Value Alignment and Trust in AI Systems: A Survey

www.surveyx.cn

Abstract

This survey explores the intricate relationship between artificial intelligence (AI) systems and human values, emphasizing the necessity of value alignment and trust. It underscores the importance of aligning AI systems with ethical standards to ensure safety and trustworthiness across domains like healthcare and autonomous vehicles. The survey discusses challenges such as the principal-agent problem, AI decision-making opacity, and encoding AI with diverse moral values. It evaluates frameworks supporting value alignment, including reinforcement learning and hybrid methods. Trust in AI systems is analyzed through transparency, explainability, and accountability, with mechanistic interpretability enhancing user confidence. Human-centered design and ethical frameworks are highlighted as critical for improving user experience and ensuring ethical AI operations. The survey identifies challenges and opportunities in human-AI interaction, advocating for interdisciplinary collaboration and emerging technologies to address these issues. It stresses the need for continuous monitoring and public discourse on advanced AI technologies, promoting proactive ethical approaches. In conclusion, aligning AI systems with human values is crucial for trust and safety, requiring stakeholders to address technical, ethical, and societal challenges to develop responsible AI systems aligned with moral standards.

1 Introduction

1.1 Overview of the Survey Structure

This survey comprehensively examines the intricate relationship between artificial intelligence (AI) systems and human values, emphasizing value alignment and trust. It begins with an introduction that underscores the necessity of aligning AI systems with human values and ethics to ensure safety and trustworthiness. The second section provides foundational background, defining key concepts such as value alignment, AI alignment, trust in AI systems, human-AI interaction, ethical AI, and autonomous systems.

The third section focuses on value alignment in AI, discussing its significance, challenges, and various approaches and frameworks for achieving alignment, including the integration of human preferences into AI systems. The fourth section analyzes trust in AI systems, highlighting the roles of transparency, explainability, and accountability, alongside the contribution of mechanistic interpretability.

Human-AI interaction is explored in the fifth section, where the dynamics of interaction are examined for their impact on value alignment and trust, emphasizing the design of AI systems that promote effective and safe human interaction. The sixth section addresses ethical AI and autonomous systems, analyzing ethical considerations and the role of frameworks and guidelines in preventing harm.

The survey concludes by identifying key challenges and potential future directions, stressing the importance of interdisciplinary collaboration and emerging technologies in overcoming these challenges. It synthesizes the main findings, highlighting the critical need for aligning AI systems with human values to foster trust and safety. Effective alignment necessitates a nuanced understanding of the

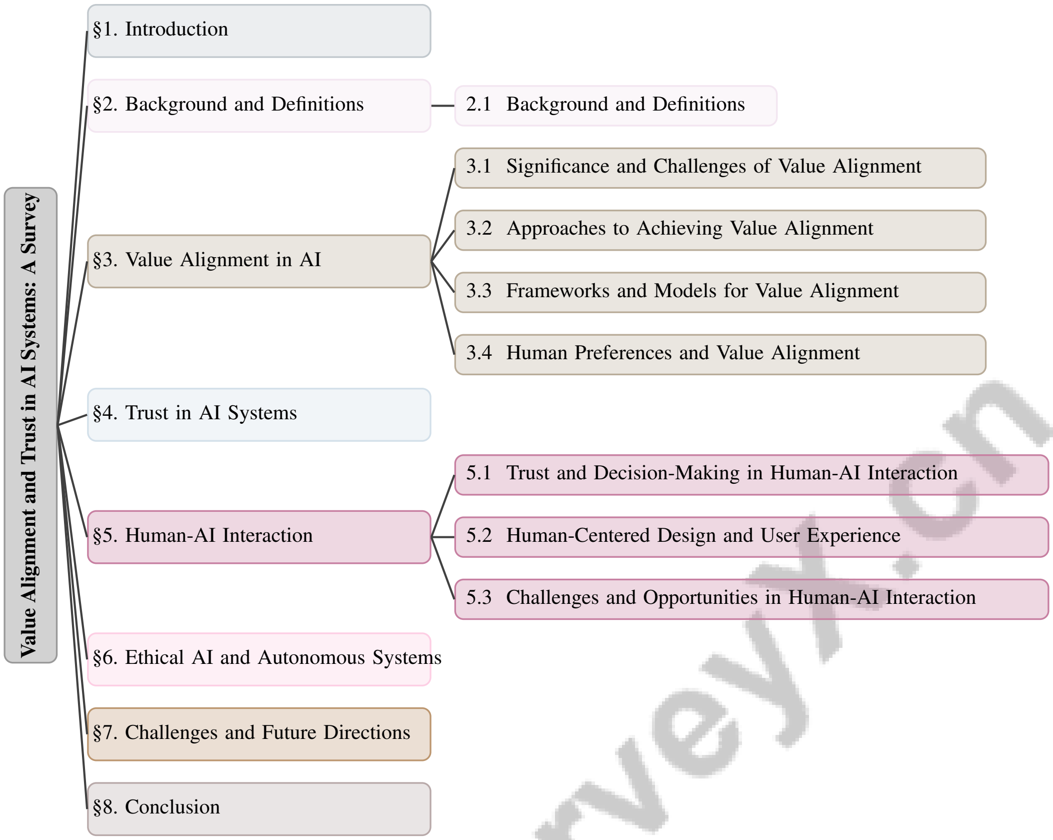
Value Alignment and Trust in AI Systems: A Survey
§1. Introduction
§2. Background and Definitions
2.1 Background and Definitions
§3. Value Alignment in AI
3.1 Significance and Challenges of Value Alignment
3.2 Approaches to Achieving Value Alignment
3.3 Frameworks and Models for Value Alignment
3.4 Human Preferences and Value Alignment
§4. Trust in AI Systems
§5. Human-AI Interaction
5.1 Trust and Decision-Making in Human-AI Interaction
5.2 Human-Centered Design and User Experience
5.3 Challenges and Opportunities in Human-AI Interaction
§6. Ethical AI and Autonomous Systems
§7. Challenges and Future Directions
§8. Conclusion

