. Gopinath, J. DeCastro, G. Rosman, E. Sumner, A. Morgan, S. Hakimi, and S. Stent. Hmiway-env: A framework for simulating behaviors and preferences to support human-AI teaming in driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [54] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.
- [55] Anhong Guo, Anuraag Jain, Shomiron Ghose, Gierad Laput, Chris Harrison, and Jeffrey P. Bigham. Crowd-AI camera sensing in the real world. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3), sep 2018.
- [56] Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. From chatGPT to threatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, 2023.
- [57] Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1112–1123, 2023.
- [58] Stefan Harrer. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90, 2023.
- [59] Luminita Hartle and Tara Kaczorowski. The positive aspects of mursion when teaching higher education students. *Quarterly Review of Distance Education*, 20(4):71–100, 2019.
- [60] Allyson I Hauptman, Beau G Schelble, Nathan J McNeese, and Kapil Chalil Madathil. Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. *Computers in Human Behavior*, 138:107451, 2023.
- [61] Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. Exploring the responses of large language models to beginner programmers' help requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1*, ICER '23, page 93–105, New York, NY, USA, 2023. Association for Computing Machinery.
- [62] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. Human-AI collaboration: The effect of AI delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 453–463, New York, NY, USA, 2023. Association for Computing Machinery.
- [63] Katharine E Henry, Rachel Kornfield, Anirudh Sridharan, Robert C Linton, Catherine Groh, Tony Wang, Albert Wu, Bilge Mutlu, and Suchi Saria. Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. *NPJ digital medicine*, 5(1):97, 2022.
- [64] Keke Hou, Tingting Hou, and Cai Lili. Exploring trust in human-AI collaboration in the context of multiplayer online games. systems. *Human–AI Teaming: Synergy, Decision-Making and Interdependency*, 2023.
- [65] Hengyuan Hu and Dorsa Sadigh. Language instructed reinforcement learning for human-AI coordination. *arXiv preprint arXiv:2304.07297*, 2023.

- [66] Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. Bias assessment and mitigation in LLM-based code generation. *arXiv preprint arXiv:2309.14345*, 2023.
- [67] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*, page 8022–8034, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [68] Ashesh Jain. ChatGPT meets video security: A new era of intelligent surveillance, 2023.
- [69] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022.
- [70] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.
- [71] Douglas B Johnson, Rachel S Goodman, J Randall Patrinely, Cosby A Stone, Eli Zimmerman, Rebecca Rigel Donald, Sam S Chang, Sean T Berkowitz, Avni P Finn, Eiman Jahangir, Elizabeth A Scoville, Tyler Reese, Debra E. Friedman, Julie A. Bastarache, Yuri F van der Heijden, Jordan Wright, Nicholas Carter, Matthew R Alexander, Jennifer H Choe, Cody A Chastain, John Zic, Sara Horst, Isik Turker, Rajiv Agarwal, Evan C. Osmundson, Kamran Idrees, Colleen M. Kiernan, Chandrasekhar Padmanabhan, Christina Edwards Bailey, Cameron Schlegel, Lola B. Chambliss, Mike Gibson, Travis J. Osterman, and Lee Wheless. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the chat-GPT model. *Research Square*, 2023.
- [72] Matthew Johnson and Alonso Vera. No AI is an island: the case for teaming intelligence. *AI magazine*, 40(1):16–28, 2019.
- [73] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- [74] Georgios A Kaassis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.
- [75] Frank Kaptein, Joost Broekens, Koen V Hindriks, and Mark Neerincx. CAAF: A cognitive affective agent programming framework. *Engineering Psychology and Cognitive Ergonomics*, 10906:317–330, 2016.
- [76] Vinothini Kasinathan, Aida Mustapha, and Imran Medi. Adaptive learning system for higher learning. In *2017 8th international conference on information technology (ICIT)*, pages 960–970. IEEE, 2017.
- [77] Muiz Ahmed Khan, Pias Paul, Mahmudur Rashid, Mainul Hossain, and Md Atiqur Rahman Ahad. An AI-based visual aid with integrated reading assistant for the completely blind. *IEEE Transactions on Human-Machine Systems*, 50(6):507–517, 2020.
