A Survey of AI Cognitive Bias Ethical Implications and Bias Mitigation in Language Models

Abstract

This survey explores the intricate relationship between artificial intelligence (AI) technologies and inherent cognitive biases, with a focus on ethical implications and gender stereotypes within large language models (LLMs). The pervasive influence of AI across various sectors necessitates a critical examination of biases encoded in AI systems, which often manifest as gender stereotypes and societal norms. The survey highlights the profound impact of these biases on AI outputs, particularly in high-stakes areas such as healthcare and education, where they can adversely affect decision-making and perpetuate inequalities. It underscores the importance of integrating ethical considerations into AI development, advocating for transparency, accountability, and inclusivity to foster trust and promote equitable outcomes. The study examines various bias mitigation strategies, including data augmentation, model fine-tuning, and algorithmic fairness, demonstrating their effectiveness in reducing biases without compromising model performance. Additionally, the survey emphasizes the need for responsible AI practices, proposing frameworks and guidelines that align AI systems with societal values and ethical standards. The findings suggest that ongoing research and interdisciplinary collaboration are crucial for advancing ethical AI practices, ensuring that AI technologies contribute positively to societal outcomes. The paper concludes by advocating for continued dialogue among researchers, policymakers, and industry stakeholders to address the ethical challenges posed by AI and promote responsible development and deployment of AI technologies.

1 Introduction

1.1 Growing Importance of AI Technologies

The rapid advancement and integration of artificial intelligence (AI) technologies across various sectors have transformed modern society. AI's influence spans healthcare to finance, enhancing efficiencies and offering novel solutions to complex problems [1]. The evolution of AI, particularly through Large Language Models (LLMs) like ChatGPT, underscores its versatility and profound impact on research and applications [2].

Global engagement with AI is reflected in its adoption across diverse cultural and geographical contexts. A survey of 10,005 participants from countries including Australia, Canada, the USA, South Korea, France, Brazil, India, and Nigeria illustrates widespread excitement and concern about AI's potential and future implications [3]. This perspective highlights not only technological advancements but also societal readiness to embrace AI innovations.

In scientific research, AI catalyzes significant advancements, enabling unprecedented speed and accuracy in exploring new frontiers [1]. Its integration exemplifies a transformative impact, facilitating discoveries previously unattainable. As AI evolves, its role in shaping the future of various industries and research domains becomes increasingly critical, necessitating ongoing evaluation to capture the complexities of real-world applications [4].

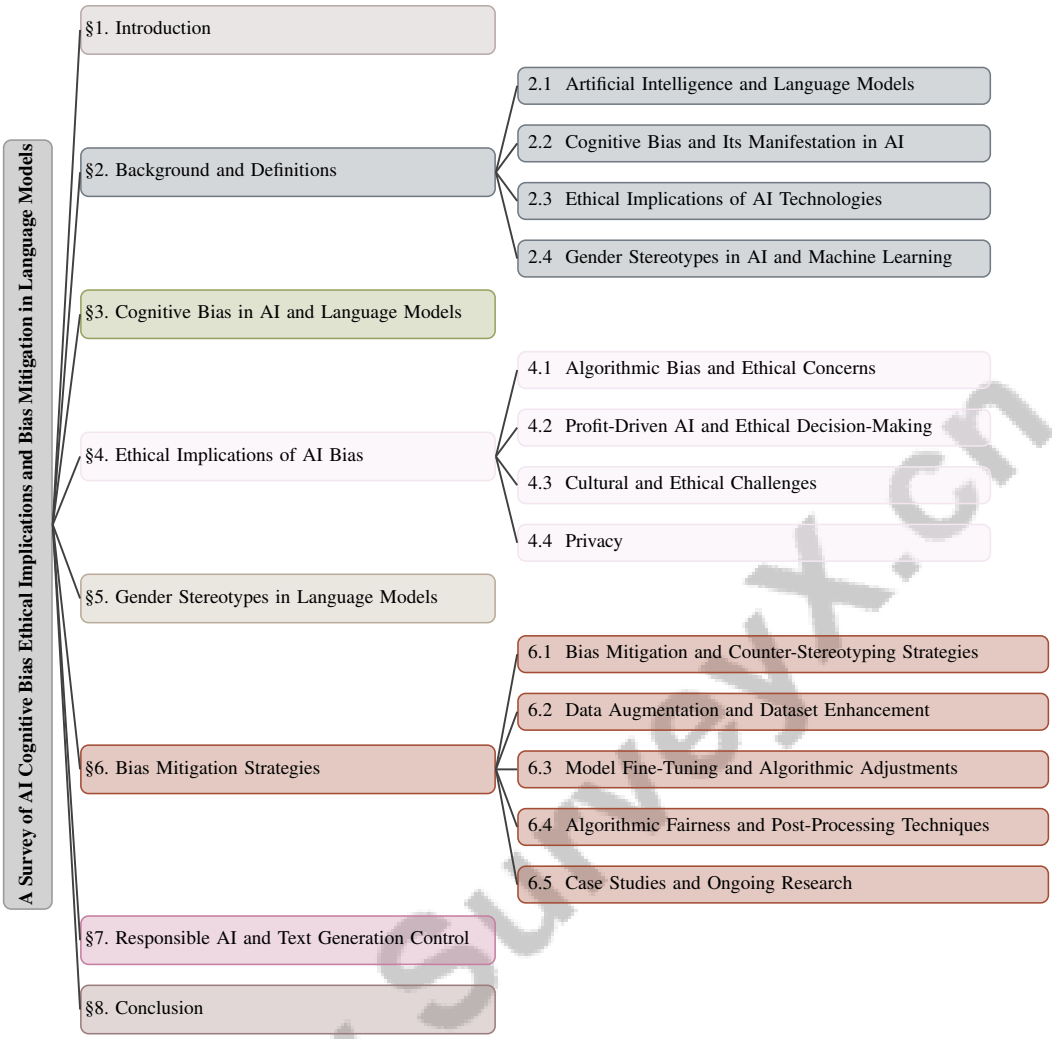

Figure 1: chapter structure

Moreover, moral and ethical considerations surrounding AI, particularly in social robots and chatbots, emphasize the importance of addressing the implications of advanced AI technologies [5]. As AI systems become integral to daily life, responsible development and deployment are paramount to ensure societal benefits.

1.2 Focus on Cognitive Biases and Ethical Implications

The integration of cognitive biases within AI systems, especially in LLMs, presents profound ethical challenges. These biases often manifest as gender stereotypes and societal norms, influencing decision-making processes and necessitating comprehensive evaluation and mitigation strategies. The personalization of AI-generated content, influenced by user identity, complicates the bias landscape, highlighting the need for nuanced understanding and intervention [6].