Figure 1: chapter structure

interplay between normative and technical aspects, ensuring that alignment goals encompass diverse human values, intentions, and preferences. A principle-based approach is essential for identifying fair alignment principles that resonate with various moral beliefs, thereby managing the risks of misalignment as AI capabilities expand through robust governance and assurance practices [1, 2].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Background and Definitions

The ethical development and deployment of artificial intelligence (AI) necessitate a nuanced understanding of its interdisciplinary nature, focusing on aligning AI systems with human values and ethical standards. This alignment involves addressing both normative and technical dimensions, specifying goals such as adherence to human instructions, intentions, or broader ethical principles, and ensuring fairness by accommodating diverse moral beliefs. A bidirectional alignment approach is crucial, ensuring AI systems reflect human values while society adapts to AI's evolving capabilities [3, 4, 5, 2]. Central themes include value alignment, AI alignment, trust in AI systems, human-AI interaction, ethical AI, and autonomous systems.

Value alignment involves aligning an agent's utility function with its principal's, a complex task due to information asymmetry and divergent values [6]. This challenge is acute in high-stakes areas like healthcare, where public trust in medical AI is essential [7]. The diversity of human values and the challenge of defining 'good' outcomes that align with ethical intuitions further complicate this task.

AI alignment extends the concept of value alignment by ensuring AI systems operate according to intended values, mitigating risks and societal harm [8]. This is especially critical in scenarios where AI decisions significantly impact human lives, highlighting the need for robust and secure systems [9].

AI-generated, for reference only.

Trust in AI systems is pivotal for societal acceptance and integration. It is established through reliability, safety, fairness, and explainability, which foster confidence in AI technologies [10]. However, overreliance on AI can erode trust, necessitating clear evaluation methods [11].

Human-AI interaction focuses on designing AI systems that promote safe and effective interactions, ensuring AI can comprehend and appropriately respond to human cues without undermining autonomy [12]. The human-centered artificial intelligence (HCAI) approach emphasizes aligning AI with human needs and capabilities to enhance productivity and well-being [13].

Ethical AI involves creating guidelines and principles to govern AI development, ensuring operations that do not cause harm. Addressing fairness, accountability, transparency, and ethics is crucial for aligning AI systems with human rights values [10]. Consideration of non-human entities and environmental well-being is also necessary to prevent ecological harm [13].

Autonomous systems, which operate with a degree of independence, raise governance and safety concerns. Understanding their mechanisms is vital to ensure alignment with human values and avert unintended risks [13]. The importance of addressing value alignment and coherent extrapolated volition is particularly notable in discussions about superintelligent AI systems [9].

Integrating ethical principles with a human rights-based approach provides a robust framework for addressing the challenges and opportunities presented by AI technologies. This integration emphasizes aligning AI development with universal human rights and offers concrete recommendations for fostering a "Good AI Society" [14, 15]. These principles are essential for developing AI systems that are safe, trustworthy, and aligned with human values, offering promising avenues for enhancing AI alignment with ethical standards.

In recent years, the alignment of artificial intelligence (AI) systems with human values has emerged as a critical area of research. This alignment is not merely a technical challenge but involves a complex interplay of ethical, societal, and technical considerations. To illustrate this multifaceted issue, Figure 2 provides a visual representation of the hierarchical structure of value alignment in AI. This figure categorizes the significance, challenges, approaches, frameworks, and the role of human preferences, underscoring the necessity of a multidisciplinary approach. By integrating these diverse elements, researchers and practitioners can better understand how to effectively align AI systems with human values, ensuring their deployment is both ethical and beneficial to society.

3 Value Alignment in AI

3.1 Significance and Challenges of Value Alignment

Value alignment is pivotal for AI systems to function in accordance with human ethical standards and societal norms, particularly in high-stakes domains like autonomous vehicles and military applications [16]. As AI technologies advance, they must adapt to evolving human preferences and ethical standards, posing significant challenges. The principal-agent problem exemplifies these challenges, where conventional methods fail to align an agent's utility with that of its principal [6]. Conflicting values further complicate encoding AI systems to reflect human morals [17]. The opacity of decision-making processes in AI, especially in deep neural networks, erodes public trust, notably in sensitive sectors like healthcare, where biases and cybersecurity risks are prevalent [7].

Machine learning systems' dependence on biased data and flawed reward functions can lead to unethical outcomes, underscoring the necessity for robust value alignment mechanisms [18]. Ensuring robot safety in dynamic environments using reinforcement learning is challenging [19]. Overcoming the naturalistic fallacy is crucial for AI systems to integrate ethical principles with empirical facts without oversimplifying moral reasoning [20].

The absence of a legal framework recognizing robots as subjects complicates liability attribution for damages caused by their actions [21]. Limitations of reinforcement learning from human feedback (RLHF) in capturing human ethics nuances and ensuring AI safety present significant challenges [22]. Potential RLHF misuse, difficulties in collecting human feedback, and the risk of perpetuating biases further complicate value alignment [23].

The Value-Alignment Problem (VAP) highlights the difficulty of ensuring autonomous agents adhere to human values in complex environments [24]. The unpredictability of human and AI behavior, alongside challenges in defining and measuring 'reasonable' AI behavior, presents further obstacles

AI-generated, for reference only.