- [78] Georgia Killcrece, Klaus-Peter Kossakowski, Robin Ruefle, and Mark Zajicek. State of the practice of computer security incident response teams (CSIRTs). Technical Report CMU/SEI-2003-TR-001, Carnegie Mellon University, Software Engineering Institute’s Digital Library, Oct 2003. Accessed: 2024-Feb-13.
- [79] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "help me help the AI": Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.
- [80] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
- [81] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34:2611–2624, 2021.

- [82] Konstantinos Konstantis, Antonios Georgas, Antonis Faras, Konstantinos Georgas, and Aristotle Tympas. Ethical considerations in working with chatGPT on a questionnaire about the future of work with chatGPT. *AI and Ethics (Online)*, 2023.
- [83] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA, 2023. Association for Computing Machinery.
- [84] Steve WJ Kozlowski and Daniel R Ilgen. Enhancing the effectiveness of work groups and teams, psychological science in the public interest. *Psychological Science in the Public Interest*, 2006.
- [85] Nitin Kumar and Anuj Jain. A deep learning based model to assist blind people in their navigation. *Journal of Information Technology Education: Innovations in Practice*, 21:095–114, 2022.
- [86] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.
- [87] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF: Scaling reinforcement learning from human feedback with AI feedback, 2023.
- [88] Séverin Lemaignan, Mathieu Warnier, E. Akin Sisbot, Aurélie Clodic, and Rachid Alami. Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence*, 247:45–69, 2017.
- [89] Bruno Lepri, Nuria Oliver, and Alex Pentland. Ethical machines: The human-centric use of artificial intelligence. *IScience*, 24(3), 2021.
- [90] Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. Human-in-the-loop debugging deep text classifiers. *arXiv preprint arXiv:2010.04987*, abs/2010.04987, 2020.
- [91] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023.
- [92] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023.
- [93] Minghao Liu, Zeyu Cheng, Shen Sang, Jing Liu, and James Davis. Tag-based annotation creates better avatars. *arXiv preprint arXiv:2302.07354*, 2023.
- [94] Minghao Liu, Jiaheng Wei, Yang Liu, and James Davis. Do humans and machines have the same eyes? human-machine perceptual differences on image classification. *arXiv preprint arXiv:2304.08733*, 2023.
- [95] Jinghui Lu, Linyi Yang, Brian Mac Namee, and Yue Zhang. A rationale-centric framework for human-in-the-loop machine learning. In *Annual Meeting of the Association for Computational Linguistics*, 2022.
- [96] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. Readingquizmaker: A human-NLP collaborative system that supports instructors to design high-quality reading quiz questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.
- [97] Chen Lv, Yutong Li, Yang Xing, Chao Huang, Dongpu Cao, Yifan Zhao, and Yahui Liu. Human-machine collaboration for automated driving using an intelligent two-phase haptic interface. *Advanced Intelligent Systems*, page 2000229, 02 2021.
- [98] Joseph B Lyons, Kevin T Wynne, Sean Mahoney, and Mark A Roebke. Trust and human-machine teaming: A qualitative study. In *Artificial intelligence for the internet of everything*, pages 101–116. Elsevier, 2019.
- [99] Qing Lyu, Josh Tan, Mike E Zapadka, Janardhana Ponnatapuram, Chuang Niu, Ge Wang, and Christopher T Whitlow. Translating radiology reports into plain language using chatGPT and GPT-4 with prompt learning: Promising results, limitations, and potential. *arXiv preprint arXiv:2303.09038*, 2023.

- [100] Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. A review on human–AI interaction in machine learning and insights for medical applications. *International Journal of Environmental Research and Public Health*, 18, 2021.
- [101] Henrique Marcos and Melina Pullin. Large language models and eu data protection: Mapping (some) of the problems. <https://digi-con.org/large-language-models-and-eu-data-protection-mapping-some-of-the-problems/>, 2023. Accessed: 2024-02-28.