Ethical implications are pronounced in high-stakes areas like healthcare, where LLMs may propagate harmful misinformation, adversely affecting patient outcomes and healthcare delivery [7]. Evaluations of LLMs' moral alignment through responses to ethical dilemmas reveal significant gaps in understanding the ethical dimensions of AI systems [8]. Bridging these gaps is crucial to ensure alignment with societal values and ethical norms.

The deployment of AI in education underscores the necessity for transparency and trust, as biases can critically influence assessment fairness and validity [9]. Ethical considerations extend to human com-

AI-generated, for reference only.

munication, where generative AI affects the identification and protection of contextual information, thereby influencing interpersonal interactions [10].

In emotion classification tasks, cognitive biases in AI systems raise ethical concerns, emphasizing the need for socially responsible AI development that acknowledges these biases [11]. Current IT-related ethical codes inadequately address the unique challenges posed by LLM-based content generation, necessitating new ethical standards tailored to modern AI complexities [12].

The differentiation between quantitative and qualitative intelligence in LLMs adds another layer of ethical complexity, essential for assessing AI systems' intelligence and decision-making capabilities [2]. These considerations underscore the need for ethical frameworks guiding responsible AI development, ensuring alignment with established ethical principles and societal expectations. The presence of unfair biases, such as gender and racial biases, in pretrained language models complicates their adoption and necessitates sustainable debiasing approaches [13]. Furthermore, ethical challenges posed by AI systems like DALL-E 2 regarding authorship, ownership, and potential misuse of generated content highlight the importance of addressing ethical issues in AI development [14].

1.3 Structure of the Survey

The survey provides a comprehensive analysis of multifaceted issues surrounding AI technologies, focusing on cognitive biases, ethical implications, and gender stereotypes within language models. Following the introduction, which highlights the growing importance of AI technologies and the central focus on cognitive biases and ethical implications, the paper systematically explores foundational concepts essential for understanding these issues. The background section defines key terms such as AI, cognitive bias, ethical implications, gender stereotypes, and language models, emphasizing their relevance to text generation and bias mitigation.

Subsequent sections examine the presence and impact of cognitive biases in AI and language models, analyzing how these biases manifest and their implications on AI outputs. This is followed by an analysis of the ethical implications of AI bias, addressing algorithmic bias, profit-driven decision-making, and cultural challenges, supported by insights from studies on public perception and moral considerations of AI. The survey investigates gender stereotypes within language models, providing examples of their encoding and impact on AI-generated content, and discussing implications for gender equality.

The paper further explores bias mitigation strategies, discussing various approaches such as data augmentation, model fine-tuning, and algorithmic fairness, highlighting successful case studies and ongoing research efforts. The section on responsible AI and text generation control outlines principles and frameworks for ethical AI practices, emphasizing transparency, accountability, and inclusivity.

This synthesis of key findings underscores the critical need to address cognitive biases and ethical implications associated with AI, highlighting the importance of ongoing research and collaborative efforts to foster responsible AI practices and effective bias mitigation strategies, particularly considering the diverse social impacts of AI technologies across different geographical contexts. By advocating for a comprehensive understanding of local cultural and social factors, the conclusion emphasizes the necessity of bridging the gap between ethical principles and practical applications, ensuring that AI systems are developed and implemented to promote equity and accountability. [15, 16, 17, 18]. This structured approach ensures a thorough examination of the complex interplay between AI technologies and societal values, providing valuable insights for researchers, practitioners, and policymakers.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Artificial Intelligence and Language Models

Artificial Intelligence (AI) replicates human cognitive functions such as learning, reasoning, and language understanding, impacting fields like national security, NLP, and computer vision [19]. Its human-like traits and prosocial behaviors underscore its societal integration [5]. Large language models (LLMs) are central to AI, enabling coherent and contextually relevant human language generation. They play a transformative role in NLP, as seen in AI-Augmented Surveys (AAS), which leverage national data to predict responses, thus enhancing communication technologies [20].

LLMs embody a dual perspective on intelligence, involving quantitative data handling and qualitative reasoning [21], crucial for economic control and information access [22]. However, the limited use of NLP for fact-checking indicates a need for AI tools that meet user needs while ensuring ethical use [23]. Understanding AI and language models' functionality and ethical implications is vital for aligning them with societal expectations. Frameworks like contextual confidence emphasize tools and policies that maintain communication integrity [24]. As AI evolves, its societal and technological impacts demand continuous study and development.

2.2 Cognitive Bias and Its Manifestation in AI

Cognitive biases, mental shortcuts leading to biased actions, significantly affect ethics and discrimination [6]. In AI, especially LLMs, these biases often stem from societal stereotypes in training data [13], leading to flawed logic and complicating ethical deployment [25]. Traditional debiasing methods often overlook root causes. Cognitive-ecological frameworks stress monitoring and control for effective AI decision-making [26], while synthetic personalities in LLMs may cause inconsistencies [27].

A major challenge is 'hallucinations'—errors in machine interpretation exacerbating biases [28]. Addressing cognitive bias requires advanced detection and mitigation strategies. Comprehensive frameworks categorizing generative AI responses by sentiment bias and authority aim to enhance accountability and transparency, aiding informed decisions in public interest and personal well-being [29, 30, 31, 17, 6]. Mitigating these biases is crucial for developing fair, inclusive, and trustworthy AI technologies.

2.3 Ethical Implications of AI Technologies

AI technologies raise ethical concerns, particularly around bias and decision-making. Detecting cognitive biases at the individual level is challenging, especially as LLMs are used in critical areas like medicine, where biases can adversely affect outcomes [6, 25]. Issues arise from mimetic models, which may lead to deception and consent violations [7]. The opacity of AI decision-making complicates ethical concerns, fostering skepticism about AI output fairness [32].

Detecting toxic language in AI outputs is another challenge, as reliance on explicit keywords often misses nuanced toxicity, potentially biasing against minorities [33]. This limitation necessitates sophisticated detection methods. Stereotypical biases in language models raise ethical concerns about their deployment, requiring sustainable debiasing for equitable AI applications [13]. AI technologies like DALL-E 2 highlight the need for frameworks addressing societal impacts of AI-generated content [14].