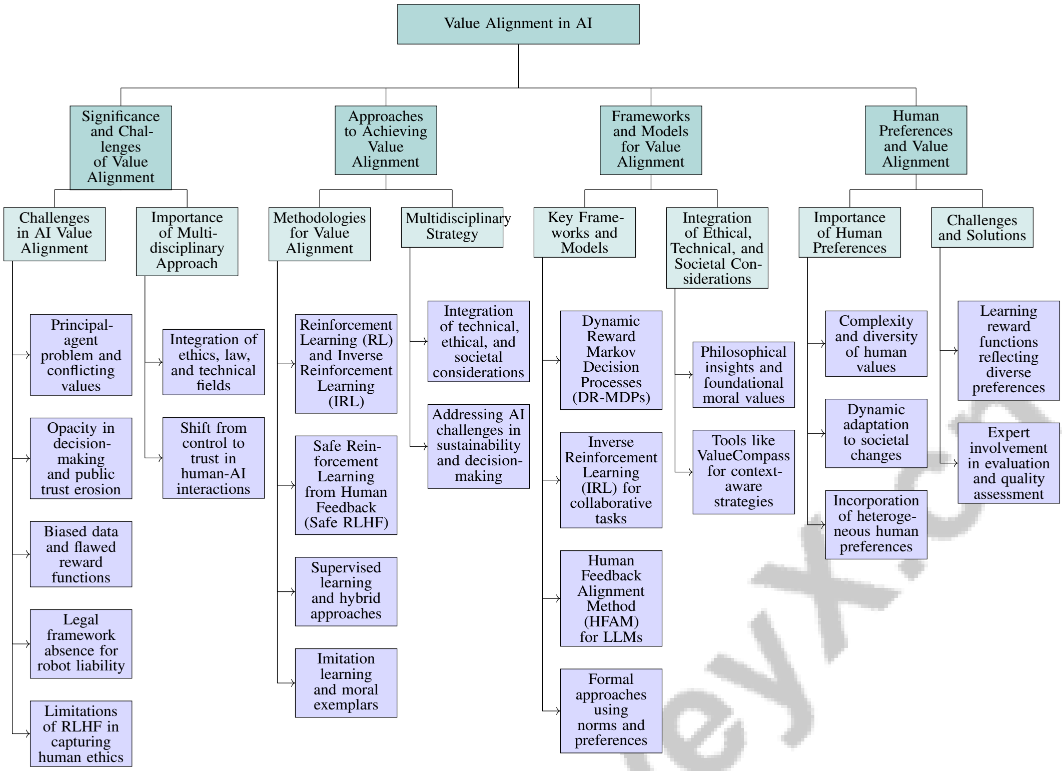
Value Alignment in AI
Significance and Challenges of Value Alignment
Approaches to Achieving Value Alignment
Frameworks and Models for Value Alignment
Human Preferences and Value Alignment
Challenges in AI Value Alignment
Importance of Multi-disciplinary Approach
Methodologies for Value Alignment
Multidisciplinary Strategy
Key Frameworks and Models
Integration of Ethical, Technical, and Societal Considerations
Importance of Human Preferences
Challenges and Solutions
Principal-agent problem and conflicting values
Opacity in decision-making and public trust erosion
Biased data and flawed reward functions
Legal framework absence for robot liability
Limitations of RLHF in capturing human ethics
Integration of ethics, law, and technical fields
Shift from control to trust in human-AI interactions
Reinforcement Learning (RL) and Inverse Reinforcement Learning (IRL)
Safe Reinforcement Learning from Human Feedback (Safe RLHF)
Supervised learning and hybrid approaches
Imitation learning and moral exemplars
Integration of technical, ethical, and societal considerations
Addressing AI challenges in sustainability and decision-making
Dynamic Reward Markov Decision Processes (DR-MDPs)
Inverse Reinforcement Learning (IRL) for collaborative tasks
Human Feedback Alignment Method (HFAM) for LLMs
Formal approaches using norms and preferences
Philosophical insights and foundational moral values
Tools like ValueCompass for context-aware strategies
Complexity and diversity of human values
Dynamic adaptation to societal changes
Incorporation of heterogeneous human preferences
Learning reward functions reflecting diverse preferences
Expert involvement in evaluation and quality assessment

Figure 2: This figure illustrates the hierarchical structure of value alignment in AI, categorizing the significance, challenges, approaches, frameworks, and the role of human preferences. It highlights the importance of a multidisciplinary approach and integration of ethical, technical, and societal considerations for aligning AI systems with human values.

[25]. Addressing algorithmic biases, ensuring accountability, and overcoming AI decision-making's 'black box' nature are critical challenges [12].

A multidisciplinary approach, incorporating ethics, law, and technical fields, is essential for designing AI systems aligned with human values and ethical principles. Shifting from control to trust may foster safer human-AI interactions [26].

3.2 Approaches to Achieving Value Alignment

Achieving value alignment in AI systems necessitates a multifaceted approach integrating various methodologies. As illustrated in Figure 3, the key approaches to achieving value alignment can be categorized into Reinforcement Learning, Conflict Management, and Supervised Learning, each encompassing specific methodologies and models. Reinforcement learning (RL), particularly Inverse Reinforcement Learning (IRL), is crucial for adapting AI values to align with human values during interactions [27]. Safe Reinforcement Learning from Human Feedback (Safe RLHF) enhances this process by separating reward and cost model training, aligning AI behavior with human values [28].

Managing conflicts in AI alignment involves external incentives and reducing information asymmetry, emphasizing transparency and mutual understanding in human-AI interactions [6]. The METUX model evaluates technology's impact on user experiences to support autonomy and achieve value alignment [10]. The BEAVER TAILS dataset independently assesses helpfulness and harmlessness, enhancing safety alignment training [11].

Supervised learning significantly contributes to value alignment by training AI systems to recognize scenarios where human values may be compromised. Enhanced methodologies for collecting and utilizing human feedback improve alignment with human values [29]. A hybrid value alignment approach, merging machine learning-based and logic-based methods, aims to create ethically sound

AI systems [18]. Quantified modal logic connects ethical reasoning with factual states to ensure a comprehensive ethical framework for AI decision-making [20].

Imitation learning allows AI systems to emulate moral exemplars and incorporate virtue ethics into decision-making processes, emphasizing moral education [30]. A formal approach to defining and computing value alignment through norms and preferences provides a structured framework for AI systems [24].

These approaches underscore the necessity of a multidisciplinary strategy, integrating technical, ethical, and societal considerations to develop AI systems aligned with human moral and ethical standards. This integration is vital for addressing AI challenges across various domains, including sustainability initiatives where AI enhances decision-making and resource management [13].