- [102] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.
- [103] Nathan J. McNeese, Beau G. Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. Who/what is my teammate? team composition considerations in human–AI teaming. *IEEE Transactions on Human-Machine Systems*, 51(4):288–299, 2021.
- [104] Lucas Memmert and Eva Bittner. Complex problem solving through human-ai collaboration: Literature review on research contexts. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2022.
- [105] Lucas Memmert and Navid Tavanapour. Towards human-AI-collaboration in brainstorming: Empirical insights into the perception of working with a generative ai. In *31st European Conference on Information Systems (ECIS)*, 2023.
- [106] Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20, 2023.
- [107] Henny Admoni Michelle Zhao, Reid Simmons. The role of adaptation in collective human–AI teaming. *Topics in Cognitive Science*, 2022.
- [108] Fan Mingming, Yang Xianyou, Yu Tsz Tung, Q. Liao Vera, and Zhao Jian. Human-AI collaboration for ux evaluation: Effects of explanation and synchronization. *arXiv.org*, 2023. Accessed: 2023-11-28.
- [109] Suvir Mirchandani, F. Xia, Peter R. Florence, Brian Ichter, Danny Driess, Montse Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023.
- [110] Brent Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, , and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data and Society*, 2016.
- [111] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. Evaluating the impact of human explanation strategies on human-AI visual decision-making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023.
- [112] I. Munyaka, Z. Ashktorab, C. Dugan, J. Johnson, and Q. Pan. Decision making strategies and team efficacy in human-AI teams. *Proceedings of the ACM on Human-Computer Interaction*, 2023.
- [113] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [114] Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurrian van Diggelen. Using perceptual and cognitive explanations for enhanced human-agent team performance. *Lecture Notes in Computer Science*, 10011:204–214, 2018.
- [115] An Ngo, Daniel Phelps, Derrick Lai, Thanyared Wong, Lucas Mathias, Anish Shivamurthy, Mustafa Ajmal, Minghao Liu, and James Davis. Tag-based annotation for avatar face creation. *arXiv preprint arXiv:2308.12642*, 2023.
- [116] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36:618–634, 2017.

- [117] Wei Niu, Zhenglun Kong, Geng Yuan, Weiwen Jiang, Jexiong Guan, Caiwen Ding, Pu Zhao, Sijia Liu, Bin Ren, and Yanzhi Wang. Real-time execution of large-scale language models on mobile, 2020.
- [118] Hyacinth S Nwana. Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, 4(4):251–277, 1990.
- [119] Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. Likelihood-based mitigation of evaluation bias in large language models. *arXiv preprint arXiv:2402.15987*, 2024.
- [120] OpenAI. Chatgpt-4. <https://openai.com/chatgpt>, 2023.
- [121] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [122] Saket Ozarkar, Raj Chetwani, Sugam Devare, Sumeet Haryani, and Nupur Giri. AI for accessibility: Virtual assistant for hearing impaired. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7, 2020.
- [123] Subharun Pal. The future of large language models: A futuristic dissection on AI and human interaction. *IJFMR-International Journal For Multidisciplinary Research*, 5(3), 2023.
- [124] SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. VLAAD: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 980–987, January 2024.
- [125] Armin Danesh Pazho, Christopher Neff, Ghazal Alinezhad Noghre, Babak Rahimi Ardabili, Shanle Yao, Mohammadreza Baharani, and Hamed Tabkhi. Ancilia: Scalable intelligent video surveillance for the artificial intelligence of things. *arXiv preprint arXiv:2301.03561*, 2023.
- [126] Michael Pflanzer, Zach Traylor, Joseph B. Lyons, Veljko Dubljević, and Chang S. Nam. Ethics in human–AI teaming: principles and perspectives. *AI and Ethics*, pages 1–19, 2022.
- [127] Joris Pries, Sandjai Bhulai, and Rob D. van der Mei. Active pairwise distance learning for efficient labeling of large datasets by human experts. *Applied Intelligence*, 53:24689 – 24708, 2023.