Addressing AI's ethical implications requires understanding inherent biases and establishing frameworks aligning technologies with societal expectations. This involves recognizing diverse ethical challenges, including social inequalities and unintended consequences, necessitating stakeholder involvement for appropriate solutions. A nuanced approach considering cultural and geographical contexts is essential for mitigating risks and ensuring AI benefits all societal segments, particularly vulnerable populations [15, 34, 17]. Interdisciplinary collaboration and continuous evaluation are vital for promoting fair AI outcomes.

2.4 Gender Stereotypes in AI and Machine Learning

Gender stereotypes are entrenched in AI and machine learning systems, reflecting societal biases in training data. These biases reinforce stereotypes and inequalities, affecting decision-making across domains. In NLP systems, gender bias is pervasive, with training data embedding societal biases that result in harmful stereotypes [35]. This is evident in media representations, where content themes highlight gender biases [36].

Gender bias is prominent in machine translation, where stereotypes misalign pronoun references with gender roles, especially in occupational contexts. LLMs perpetuate these biases, influencing educational choices in male-dominated STEM fields. Bias in language corpora raises concerns about real-world applications, including recruitment and decision-making. While efforts to mitigate these biases exist, challenges remain for morphologically rich languages, often yielding ungrammatical out-

puts. Innovative approaches have shown promise in reducing gender stereotyping without sacrificing grammaticality, achieving accuracy in languages like Spanish and Hebrew [37, 38, 39].

In human-robot interactions, conversational agent design can trigger gender biases, influencing user perceptions. Societal attractiveness perceptions impact gender classification systems, reinforcing stereotypes. AI models like Stable Diffusion 2.1 show variations in gender classification accuracy linked to perceived attractiveness, reflecting social prejudices. This necessitates a multidisciplinary approach to AI development, incorporating insights from cognitive psychology and feminist legal theory to mitigate harmful stereotypes and promote gender diversity in training datasets [40, 41, 42]. The reinforcement of gender stereotypes in models like GPT-2 and GPT-3.5 underscores entrenched biases in training data.

Addressing implicit gender bias in language generation involves developing benchmarks to evaluate stereotypes across contexts, particularly in gendered occupations, as seen in studies of LLMs like GPT-2 and GPT-3.5. These benchmarks uncover gendered word associations and biased narratives, exploring implications for societal perceptions and marginalized communities. Significant variations in gender bias across languages emphasize the need for interdisciplinary approaches to mitigate biases and promote equitable AI systems [43, 44, 38, 45, 46]. Challenges persist in achieving consensus on fairness definitions and auditing opaque NLP models. Acknowledging and addressing these biases can foster a more equitable and inclusive society, promoting fairness and reducing harmful stereotypes.

In recent years, the exploration of cognitive biases within artificial intelligence (AI) and language models has gained significant attention. Understanding these biases is crucial for improving the reliability and ethical implications of AI outputs. Figure 2 illustrates the hierarchical categorization of cognitive biases in AI and language models, highlighting the impact on AI outputs and the challenges associated with identifying and measuring these biases. This figure emphasizes various forms of biases, ethical implications, hurdles encountered in the field, innovative solutions being proposed, as well as the origins, challenges, consequences, and mitigation strategies related to cognitive biases in AI. By examining this framework, researchers can better navigate the complexities of bias in AI systems and develop more effective approaches to address these critical issues.

3 Cognitive Bias in AI and Language Models

3.1 Impact of Cognitive Biases on AI Outputs

Cognitive biases deeply embedded in AI systems, especially large language models (LLMs), significantly impact the fairness and reliability of their outputs, often resulting in skewed and adverse outcomes. These biases manifest in various forms, notably gender biases, which can influence decision-making processes even without explicit gender references. Research highlights that LLMs can expose these biases through specific inputs, indicating the subtle encoding of gender stereotypes [39]. In machine translation, these biases compromise translation accuracy and fairness, often defaulting to masculine forms or using vocal traits for gender determination, particularly for speaker-dependent words [1]. Larger models tend to exhibit stereotypical reasoning, complicating bias mitigation efforts [21]. Cognitive biases also affect tasks such as question answering, where ingrained gender stereotypes can skew model performance, especially within contexts like fairytale narratives, impacting response accuracy and narrative integrity [28].

The ethical implications of these biases are profound, as studies show that language models' responses to moral dilemmas are influenced by biases affecting moral permissibility and intention, raising concerns about AI alignment with societal values [6]. Addressing these biases requires identifying specific model weights encoding stereotypical gender bias, essential for effective debiasing strategies that enhance output fairness and reliability [35]. A comprehensive approach involves robust training frameworks, clear explanation structures, and interdisciplinary collaboration to align AI with societal values and ethical standards [19]. The dominance of ethical discussions from the Global North may perpetuate inequalities, highlighting the necessity for inclusive frameworks that incorporate diverse perspectives [33].