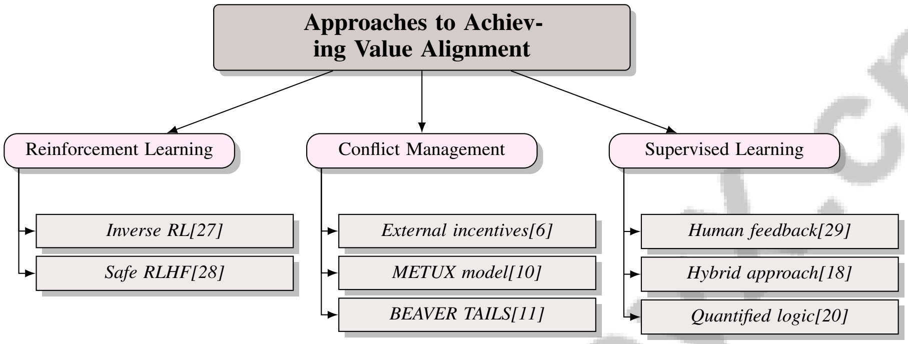

Figure 3: This figure illustrates the key approaches to achieving value alignment in AI systems, categorized into Reinforcement Learning, Conflict Management, and Supervised Learning, each with specific methodologies and models.

3.3 Frameworks and Models for Value Alignment

Method Name	Alignment Mechanisms	Ethical Considerations	Adaptability and Learning
MDP[31]	Adaptive-learner Strategy	Personalized Value Alignment	Dynamically Adjusts
AVAM[27]	Inverse Reinforcement Learning	Trust, Human Values	Dynamically Align Values
HFAM[29]	Structured Feedback Collection	Clearer Evaluation Criteria	Iteratively Improve Summarization
PRO[32]	Preference Ranking Optimization	Sensitive Content Handling	Dynamically Learn Preferences

Table 1: Comparison of various AI alignment frameworks and models, highlighting their alignment mechanisms, ethical considerations, and adaptability in learning. The table provides an overview of methods such as MDP, AVAM, HFAM, and PRO, emphasizing their strategies for aligning AI systems with human values and preferences.

Developing frameworks and models for value alignment in AI systems ensures these technologies operate according to human ethical standards and societal values. Table 1 presents a comprehensive comparison of different frameworks and models for value alignment in AI systems, illustrating their mechanisms, ethical considerations, and adaptability features. Dynamic Reward Markov Decision Processes (DR-MDPs) provide a formal structure for aligning AI systems with evolving human preferences [33]. Adaptive-learner strategies enhance AI systems' ability to learn and align with human reward functions [31].

Inverse Reinforcement Learning (IRL) facilitates AI alignment with human values in collaborative tasks, enhancing trust and performance [27]. The AVAM framework dynamically aligns a robot's values with humans, improving collaborative environments [27].

Integrating Safe Reinforcement Learning with RLHF allows dynamic adjustments between helpfulness and harmlessness during AI fine-tuning [28]. This ensures AI systems adapt their behavior to align with human values while maintaining safety and ethical standards. Research on RLHF and RLAIF emphasizes the need for an integrated approach addressing both technical and ethical dimensions [22].

The Human Feedback Alignment Method (HFAM) systematically collects human feedback to train large language models (LLMs), ensuring outputs align with human values [29]. The PRO framework

AI-generated, for reference only.

optimizes LLMs to align response rankings with human preferences, enhancing AI attunement to human values [32].

A formal approach to defining and computing value alignment through norms and preferences offers a structured framework for aligning AI systems with human values [24]. The theoretical perspective involves a deontological framework using quantified modal logic to express ethical principles for AI actions [18]. This framework analyzes ethical reasoning and empirical observation, establishing a foundation for value alignment systems [20].

The theoretical framework proposed by [17] emphasizes AI solutions' social context, highlighting broader societal impacts and ethical implications beyond technical content.

These frameworks and models underscore the necessity of integrating ethical, technical, and societal considerations in AI development. Incorporating philosophical insights and foundational moral values into AI alignment frameworks establishes a comprehensive basis for ensuring AI systems resonate with diverse human moral and ethical standards. By articulating clear, defensible values such as survival, sustainable intergenerational existence, and truth, these frameworks facilitate systematic alignment, addressing varying human beliefs. Tools like ValueCompass assess AI alignment across contexts, revealing potential misalignments and emphasizing context-aware strategies to uphold societal values in AI development [34, 35, 2]. Such integration addresses AI challenges in various domains, facilitating societal integration that respects human values.

3.4 Human Preferences and Value Alignment

Integrating human preferences into AI systems is crucial for effective value alignment, ensuring technologies operate harmoniously with diverse human values and ethical standards. Understanding human values' complexity is essential, as they serve as benchmarks for evaluating actions and states, significantly influencing AI decision-making [24]. AI systems must accommodate a broad spectrum of human values, rather than adhering to a singular moral framework [18].

The dynamic nature of human values requires AI systems to be adaptable, evolving alongside societal changes and individual priorities. Reinforcement Learning from Human Feedback (RLHF) enhances information integrity and aligns AI systems with human values, reducing bias in outputs [23]. Methodologies that automate alignment processes minimize manual curation, enabling rapid adaptation to varied value systems [29].

Incorporating heterogeneous human preferences involves understanding diverse and sometimes conflicting societal values, crucial for AI systems operating across different cultural and ethical contexts. The interplay between facts and values is critical in AI ethics, necessitating systems that are technically robust and ethically aligned [18].

Challenges in learning reward functions that reflect diverse human preferences are addressed by frameworks considering individual values and priorities, ensuring AI systems respond to varied ethical landscapes [20]. Expert involvement in evaluation establishes a reliable baseline for assessing AI output quality, particularly in tasks like summarization, where human judgment is crucial [29].

These insights highlight the importance of considering human preferences in value alignment, emphasizing the need for flexible, adaptive AI systems capable of integrating a wide range of human values. Implementing robust ethical frameworks ensures AI systems function responsibly and effectively across societal contexts, fostering an inclusive, democratic approach to technology design that prioritizes human rights, autonomy, and harm mitigation [10, 36, 14].