- [128] Zheng Qingxiao, Tang Yiliu, Liu Yiren, Liu Weizi, and Huang Yun. UX research on conversational human-AI interaction: A literature review of the ACM digital library. *arXiv preprint arXiv*, 2023. Accessed: 2023-11-28.
- [129] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.
- [130] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.
- [131] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.
- [132] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. Supporting human-AI collaboration in auditing LLM with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 913–926, 2023.
- [133] Rowena Rodrigues. Legal and human rights issues of ai: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, 2020.
- [134] Stuart J. Russell, Peter Norvig, Ming-Wei Chang, Jacob Devlin, Anca Dragan, David Forsyth, Ian Goodfellow, Jitendra Malik, Vikash Mansinghka, Judea Pearl, and Michael J Wooldridge. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2021.

- [135] Eduardo Salas, Carolyn Prince, David P Baker, and Lisa Shrestha. Situation awareness in team performance: Implications for measurement and training. *Human Factors*, 37, 2017.
- [136] SP Samyuktha, P Kavitha, VA Kshaya, P Shalini, and R Ramya. A survey on cyber security meets artificial intelligence: AI–driven cyber security. *Journal of Cognitive Human-Computer Interaction*, 2(2):50–55, 2022.
- [137] Glorin Sebastian. Privacy and data protection in chatgpt and other AI chatbots: Strategies for securing user information. Available at SSRN 4454761, 2023.
- [138] Burr Settles. *Active Learning Literature Survey*. Department of Computer Sciences, University of Wisconsin-Madison, 2009.
- [139] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition. In *Proceedings of the ACM Web Conference 2022*, WWW ’22, page 882–892, New York, NY, USA, 2022. Association for Computing Machinery.
- [140] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, 2023.
- [141] Sabina Leonelli Silvia Milano, Joshua A. McGrane. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 2023.
- [142] Adam Sobieszek and Tadeusz Price. Playing games with AIs: the limits of gpt-3 and similar large language models. *Minds and Machines*, 32(2):341–364, 2022.
- [143] B.C. Stahl, A. Andreou, P. Brey, T. Hatzakis, A. Kirichenko, K. Macnish, S. Laulhé Shaelou, A. Patel, M. Ryan, and D. Wright. Artificial intelligence for human flourishing – beyond principles for machine learning. *Journal of Business Research*, 2021.
- [144] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.
- [145] Nintendo Switch. MarioKart 8 deluxe, 2017. Video game.
- [146] Aaquib Tabrez, Matthew Luebbers, and Bradley Hayes. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports*, 1, 12 2020.
- [147] Atieh Taheri, Mohammad Izadi, Gururaj Shiram, Negar Rostamzadeh, and Shaun Kane. Breaking barriers to creative expression: Co-designing and implementing an accessible text-to-image interface. *arXiv preprint arXiv:2309.02402*, 2023.
- [148] Ekenobi ThankGod Chinonso. The impact of chatGPT on privacy and data protection laws. Available at SSRN: <https://ssrn.com/abstract=4556181>, 2023.
- [149] Surendrabikram Thapa and Surabhi Adhikari. ChatGPT, bard, and large language models for biomedical research: opportunities and pitfalls. *Annals of Biomedical Engineering*, 51(12):2647–2651, 2023.
- [150] Graham Todd, Sam Earle, Muhammad Umair Nasir, Michael Cerny Green, and Julian Togelius. Level generation through large language models. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, pages 1–8, 2023.
- [151] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [152] Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. Does human collaboration enhance the accuracy of identifying LLM-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pages 163–174, 2023.
- [153] Imdad Ullah, Najm Hassan, Sukhpal Singh Gill, Basem Suleiman, Tariq Ahamed Ahanger, Zawar Shah, Junaid Qadir, and Salil S Kanhere. Privacy preserving large language models: ChatGPT case study based vision and framework. *arXiv preprint arXiv:2310.12523*, 2023.
- [154] Karel van den Bosch, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx. Six challenges for human-AI co-learning. In *Adaptive Instructional Systems: First International*

- Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*, pages 572–589. Springer, 2019.