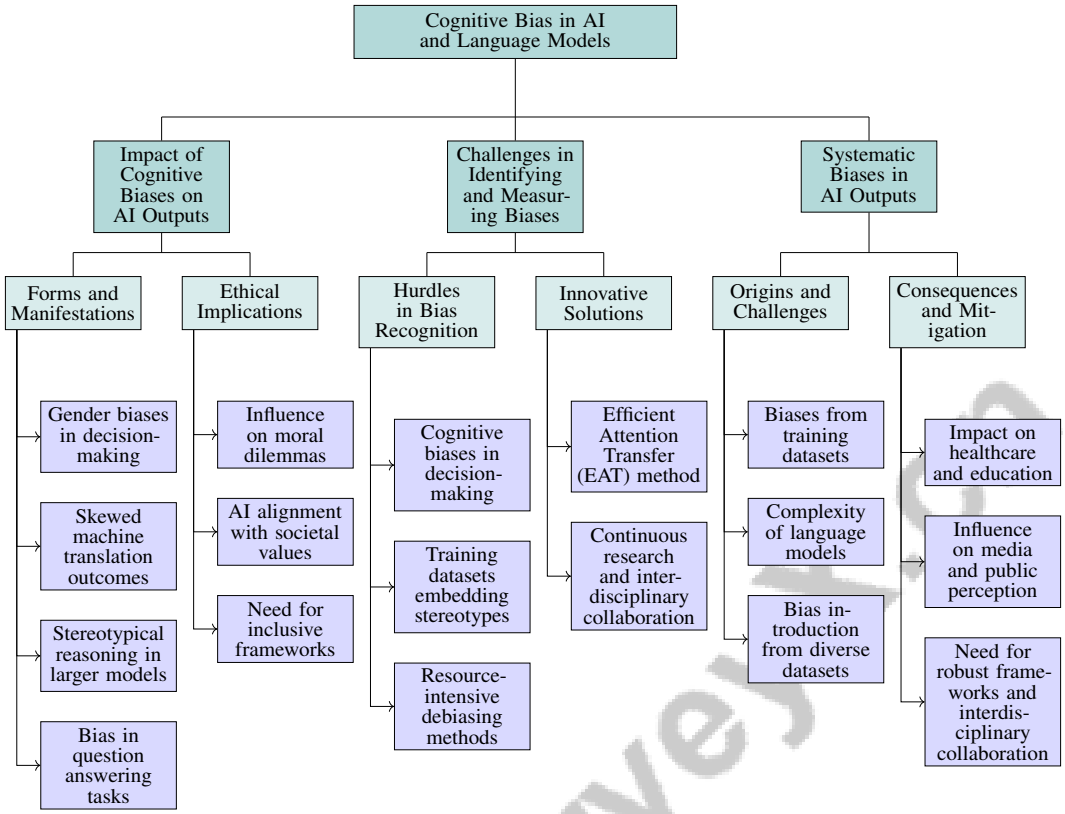

Figure 2: This figure illustrates the hierarchical categorization of cognitive biases in AI and language models, highlighting the impact on AI outputs, challenges in identifying and measuring biases, and systematic biases in AI outputs. It emphasizes forms, ethical implications, hurdles, innovative solutions, origins, challenges, consequences, and mitigation strategies.

3.2 Challenges in Identifying and Measuring Biases

Identifying and measuring biases in AI systems, particularly language models, presents significant challenges that impede fairness and reliability efforts. A major hurdle is effectively recognizing and addressing cognitive biases within decision-making processes, crucial in areas like clinical settings [47]. These biases often originate from training datasets, embedding societal norms and stereotypes that AI systems may inadvertently replicate or amplify [48]. Debiasing AI models is further complicated by the computational resources and manual annotations required for existing methods, often limiting their applicability across various biases [49]. This underscores the need for versatile debiasing techniques that address biases comprehensively without compromising AI performance.

Innovations like the Efficient Attention Transfer (EAT) method offer promising solutions, providing computational efficiency and improved fairness in AI outputs without significant performance loss [50]. The ongoing challenges in bias identification and quantification, especially in LLMs, necessitate continuous research and innovative solutions. These efforts are crucial for developing comprehensive frameworks that effectively mitigate biases from systemic discrimination and cognitive biases, enhancing ethical AI deployment and promoting transparency and accountability in various applications [25, 30, 51, 11, 17]. Interdisciplinary collaboration, drawing insights from cognitive science, ethics, and computer science, is essential for creating AI technologies that align with societal values and promote fairness and inclusivity.

3.3 Systematic Biases in AI Outputs

Systematic biases in AI outputs significantly affect the quality and fairness of AI-generated content, often stemming from training datasets that encode societal stereotypes and norms. This results in

outputs that reflect and perpetuate existing inequalities [35]. Such biases are particularly evident in language models, where gender and racial stereotypes are embedded in the model's parameters, influencing generated text and potentially causing harm [13]. Addressing systematic biases is challenging due to the complexity of language models, which obscures bias origins and mechanisms within their architecture, complicating effective identification and mitigation efforts [28]. Furthermore, using large, diverse datasets can inadvertently introduce biases from various sources, exacerbating the problem and misaligning outputs with societal values and ethical standards [1].

The consequences of systematic biases are extensive, impacting sectors like healthcare, education, and media. In healthcare, biased AI systems can lead to misdiagnoses or unequal treatment recommendations, disproportionately affecting marginalized groups [6]. In education, biased language models can influence assessment fairness and perpetuate stereotypes, adversely affecting students' learning experiences and outcomes [9]. Additionally, AI-generated content reflecting systematic biases can reinforce harmful stereotypes in media, shaping public perception and contributing to societal divisions [36].

To illustrate these dynamics, Figure 3 presents a comprehensive overview of the sources, impacts, and mitigation strategies of systematic biases in AI outputs. This figure highlights the role of training datasets and language models as primary sources of bias, while also identifying healthcare, education, and media as key impact areas. Furthermore, it suggests strategies such as improved data collection, algorithmic fairness techniques, and interdisciplinary collaboration to effectively address these biases. Mitigating systematic biases requires robust frameworks addressing bias root causes within AI systems, including improved data collection for diverse, representative datasets, algorithmic fairness techniques, and interdisciplinary collaboration to create AI technologies aligned with human values, promoting fairness and inclusivity [52]. Addressing these biases can enhance AI system equity and trustworthiness, contributing to positive societal outcomes and responsible AI development and deployment.

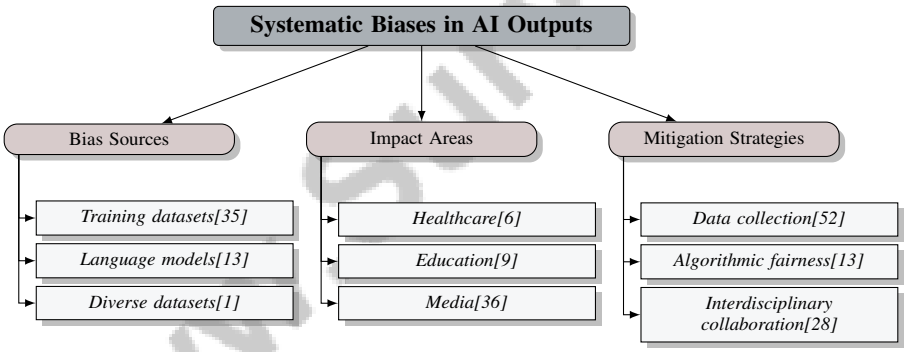

Figure 3: This figure illustrates the sources, impacts, and mitigation strategies of systematic biases in AI outputs, highlighting the role of training datasets, language models, and diverse datasets as bias sources. It identifies healthcare, education, and media as key impact areas and suggests strategies like improved data collection, algorithmic fairness techniques, and interdisciplinary collaboration to address these biases.

4 Ethical Implications of AI Bias

4.1 Algorithmic Bias and Ethical Concerns

Algorithmic bias in AI systems poses ethical challenges by reinforcing societal stereotypes and inequalities, particularly through biased word embeddings in training data [35]. This bias impacts public perception and decision-making, notably in emotion analysis where gender stereotypes are pervasive [36]. Addressing algorithmic bias necessitates understanding moral judgments and accountability, complicated by machine learning's interpretive nature, which can obscure AI outputs [28]. The gender bias propagation in language models highlights the need for ethical frameworks to prevent harmful stereotypes.