4 Trust in AI Systems

Establishing trust in artificial intelligence (AI) is crucial for its successful integration across various societal sectors. Trust is a multifaceted construct involving transparency, explainability, accountability, fairness, auditability, and safety. These dimensions are essential for fostering stakeholder confidence in AI systems, addressing unique challenges throughout their lifecycle. Enhancing these aspects is vital for societal acceptance, particularly in high-stakes environments where human-AI collaboration is necessary for optimal outcomes [37, 38, 39]. This section explores the roles of transparency, explainability, and accountability as foundational elements that underpin user confidence and societal acceptance.

4.1 Transparency, Explainability, and Accountability

Transparency, explainability, and accountability are critical in building trust in AI systems, addressing the ethical and operational complexities inherent in these technologies. Transparency involves clarifying decision-making processes, enhancing user comprehension, and fostering trust, particularly in robotics and autonomous systems where it can alleviate public skepticism post-accident [40]. Incorporating ethical principles into decision-making frameworks, such as those in autonomous vehicles, enhances transparency by facilitating nuanced ethical deliberations [41].

Explainability, closely linked to transparency, refers to the capacity of AI systems to provide comprehensible outputs. The MATCH conceptual model illustrates how trustworthiness is communicated and processed, emphasizing the need for tailored explanations that accommodate diverse cognitive processes in trust judgments [42]. This diversity necessitates designing AI systems that cater to various end-users, ensuring explanations are accessible and meaningful.

As depicted in Figure 4, the hierarchical structure of key concepts in AI systems focusing on transparency, explainability, and accountability is illustrated, highlighting the primary categories, their subcategories, and details. This visual representation showcases the relationships and significance of each aspect in building trust and addressing ethical challenges in AI.

Accountability establishes clear mechanisms for attributing responsibility for AI actions and decisions, crucial for maintaining trust, especially in sensitive domains. The Safe Reinforcement Learning Framework (SRRL) exemplifies this by integrating human feedback and interactive behaviors, ensuring alignment with ethical standards [19]. This alignment is reinforced by frameworks emphasizing compliance, reporting, oversight, and enforcement as accountability goals.

Integrating transparency, explainability, and accountability is essential for addressing the challenges posed by AI systems and ensuring they operate in alignment with human values and ethical standards. By enhancing factors such as fairness, explainability, auditability, and safety, these technologies can foster stakeholder trust, which is vital for their responsible and effective societal adoption [37, 38, 43, 39, 44].

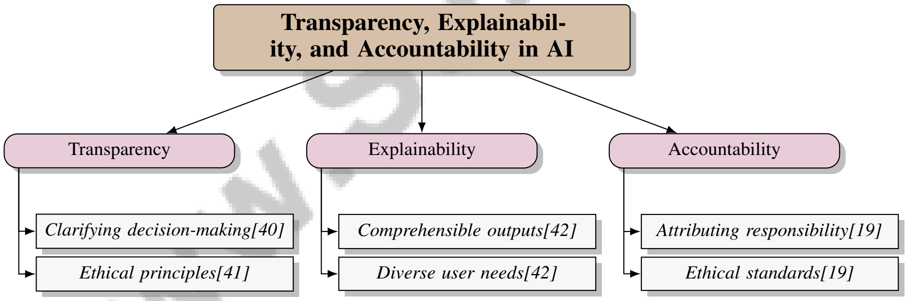
Transparency, Explainability, and Accountability in AI
Transparency
Explainability
Accountability
Clarifying decision-making[40]
Ethical principles[41]
Comprehensible outputs[42]
Diverse user needs[42]
Attributing responsibility[19]
Ethical standards[19]

Figure 4: This figure illustrates the hierarchical structure of key concepts in AI systems focusing on transparency, explainability, and accountability. It highlights the primary categories, their subcategories, and details, showcasing the relationships and significance of each aspect in building trust and addressing ethical challenges in AI.

4.2 Mechanistic Interpretability and Trust

Mechanistic interpretability significantly enhances trust in AI systems by providing insights into their internal workings, thereby fostering transparency and accountability. The survey by [45] highlights the necessity of mechanistic interpretability for ensuring AI safety and alignment, advocating for clearer concepts and standards in the field. This aligns with the broader call for transparency and accountability to build public trust in AI systems [46].

A critical aspect of mechanistic interpretability is its ability to quantify the influence of AI recommendations on human responses, reflecting the trust placed in these systems [47]. This quantification is vital for understanding and calibrating trust, focusing on case-specific trust calibration rather than general enhancement [39]. By elucidating how AI systems derive recommendations, mechanistic interpretability enables users to develop a more nuanced trust.

AI-generated, for reference only.

The concept of 'meaningful human control' is integral to maintaining moral responsibility over autonomous systems, ensuring human accountability for AI decisions. This underscores the importance of designing AI systems capable of tracking human moral reasoning, thereby reinforcing accountability and trust. Ethical considerations must also be embedded within AI systems to cultivate public trust and accountability, as argued by [48].

Proposed practices to enhance the safety and accountability of agentic AI systems are crucial for building societal trust [49]. Implementing an 'ethical black box' in robots to continuously record sensor data and internal statuses can significantly aid in accident investigations and bolster public confidence in AI systems [40].

5 Human-AI Interaction

5.1 Trust and Decision-Making in Human-AI Interaction

Trust is pivotal in human-AI interactions, shaping decision-making and user outcomes, particularly in high-stakes environments where collaboration is vital [37, 9]. Enhanced collaboration in human-AI interactions improves decision-making and operational efficiency across domains [50]. In cooperative settings, such as multi-agent AI systems, trust is crucial for effective collaboration [51]. Frameworks that integrate human and AI capabilities further boost creativity and efficiency [52].

Systematic reward explanation techniques enhance transparency, fostering trust and improving collaborative decision-making, especially in critical scenarios like search and rescue [53, 54]. However, restoring trust after violations due to errors or unexpected behaviors remains challenging, necessitating a deep understanding of user interactions and trust cues, as seen with AI symptom checkers [55, 42]. Ensuring fairness and ethical alignment in AI behavior, as demonstrated by systems like GPT-4o, is essential for maintaining trust and supporting effective decision-making across diverse contexts [56].