- [155] Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221, 2011.
 - [156] Susanna Värtinen, Perttu Hämäläinen, and Christian Guckelsberger. Generating role-playing game quests with GPT language models. *IEEE Transactions on Games*, 2022.
 - [157] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. Prevalence and prevention of large language model use in crowd work. *arXiv preprint arXiv:2310.15683*, abs/2310.15683, 2023.
 - [158] Emmanuelle Walkowiak and Trent MacDonald. Generative ai and the workforce: What are the risks? Available at SSRN, 2023.
 - [159] Dylan Walsh. Legal issues presented by generative AI. <https://mitsloan.mit.edu/ideas-made-to-matter/legal-issues-presented-generative-ai>, 2023. Accessed: 2024-02-28.
 - [160] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-AI collaboration in data science: Exploring data scientists’ perceptions of automated AI. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–24, 2019.
 - [161] Rose Wang and Dorottya Demszky. Is chatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *18th Workshop on Innovative Use of NLP for Building Educational Applications*, June 2023.
 - [162] Yifei Wang. The large language model (LLM) paradox: Job creation and loss in the age of advanced AI. TechRxiv Preprint, 2023. Available at TechRxiv: [URL].
 - [163] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.
 - [164] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint arXiv:2302.10205*, 2023.
 - [165] Justin D. Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I. Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. Perfection not required? human-AI partnerships in code translation. In *26th International Conference on Intelligent User Interfaces, IUI ’21*, page 402–412, New York, NY, USA, 2021. Association for Computing Machinery.
 - [166] Feng Wen, Zixuan Zhang, Tianyi He, and Chengkuo Lee. AI enabled sign language recognition and vr space bidirectional communication using triboelectric smart glove. *Nature Communications*, 12:5378, 09 2021.
 - [167] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi Bai, Xinyu Cai, Min Dou, Shuanglu Hu, Botian Shi, and Yu Qiao. On the road with GPT-4V(ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*, 2023.
 - [168] Elizabeth Whitaker, Ethan Trewhitt, Matthew Holtsinger, Christopher Hale, Elizabeth Veinott, Chris Argenta, and Richard Catrambone. The effectiveness of intelligent tutoring on training in a video game. In *2013 IEEE International Games Innovation Conference (IGIC)*, pages 267–274. IEEE, 2013.
 - [169] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. The role and limits of principles in ai ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 195–200, 2019.
 - [170] Allison Woodruff, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. How knowledge workers think generative AI will (not) transform their industries. *arXiv preprint arXiv:2310.06778*, 2023.

- [171] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.
- [172] Tongshuang Sherry Wu, Michael Terry, and Carrie J. Cai. Ai chains: Transparent and controllable human-AI interaction by chaining large language model prompts. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2021.
- [173] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*, 2023.
- [174] Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. Human-centric autonomous systems with LLMs for user command reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 988–994, January 2024.
- [175] Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, and Dakuo Wang. Beyond labels: Empowering human annotators with natural language explanations through a novel active-learning architecture. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11629–11643, Singapore, December 2023. Association for Computational Linguistics.
- [176] Chunyong Yin, Biao Zhou, Zhichao Yin, and Jin Wang. Local privacy protection classification based on human-centric computing. *Human-centric Computing and Information Sciences*, 9:1–14, 2019.
- [177] Lihan Zha, Yuchen Cui, Li-Heng Lin, Minae Kwon, Montse Gonzalez Arenas, Andy Zeng, Fei Xia, and Dorsa Sadigh. Distilling and retrieving generalizable knowledge for robot manipulation via language corrections. *ArXiv*, abs/2311.10678, 2023.
- [178] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. Investigating AI teammate communication strategies and their impact in human-ai teams for effective teamwork. *Proceedings of the ACM on Human-Computer Interaction*, 7:1–31, 2023.
- [179] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. "An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–25, 2021.
- [180] Jianlong Zhou and Fang Chen. Towards trustworthy human-AI teaming under uncertainty. In *IJCAI 2019 workshop on explainable AI (XAI)*, 2019.
- [181] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.