Algorithmic bias also affects non-native English writers, leading to unfair assessments and perpetuating disparities. AI detection tools' false positives disproportionately impact non-native speakers,

AI-generated, for reference only.

raising fairness and reliability concerns. Bias reduction techniques may inadvertently perpetuate biases, requiring careful consideration in AI development [27]. Comprehensive evaluation methods reflecting real-world complexities are essential. Sustainable modular debiasing offers efficient debiasing without altering all model parameters, providing insights into reasoning mechanisms and potential biases [13]. The societal impacts of AI technologies, like DALL-E 2, emphasize the need for ethical frameworks addressing AI-generated content [14].

Mitigating algorithmic bias requires implementing Value-Sensitive Design (VSD) methodologies and fostering interdisciplinary collaboration among AI, ethics, psychology, and human-computer interaction researchers. This aligns AI systems with human values, addressing cognitive biases influencing human decision-making and algorithmic outputs. Establishing guidelines for ethical research practices involving human participants enhances transparency and accountability in AI development [25, 53, 30, 11, 8]. By promoting fairness and inclusivity, AI technologies can positively impact societal outcomes and advance responsible AI practices. Prioritizing responsible AI policies and developing multi-stakeholder frameworks are crucial for navigating ethical challenges posed by algorithmic bias.

4.2 Profit-Driven AI and Ethical Decision-Making

The tension between profit motives and ethical decision-making in AI development creates a complex landscape where financial objectives often overshadow ethical considerations. Models like GreedLlama exemplify how profit-oriented training undermines moral decision-making capabilities. Deploying such models in business contexts without sufficient ethical oversight poses substantial risks, as financial incentives may prioritize performance over ethical integrity [54].

Aligning AI systems with ethical standards is challenging due to inherent biases in AI assessments that can skew decision-making processes and lead to ethically questionable outcomes. For instance, limitations in evaluating personality traits within AI systems underscore the difficulties in ensuring unbiased and ethical AI interactions [55]. These biases exacerbate the disconnect between ideal ethical practices and their practical implementation, especially in applying explainable AI (XAI) across diverse contexts [56].

The ethical implications of personalized interactions with large language models (LLMs) highlight the necessity for responsible alignment with human-like traits. While prompt engineering enhances LLM capabilities, achieving consistent empathy and contextual relevance remains a challenge. This inconsistency undermines ethical integrity, emphasizing the need for balancing performance optimization with ethical considerations, particularly in AI-driven processes like peer review, where alignment with scientific norms and transparency is crucial [57, 52, 58, 16].

Navigating the ethical complexities of profit-driven AI requires recognizing the distinct roles of fairness and solidarity in AI applications. By disentangling these concepts and integrating them into AI development, stakeholders can promote ethical practices aligned with societal values and expectations [59]. This effort necessitates embedding ethical frameworks into AI technologies' design and deployment to ensure profit motives do not compromise ethical integrity.

4.3 Cultural and Ethical Challenges

Cultural and ethical challenges posed by biased AI systems are multifaceted, reflecting the diverse contexts of their deployment. A primary challenge is the lack of cross-cultural research on AI perceptions, particularly in non-Western contexts like India, leading to biases in understanding AI's social acceptability and broader implications [60]. These cultural incongruences can result in significant harms and misunderstandings in AI-human interactions, underscoring ethical concerns [61].

As illustrated in Figure 4, which highlights the primary cultural and ethical challenges in AI, the categories of cultural biases, ethical concerns, and issues of AI accountability are interconnected. Each category is supported by relevant literature, emphasizing the need for cross-cultural understanding, ethical transparency, and improved accountability in AI systems. The complexity and opacity of AI systems create information asymmetries, exacerbating cultural and ethical challenges and complicating efforts to ensure transparency and accountability [57]. This opacity is compounded by deeply rooted gender stereotypes in society, complicating recognition and alteration of such biases

in AI-generated content [62]. Current studies often overlook the intersectionality of gender and geographic representation, resulting in a narrow focus that fails to address marginalized groups' needs [63].

Integrating AI into privacy policy assessments presents significant cultural and ethical challenges, as large language models (LLMs) must navigate complex legal, ethical, and data science considerations to ensure privacy and data protection [64]. These challenges highlight the necessity for AI systems to be designed with a nuanced understanding of cultural contexts and ethical frameworks, promoting fairness and inclusivity in their applications.

The role of explanations in AI decision-making processes is critical for ensuring fairness and appropriate reliance on AI systems. Understanding how explanations impact these processes can help mitigate biases and enhance the ethical deployment of AI technologies [65]. However, AI systems lack moral agency due to the absence of constitutive symmetry and shared rational capacities with humans, raising ethical questions about the extent to which AI can be held accountable for its actions [56].

Future research should focus on enhancing LLMs' grounding capabilities, exploring multimodal integrations, and addressing biases to improve their applicability across diverse fields [2]. By tackling these cultural and ethical challenges, AI systems can become more equitable and trustworthy, contributing positively to societal outcomes and advancing responsible development and deployment of AI technologies.

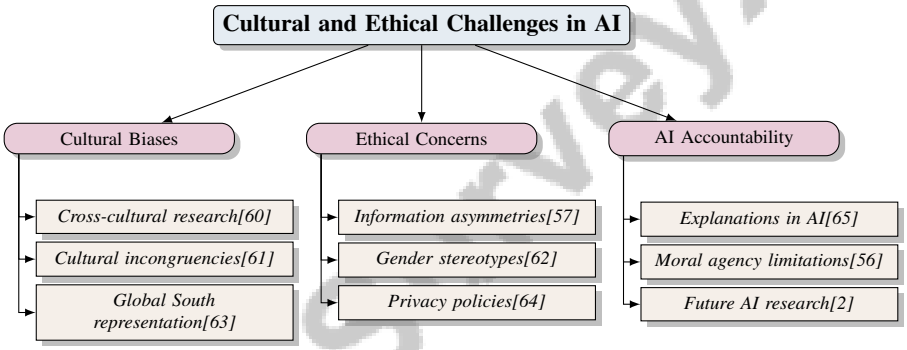

Figure 4: This figure illustrates the primary cultural and ethical challenges in AI, highlighting cultural biases, ethical concerns, and issues of AI accountability. Each category is supported by relevant literature, emphasizing the need for cross-cultural understanding, ethical transparency, and improved accountability in AI systems.