5.2 Human-Centered Design and User Experience

Human-centered design enhances user experience and trust in AI systems by focusing on user needs and behaviors, ensuring systems are intuitive and aligned with human values [57]. This approach is critical for improving user experience and safety. A reflective sociotechnical approach in developing explainable AI systems ensures both social and technical dimensions meet diverse user needs while maintaining transparency and accountability [58]. Incorporating user feedback into the design process enhances satisfaction and trust, leading to ethical AI deployments.

Iterative testing and refinement are crucial for aligning AI systems with user expectations and ethical standards, addressing usability issues to minimize frustration and mistrust. By emphasizing user experience and ethical considerations, human-centered design contributes to developing AI systems that enhance effectiveness while promoting societal values. This approach encourages interdisciplinary collaboration, ensuring AI technologies empower users, support human rights, and align with sustainable development goals [59, 12, 60, 61, 57].

5.3 Challenges and Opportunities in Human-AI Interaction

Integrating AI into human life presents challenges and opportunities. A key challenge is the complexity of large AI models, which often struggle with generalizability across architectures and tasks [45]. This complexity can hinder effective deployment, necessitating adaptable, robust models for diverse environments. The opacity of AI decision-making processes can lead to mistrust and reluctance to rely on these systems. Enhancing transparency and explainability is crucial for fostering trust, as fairness, explainability, and auditability significantly influence user perceptions. Understanding these elements is vital for developing reliable, ethically sound AI systems, especially in high-stakes domains like healthcare and justice [37, 38, 39, 48]. Advanced interpretability techniques can elucidate AI models' workings, fostering user confidence and facilitating seamless interactions.

Despite challenges, significant innovation opportunities exist in human-AI interaction. AI integration can enhance productivity, improve decision-making, and foster creativity. Leveraging AI across sectors like business, education, and healthcare can yield new business models and competitive advantages, enhancing value outcomes. Successful implementation requires careful resource orchestration

AI-generated, for reference only.

and governance to navigate complexities, ensuring AI serves human interests and promotes fairness [12, 52, 44].

The field of human-AI interaction fosters interdisciplinary collaboration, uniting experts from various domains—such as human-computer interaction, ethics, and machine learning—to comprehensively address the ethical, technical, and societal implications of AI technologies. This collaboration is essential for developing human-centered AI frameworks that prioritize user values, promote trust and safety, and ensure equitable access to AI systems, guiding responsible AI integration into society [62, 59, 3, 15]. Such collaboration can lead to more ethical and user-centered AI systems that align with human values and societal norms.

6 Ethical AI and Autonomous Systems

The ethical landscape of artificial intelligence (AI) and autonomous systems is increasingly complex, requiring robust frameworks to navigate societal challenges such as value alignment, privacy, and misinformation [62, 63, 64]. These frameworks are essential for guiding the responsible design and implementation of AI technologies, ensuring adherence to societal norms and minimizing potential harms.

6.1 Ethical Frameworks and Guidelines

Ethical frameworks are crucial for ensuring AI systems operate responsibly and mitigate harm. Current frameworks often reflect diverse perspectives, leading to fragmentation [21]. A unified approach incorporating ethical considerations throughout the AI lifecycle is imperative. Meaningful human control (MHC) is vital for maintaining ethical standards, especially in autonomous systems like weapon systems, ensuring human oversight and accountability [37, 38, 64]. Governance structures must address AI's complexities and opacity, employing ethics-based auditing to ensure alignment and accountability.

Incorporating diverse expert and public perspectives enhances ethical frameworks, aligning them with fairness, accountability, and transparency principles [62, 64, 14, 15]. The concept of Green AI emphasizes sustainability, highlighting the need to minimize the environmental impact of AI technologies. Limitations like lack of transparency and biases necessitate robust ethical guidelines, such as implementing ethical black boxes to enhance transparency and accountability in autonomous systems [21].

6.2 Trust and Accountability in Autonomous Systems

Trust and accountability are critical for the deployment of autonomous systems. Automation bias, where users overly rely on AI, underscores the need for mechanisms that foster informed trust [65]. Accountability frameworks must integrate explicit documentation and audit trails, ensuring autonomous systems can be held responsible for their actions. These frameworks should address accountability's multifaceted nature, promoting ethical alignment and transparency [66, 67, 68, 64].

Meaningful human control is essential for ensuring accountability, mandating human authority over critical decisions, especially in high-stakes contexts. This principle addresses potential "responsibility gaps" from delegating decision-making to machines, fostering a framework prioritizing human rights and mitigating automation risks [16, 69, 14, 70, 57].

6.3 Technical and Ethical Challenges in Autonomous Systems

Autonomous systems present technical and ethical challenges that must be addressed to ensure alignment with human values. A primary technical challenge is articulating action reasons, complicating accountability [40]. Ethically, embedding values into design is complex, exacerbated by limitations in AI alignment practices and the influence of Western-dominated tech industry perspectives.

Biases in AI can affect educational outcomes and displace traditional roles, raising ethical concerns about AI deployment in sensitive areas [1, 71, 14]. Governance practices and ethical frameworks grounded in human rights are essential to mitigate risks and prioritize human values. The appropriate-

AI-generated, for reference only.

ness framework, while valuable, remains vulnerable to manipulation, necessitating comprehensive governance of chatbot interactions [68, 72].

Existing studies often lack a comprehensive understanding of ethical considerations in autonomous systems, leading to deployment risks [55]. Significant gaps remain in understanding trust repair conditions across contexts and demographics. The framework by [19] identifies challenges like robustness against erroneous behavior and maintaining transparency, validated through autonomous vehicle and unmanned air system examples.

Developing Trustworthy Autonomous Systems (TAS) requires a multidisciplinary approach integrating philosophy, ethics, sociology, and engineering to address responsibility, accountability, and alignment with human-centered values [73, 66, 46, 74]. This collaboration ensures ethical and effective integration of autonomous systems into society.