4.4 Privacy, Consent, and Data Misuse

Privacy, consent, and potential data misuse in AI systems are critical ethical concerns requiring robust frameworks. AI technologies generating synthetic data, like synthetic MRI images, pose significant privacy and consent challenges. The potential misuse of such data underscores the importance of addressing these ethical issues to safeguard individual rights and maintain public trust [66].

AI-augmented surveys using large language models (LLMs) to enhance data collection face limitations in predicting unasked opinions and struggle with demographic representativeness, particularly affecting minority groups. The lack of accurate representation can lead to biased outcomes and exacerbate existing inequalities [20]. The ethical considerations surrounding demographic representativeness highlight the need for AI systems to incorporate diverse perspectives and ensure equitable treatment across different population segments.

The intersection of privacy and AI technologies raises concerns about data misuse, where unauthorized access or exploitation of AI-generated data can lead to significant ethical breaches. Adhering to strict privacy protocols and obtaining explicit user consent is essential to prevent misuse and protect sensitive information. This necessitates developing comprehensive privacy policies and consent mechanisms that are transparent and easily understood by users, fostering trust and accountability in AI applications.

Addressing the ethical implications of privacy, consent, and data misuse involves interdisciplinary collaboration and implementing ethical guidelines aligning with societal values and legal standards. By prioritizing individual rights protection through robust privacy policies and promoting responsible AI practices, such as pre-deployment audits and post-deployment accountability, stakeholders can effectively mitigate data misuse risks. This strategic approach enhances AI systems' ethical integrity, fosters transparency, encourages stakeholder involvement, and supports developing diverse and inclusive AI solutions [64, 67, 57, 68, 52].

5 Gender Stereotypes in Language Models

5.1 Encoding and Manifestation of Gender Stereotypes

The encoding of gender stereotypes in language models presents notable challenges in AI development, rooted in training datasets that mirror societal biases [61]. The Gender Equality Presentation (GEP) metric offers a quantitative approach to evaluating gender presentation disparities, surpassing traditional evaluation methods [69]. Large language models (LLMs) often display implicit gender biases, particularly in occupational contexts, as evidenced by studies revealing gender disparities in responses to prompts [45]. This underscores the need for diverse datasets to mitigate biases related to race and language, supporting research that advocates for the use of edge cases to disentangle complex biases [59].

In dialogue settings, stereotypes shape the portrayal of traits and cues, affecting interactions [70]. From a social psychology perspective, stereotypes influence cognitive and affective responses, impacting behavior [71]. Addressing these stereotypes requires comprehensive frameworks categorizing AI applications by use cases, benefits, and ethical considerations, particularly in educational contexts [72]. Understanding the encoding and manifestation of gender stereotypes enables researchers to promote fairness and inclusivity in AI, necessitating ongoing refinement of datasets and methodologies.

5.2 Impact on AI-Generated Content

Gender stereotypes significantly influence AI-generated content, shaping perception and functionality. LLMs trained on datasets reflecting societal biases perpetuate these stereotypes, particularly in occupational roles, reinforcing traditional norms [61, 45]. This perpetuation leads to biased interpretations, affecting decision-making across domains. In educational settings, gender biases in AI systems can compromise assessment fairness, disadvantaging specific groups [72]. Similarly, AI-generated narratives reflecting stereotypes shape public perceptions and media representations [69].

In dialogue systems, stereotypes influence the portrayal of gendered traits and cues, affecting human-AI interactions [70]. Recognizing cognitive and affective responses to these stereotypes highlights the need for AI systems to consider the psychological dimensions of gender biases [71]. Efforts to mitigate gender stereotypes involve developing frameworks that categorize AI applications by use cases, benefits, and ethical considerations, promoting fairness and inclusivity in AI technologies [59].

5.3 Implications for Gender Equality

The implications of gender stereotypes in language models extend beyond technical aspects, influencing societal norms and gender equality. AI systems perpetuating stereotypes reinforce traditional gender roles and contribute to ongoing disparities [73]. In occupational settings, AI systems associate specific professions with genders, reinforcing traditional roles and limiting opportunities for non-conforming individuals. While fair AI algorithms can mitigate career recommendation biases, user resistance to biased systems remains a challenge. Gender stereotypes persist across digital media, with algorithm-driven content curation minimally effective in challenging biases. Human involvement in content creation and curation is more likely to disrupt ingrained biases, suggesting a need for inclusive AI design [40, 74, 36].

Societal norms perpetuated through gender-biased AI outputs influence public perceptions and expectations, affecting social dynamics and hindering gender equality progress. The contrast between fairness-focused and justice-oriented frameworks highlights the limitations of achieving fairness in AI systems without addressing societal issues contributing to inequality [73]. Addressing these implications requires technical debiasing strategies and cultural interventions challenging pervasive

gender role beliefs. This includes leveraging insights from studies on automated counter-stereotypes and examining gender-specific media content, emphasizing comprehensive strategies beyond technical solutions [40, 41, 75, 76]. Adopting justice-oriented frameworks to address inequality root causes is crucial. By integrating these frameworks into AI development, stakeholders can ensure AI systems positively influence societal norms and advance gender equality, fostering a more equitable society.

6 Bias Mitigation Strategies

6.1 Bias Mitigation and Counter-Stereotyping Strategies

Implementing effective bias mitigation strategies in AI is crucial for promoting fairness and inclusivity. A holistic approach involves various methodologies throughout AI development. Algorithmic debiasing techniques are designed to reduce gender bias in language models by refining algorithms while maintaining performance [44]. Incorporating metacognitive regulatory processes into LLMs addresses bias root causes, enhancing practical applications [26].

Frameworks informed by gender theory systematically assess training data for gender bias, ensuring alignment with societal values [11]. Counterfactual data augmentation generates grammatically correct sentences to mitigate stereotypes, particularly in morphologically rich languages [39]. The ADELE method uses debiasing adapters to effectively reduce bias while preserving model parameters [13].

Counter-stereotyping strategies are essential. The Matchmaking for AI co-design process enables stakeholders to collaboratively create AI tools tailored to their needs, enhancing inclusivity [23]. Personalized bias detection emphasizes customizing mechanisms for individual users to effectively mitigate cognitive biases [6].

Strategies minimizing reliance on large labeled datasets, such as those for toxicity detection, address data scarcity and provide robust frameworks for bias mitigation [33]. By employing diverse strategies, AI systems can align with ethical standards and societal expectations, promoting fairness and inclusivity.

6.2 Data Augmentation and Dataset Enhancement

Data augmentation and dataset enhancement are pivotal in reducing bias in AI systems, particularly language models. These techniques expand and manipulate training datasets to create more diverse, representative data, mitigating biased dataset influences on AI outputs. Enhanced variability in training data effectively addresses inherent biases, leading to more equitable AI systems [39].