7 Challenges and Future Directions

7.1 Interdisciplinary Collaboration and Emerging Technologies

Advancing artificial intelligence (AI) technologies necessitates interdisciplinary collaboration, integrating insights from engineering, philosophy, psychology, social sciences, and ethics, to address complex challenges and align AI with human values and societal norms [75]. Such collaboration among researchers, developers, and policymakers is essential for crafting comprehensive regulatory frameworks that address the ethical, technical, and societal implications of AI [61].

Emerging technologies enhance AI capabilities and address existing challenges by integrating advanced algorithms and machine learning techniques, thereby improving performance and reliability in complex reasoning tasks [76]. Future research should focus on optimizing interactions between evaluators and generators to enrich training datasets, enhancing AI systems' reasoning abilities [76].

In AI-driven healthcare, interdisciplinary collaboration is crucial for developing regulatory frameworks that balance innovation with ethical considerations, addressing technical challenges, and understanding the interplay of language, culture, and ethics in large language models (LLMs) for real-world applications [77, 78]. Establishing ethical guidelines for AI manipulation in human-AI interactions requires defining acceptable behavior boundaries to ensure AI systems operate ethically and contribute positively to human well-being [54].

Refining accountability frameworks specific to AI contexts and developing practical implementation guidelines are critical for future research [67]. Additionally, investigating stakeholder roles in AI governance is essential for responsible AI development and deployment [79]. This involves comprehensive empirical research on AI governance implementation and engaging diverse stakeholders to understand their perspectives and needs [79].

8 Conclusion

This survey delves into the intricate interplay between artificial intelligence (AI) systems and human values, underscoring the imperative of aligning AI with ethical norms to ensure safety and trust across diverse sectors, including healthcare and autonomous systems. Addressing the principal-agent dilemma, the opacity in AI decision-making, and the complexity of encoding varied human morals are pivotal challenges in achieving this alignment. The exploration of methodologies such as reinforcement learning, inverse reinforcement learning, and hybrid approaches illustrates the multifaceted strategies available to support value alignment.

The role of transparency, explainability, and accountability is crucial in fostering trust in AI, with mechanistic interpretability enhancing user confidence. Emphasizing human-centered design and ethical frameworks is vital for optimizing user interaction and ensuring ethical AI deployment.

Furthermore, the survey highlights the necessity for interdisciplinary collaboration and the integration of cutting-edge technologies to navigate the challenges and opportunities in human-AI interaction. Advocating for continuous oversight and public engagement regarding the societal impacts of advanced AI technologies, the survey promotes a forward-thinking stance on ethical AI development.

AI-generated, for reference only.

References

[1] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2024.

[2] Iason Gabriel. Artificial intelligence, values and alignment, 2020.

[3] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, 2024.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.

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

[6] Steve Phelps and Rebecca Ranson. Of models and tin men: A behavioural economics study of principal-agent problems in ai alignment using large-language models, 2023.

[7] Journal of the american medical.

[8] Sunayana Rane, Polyphony J. Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas L. Griffiths. Concept alignment, 2024.

[9] Hollen Barmer, Rachel Dzombak, Matthew Gaston, Vijaykumar Palat, Frank Redner, Carol Smith, and Tanisha Smith. Human-centered ai. 2021.

[10] Rafael A Calvo, Dorian Peters, Karina Vold, and Richard M Ryan. Supporting human autonomy in ai systems: A framework for ethical enquiry. *Ethics of digital well-being: A multidisciplinary approach*, pages 31–54, 2020.

[11] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

[12] David Mhlanga. Human-centered artificial intelligence: The superlative approach to achieve sustainable development goals in the fourth industrial revolution. *Sustainability*, 14(13):7804, 2022.

[13] Jayden Khakurel, Birgit Penzenstadler, Jari Porras, Antti Knutas, and Wenlu Zhang. The rise of artificial intelligence under the lens of sustainability. *Technologies*, 6(4):100, 2018.

[14] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. A human rights-based approach to responsible ai, 2022.

[15] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28:689–707, 2018.

[16] Daniele Amoroso and Guglielmo Tamburrini. Autonomous weapons systems and meaningful human control: ethical and legal issues. *Current Robotics Reports*, 1:187–194, 2020.

[17] Betty Li Hou and Brian Patrick Green. A multi-level framework for the ai alignment problem, 2023.

[18] Tae Wan Kim, John Hooker, and Thomas Donaldson. Taking principles seriously: A hybrid approach to value alignment, 2020.

[19] Shangding Gu, Alap Kshirsagar, Yali Du, Guang Chen, Jan Peters, and Alois Knoll. A human-centered safe robot reinforcement learning framework with interactive behaviors, 2023.

[20] Tae Wan Kim, Thomas Donaldson, and John Hooker. Grounding value alignment with ethical principles, 2019.

[21] Gerhard Wagner. Robot, inc.: personhood for autonomous systems? *Fordham L. Rev.*, 88:591, 2019.

[22] Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. Ai alignment through reinforcement learning from human feedback? contradictions and limitations, 2024.

[23] Gabrielle Kaili-May Liu. Perspectives on the social impacts of reinforcement learning with human feedback, 2023.

[24] Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perelló. Value alignment: a formal approach, 2021.

[25] Sunayana Rane. The reasonable person standard for ai, 2024.

[26] Yuzhu Cai, Sheng Yin, Yuxi Wei, Chenxin Xu, Weibo Mao, Felix Juefei-Xu, Siheng Chen, and Yanfeng Wang. Ethical-lens: Curbing malicious usages of open-source text-to-image models, 2025.

[27] Shreyas Bhat, Joseph B. Lyons, Cong Shi, and X. Jessie Yang. Value alignment and trust in human-robot interaction: Insights from simulation and user study, 2024.

[28] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback, 2023.

[29] Thilo Hagendorff and Sarah Fabi. Methodological reflections for ai alignment research using human feedback, 2022.

[30] Gopal P. Sarma. Brief notes on hard takeoff, value alignment, and coherent extrapolated volition, 2018.

[31] Shreyas Bhat, Joseph B. Lyons, Cong Shi, and X. Jessie Yang. Evaluating the impact of personalized value alignment in human-robot interaction: Insights into trust and team performance outcomes, 2023.

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

[33] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions, 2024.