Generating counterfactual examples challenges and corrects stereotypical associations in training data, fostering balanced gender representation [39]. Dataset enhancement efforts improve quality and diversity by including underrepresented groups, ensuring AI systems reflect a broad range of cultural, linguistic, and demographic characteristics [23].

Integrating data augmentation with machine learning frameworks enhances AI fairness and inclusivity by addressing biases, improving model performance, and ensuring diverse perspectives are represented in AI outputs. This approach promotes ethical AI development, fostering accountability and transparency in machine learning processes [11, 77, 17]. Ongoing refinement of data augmentation and dataset enhancement strategies is crucial for advancing responsible and ethical AI deployment across various domains.

6.3 Model Fine-Tuning and Algorithmic Adjustments

Model fine-tuning and algorithmic adjustments are critical for mitigating bias within LLMs and enhancing fairness. The 'Tiny Heap' method fine-tunes LLMs by replacing gender-exclusive terms with gender-neutral alternatives, reducing gender bias and promoting inclusivity [78]. This highlights the importance of linguistic adjustments in addressing language model biases.

Causal inference techniques offer a robust framework for bias mitigation, exemplified by the Causal Framework for Controllable Text Generation (CFC-TG), which enhances control over generated attributes for precise bias adjustments [79]. Local contrastive editing modifies specific model weights

Method Name	Bias Mitigation Techniques	Model Adjustment Methods	Fairness and Performance Balance
TH[78]	Gender-neutral Variants	Fine-tuning Dataset	Reduced Gender Stereotyping
CFC-TG[79]	Causal Inference Techniques	Confounder Balancing Objectives	High Control Accuracy
LCE[80]	Local Contrastive Editing	Weight Interpolation	Parameter-efficient Editing
REFINE[49]	Reinforcement Learning	Reinforcement Learning	Preserving Model Performance
EAT[50]	Temperature Scaling	Modulating Attention Weights	Minimal Degradation Performance
BAMOL[81]	Multi-objective Learning	Joint Learning Approach	Harmonic Mean
PMSM[27]	-	Algorithmic Adjustments	-

Table 1: Overview of various bias mitigation methods, model adjustment techniques, and their impact on fairness and performance balance in large language models. The table details specific methods such as gender-neutral variants, causal inference techniques, and reinforcement learning, highlighting their effectiveness in reducing bias while maintaining model performance.

encoding gender stereotypes based on reference models, effectively reducing bias while maintaining performance [80].

The REFINE LM method uses reinforcement learning to adjust language model predictive probability distributions, mitigating biases without compromising accuracy [49]. Additionally, the Efficient Attention Transfer (EAT) technique modulates attention distribution entropy in NLP models, enhancing fairness while retaining performance [50].

Multi-objective learning approaches like BAMOL demonstrate achieving fairness in predictions without sacrificing accuracy [81]. The Personality Measurement and Shaping Methodology (PMSM) fine-tunes LLMs through psychometric tests to align AI systems with human-like traits and ethical standards [27]. By employing these diverse techniques, developers can create AI systems that are accurate, reliable, and aligned with societal values, promoting fairness and inclusivity in applications. Table 1 provides a comprehensive summary of different strategies employed for bias mitigation and model adjustment in large language models, emphasizing their contribution to achieving a balance between fairness and performance.

6.4 Algorithmic Fairness and Post-Processing Techniques

Algorithmic fairness is crucial in AI development for mitigating biases and ensuring equitable outcomes. Post-processing techniques modify AI model outputs to comply with ethical standards and societal expectations, addressing bias, discrimination, and misinformation in AI-generated content. These techniques enhance AI applications' reliability and accountability, aligning results with industry best practices, particularly in recruitment and education [52, 82, 16, 77, 17].

A primary goal of algorithmic fairness is preventing AI systems from perpetuating societal inequalities or introducing new biases. This requires a comprehensive understanding of AI technologies' ethical implications, including transparency and consent in implementation [83]. Ensuring transparent AI operations is essential for maintaining public trust and promoting equitable outcomes.

Post-processing techniques encompass diverse methods aimed at refining AI outputs to enhance reliability and ethical compliance. These methods include recalibrating prediction scores, re-ranking outputs for relevance, and implementing fairness constraints to align final results with ethical standards. Such adjustments address biases detected post-deployment, enabling real-time modifications that enhance fairness without extensive retraining [52, 16, 17].

Integrating algorithmic fairness into AI systems requires continuous assessment and adaptation to align with evolving societal norms and ethical standards. The rapid evolution of AI technologies necessitates proactive measures to ensure responsible and ethical use, including frameworks for ongoing monitoring of AI outputs to promptly identify and address biases. Prioritizing algorithmic fairness and leveraging post-processing techniques enables developers to create systems that perform effectively while adhering to the ethical principles of fairness, transparency, and inclusivity.

6.5 Case Studies and Ongoing Research

The field of bias mitigation in AI is evolving, with numerous case studies and ongoing research efforts enhancing the understanding and reduction of biases in AI systems. The ADELE framework effectively mitigated bias across multiple benchmarks while preserving model performance and

knowledge, demonstrating its versatility in debiasing applications [13]. This approach highlights the importance of maintaining model integrity while addressing biases.

In text toxicity mitigation, the CF-Detoxtigtec approach achieved competitive results by balancing technical efficacy with interpretability, bridging automatic toxicity processing and explainable AI [33]. This underscores the necessity for AI systems to provide understandable outputs, fostering trust and accountability.

The BiasMedQA framework evaluates cognitive biases in medical LLMs, enhancing reliability by addressing limitations [27]. This is crucial for ensuring effective and ethically sound AI applications in healthcare, promoting better patient outcomes.

Future research emphasizes developing comprehensive frameworks for understanding cognitive biases in AI, enhancing transparency, and establishing best practices for ethical AI adoption across sectors [25]. Exploring non-binary gender strategies in translation systems and examining the broader implications of gender bias in machine translation are critical for promoting inclusivity in AI-generated content [70].

Additionally, exploring intersectional perspectives and developing frameworks that account for non-binary identities are essential for enhancing digital media representation inclusivity [36]. Addressing these diverse research areas allows AI technologies to align with societal values and ethical standards, promoting fairness and inclusivity.