[34] Betty Li Hou and Brian Patrick Green. Foundational moral values for ai alignment, 2023.

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

[36] Virginia Dignum. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, 20(1):1–3, 2018.

[37] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 272–283, 2020.

[38] Steven Lockey, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. 2021.

[39] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305, 2020.

[40] Alan FT Winfield and Marina Jirotka. The case for an ethical black box. In *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Guildford, UK, July 19–21, 2017, Proceedings 18*, pages 262–273. Springer, 2017.

[41] Maximilian Geisslinger, Franziska Poszler, Johannes Betz, Christoph Lütge, and Markus Lienkamp. Autonomous driving ethics: From trolley problem to ethics of risk. *Philosophy & Technology*, 34(4):1033–1055, 2021.

[42] Q. Vera Liao and S. Shyam Sundar. Designing for responsible trust in ai systems: A communication perspective, 2022.

[43] Arto Laitinen and Otto Sahlgren. Ai systems and respect for human autonomy. *Frontiers in artificial intelligence*, 4:705164, 2021.

[44] Nikolaos-Alexandros Perifanis and Fotis Kitsios. Investigating the influence of artificial intelligence on business value in the digital era of strategy: A literature review. *Information*, 14(2):85, 2023.

[45] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024.

[46] Vicky Charisi, Louise Dennis, Michael Fisher, Robert Lieck, Andreas Matthias, Marija Slavkovik, Janina Sombetzki, Alan FT Winfield, and Roman Yampolskiy. Towards moral autonomous systems. *arXiv preprint arXiv:1703.04741*, 2017.

[47] Ali Shafti, Victoria Derks, Hannah Kay, and A. Aldo Faisal. The response shift paradigm to quantify human trust in ai recommendations, 2022.

[48] Abhishek Kumar, Tristan Braud, Sasu Tarkoma, and Pan Hui. Trustworthy ai in the age of pervasive computing and big data, 2020.

[49] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic ai systems. *Research Paper, OpenAI*, 2023.

[50] Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. Transitioning to human interaction with ai systems: New challenges and opportunities for hci professionals to enable human-centered ai. *International Journal of Human–Computer Interaction*, 39(3):494–518, 2023.

[51] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

[52] Alessandro Pagano, Anders Mørch, Vita Santa Barletta, and Renate Andersen. Ai for humans and humans for ai: Towards cultures of participation in the digital age with human-centered artificial intelligence. 2024.

[53] Lindsay Sanneman and Julie Shah. Explaining reward functions to humans for better human-robot collaboration, 2021.

[54] Tathagata Chakraborti and Subbarao Kambhampati. Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-ai collaboration, 2018.

[55] Ewart J De Visser, Richard Pak, and Tyler H Shaw. From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, 61(10):1409–1427, 2018.

[56] Yu Lei, Hao Liu, Chengxing Xie, Songjia Liu, Zhiyu Yin, Canyu Chen, Guohao Li, Philip Torr, and Zhen Wu. Fairmindsim: Alignment of behavior, emotion, and belief in humans and llm agents amid ethical dilemmas, 2024.

[57] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[58] Upol Ehsan and Mark O Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, pages 449–466. Springer, 2020.

[59] Tara Capel and Margot Brereton. What is human-centered about human-centered ai? a map of the research landscape. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–23, 2023.

[60] Athanasios Mazarakis, Christian Bernhard-Skala, Martin Braun, and Isabella Peters. What is critical for human-centered ai at work?–toward an interdisciplinary theory. *Frontiers in Artificial Intelligence*, 6:1257057, 2023.

[61] Ben Shneiderman. Human-centered ai: A new synthesis. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part I 18*, pages 3–8. Springer, 2021.

[62] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. The ethics of advanced ai assistants, 2024.

[63] Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507, 2019.

[64] Jakob Mökander and Luciano Floridi. Ethics-based auditing to develop trustworthy ai. *Minds and Machines*, 31(2):323–327, 2021.

[65] Mehdi Khamassi, Marceau Nahon, and Raja Chatila. Strong and weak alignment of large language models with human values, 2024.

[66] Vahid Yazdanpanah, Enrico Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M Jonker, and Timothy Norman. Responsibility research for trustworthy autonomous systems. 2021.

[67] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. Accountability in artificial intelligence: what it is and how it works. *Ai & Society*, 39(4):1871–1882, 2024.

[68] Muneera Bano, Didar Zowghi, Pip Shea, and Georgina Ibarra. Investigating responsible ai for scientific research: An empirical study, 2023.

[69] Joanna J Bryson and Andreas Theodorou. How society can maintain human-centric artificial intelligence. *Human-centered digitalization and services*, pages 305–323, 2019.

[70] Filippo Santoni de Sio and Jeroen Van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:15, 2018.

[71] Research.

[72] Hendrik Kempt, Alon Lavie, and Saskia K. Nagel. Appropriateness is all you need!, 2023.

[73] Ville Vakkuri, Kai-Kristian Kemell, Joni Kultanen, Mikko Siponen, and Pekka Abrahamsson. Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *arXiv preprint arXiv:1906.07946*, 2019.

[74] Alexis Roger, Esma Aïmeur, and Irina Rish. Towards ethical multimodal systems, 2024.

[75] Damien Trentesaux and Stamatis Karnouskos. Engineering ethical behaviors in autonomous industrial cyber-physical human systems. *Cognition, Technology & Work*, 24(1):113–126, 2022.

[76] Zhiqing Sun, Longhui Yu, Yikang Shen, Weiyang Liu, Yiming Yang, Sean Welleck, and Chuang Gan. Easy-to-hard generalization: Scalable alignment beyond human supervision, 2024.

[77] Steven M Williamson and Victor Prybutok. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in ai-driven healthcare. *Applied Sciences*, 14(2):675, 2024.

[78] Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. Ethical reasoning and moral value alignment of llms depend on the language we prompt them in, 2024.

[79] Teemu Birkstedt, Matti Minkkinen, Anushree Tandon, and Matti Mäntymäki. Ai governance: themes, knowledge gaps and future agendas. *Internet Research*, 33(7):133–167, 2023.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.