These case studies and research initiatives underscore the significance of interdisciplinary collaboration and continuous innovation in bias mitigation strategies. By exploring various strategies, including ethical considerations in AI text generation and the global social impacts of AI technologies, researchers can ensure the responsible and equitable deployment of AI. This multifaceted approach addresses the potential for AI to exacerbate social inequalities, particularly in marginalized communities, while emphasizing the importance of diverse perspectives in shaping ethical frameworks. This commitment to ethical AI development can harness AI as a force for positive societal change, fostering innovation while mitigating risks and promoting fairness [63, 15, 34, 17].

7 Responsible AI and Text Generation Control

The dynamic landscape of artificial intelligence (AI) necessitates a framework for responsible AI to guide ethical development and deployment. As AI systems become integrated into societal structures, aligning them with ethical standards and societal values is crucial. This section explores the principles underpinning responsible AI, emphasizing their importance in generative AI systems and their implications for users and developers.

7.1 Principles of Responsible AI

Principles guiding responsible AI development are essential for ensuring ethical operation and societal alignment. Emphasizing transparency, accountability, fairness, and inclusivity, these principles advocate for continuous evaluation to meet evolving ethical challenges. Explainability and customizability in generative AI systems are vital for reducing user metacognitive demands and fostering trust [84]. Robust policies, including pre-deployment audits and post-deployment accountability, are critical for ethical practices throughout the AI lifecycle [67]. Despite high adoption barriers, these measures are crucial for responsible AI integration.

Ensuring fairness and inclusivity from data collection to deployment prevents bias reinforcement and discrimination. Frameworks like FR-Train and MLClean, along with policies such as audits and community involvement, guide stakeholders in designing AI systems that uphold ethical standards [17, 85, 67, 16]. Adhering to these principles enables the creation of systems that perform effectively while aligning with societal expectations.

7.2 Frameworks and Guidelines for Ethical AI Practices

Developing frameworks for ethical AI practices is crucial for aligning AI systems with societal values. These frameworks provide structured methodologies to address ethical challenges, emphasizing transparency, accountability, and inclusivity. By categorizing ethical concerns, they highlight the

necessity of diverse stakeholder engagement and advocate for integrating established ethical standards to enhance oversight [34, 52, 68].

Key components include guidelines promoting transparency in decision-making, fostering user trust [32]. Accountability mechanisms, such as audits and monitoring, ensure developers are responsible for their AI systems' ethical implications [67]. Inclusivity emphasizes avoiding bias and promoting equitable outcomes across applications [61].

Ethical AI frameworks often involve interdisciplinary collaboration, drawing insights from various fields to address multifaceted ethical dilemmas. Integrating diverse perspectives enhances AI applications' integrity, ensuring alignment with moral and epistemic norms [86, 52, 16, 17]. Establishing comprehensive frameworks is vital for addressing AI technologies' societal and ethical challenges, promoting accountability, and involving diverse voices to prevent reinforcing inequalities [18, 34, 57, 52, 63].

7.3 Controlling Text Generation

Controlling AI text generation ensures outputs align with ethical standards. Strategies like prompt engineering guide models to adhere to ethical guidelines, reducing biased content [87]. Causal inference frameworks, such as CFC-TG, enhance control over generated attributes, allowing precise adjustments for ethical standards [79]. Local contrastive editing refines text generation by modifying model weights, promoting fairness without compromising performance [80].

Reinforcement learning strategies, like REFINE LM, adjust predictive probability distributions to mitigate biases while maintaining accuracy [49]. These strategies ensure AI systems produce ethically aligned outputs, advancing responsible AI practices.

7.4 Transparency and Accountability in AI Development

Transparency and accountability are fundamental in AI development, ensuring ethical operation and societal alignment. Transparency involves making AI systems understandable, fostering trust and informed decision-making [32]. Accountability mechanisms, including audits and monitoring, hold developers responsible for ethical implications [67].

Integrating transparency and accountability requires robust frameworks emphasizing inclusivity and fairness, preventing bias perpetuation [61]. These principles ensure AI systems align with societal values and contribute positively to outcomes. Emphasizing transparency and accountability enhances ethical standards, addressing AI systems' unique challenges and promoting equitable outcomes [52, 67].

7.5 Inclusivity and Cultural Cognizance

AI development requires inclusivity and cultural awareness, acknowledging diverse cultural contexts and values. AI systems often emerge from limited countries, risking cultural incongruencies and ethical disparities, especially in low- and middle-income countries. Engaging diverse voices, particularly from underrepresented groups, ensures equitable service and avoids reinforcing marginalization [63, 15, 61].

Cultural cognizance ensures AI systems respect community values. Incorporating diverse perspectives aligns technologies with ethical standards, addressing cultural dependencies [31, 61]. Efforts to enhance inclusivity include developing frameworks emphasizing diversity in training data, mitigating bias and cultural incongruence risks [61, 17].

Interdisciplinary collaboration advances inclusivity and cultural cognizance, integrating insights from ethics and cultural studies to address AI technologies' ethical challenges [16, 17, 18]. Prioritizing these principles ensures AI systems align with societal values, promoting equitable opportunities [63, 67, 61].

8 Conclusion

This survey highlights the critical importance of addressing cognitive biases and ethical challenges in AI technologies, with a particular focus on large language models (LLMs). The presence of ingrained gender biases and stereotypes within these systems underscores the necessity for improved alignment strategies to mitigate such biases. Incorporating ethical considerations from the initial design stages of AI systems is essential to enhance user empowerment and reduce biases. Transparency and continuous advancements in fairness practices are identified as crucial, with certification acting as a safeguard in high-stakes scenarios.

The survey also points to the uneven distribution of LLM access, predominantly favoring wealthier nations and organizations, which exacerbates existing disparities. This inequity highlights the urgent need for developing tools and policies that bolster contextual understanding to effectively manage the complexities introduced by generative AI. The framework introduced by Zmigrod et al. showcases the potential to diminish cognitive biases in AI predictions, achieving significant fairness improvements without compromising accuracy.

An in-depth exploration of cognitive biases as predictors of user engagement further emphasizes the need to investigate cognitive influences in AI and digital content. The ECCOLA deployment model offers a valuable approach for embedding ethical considerations into AI development, fostering stakeholder participation and establishing mechanisms for evaluating ethicality throughout the AI product lifecycle. Additionally, the importance of integrating gender perspectives in robot design and interaction studies is underscored, advocating for more inclusive research that acknowledges diverse gender identities.

References

[1] Jiyan He, Weitao Feng, Yaosen Min, Jingwei Yi, Kunsheng Tang, Shuai Li, Jie Zhang, Kejiang Chen, Wenbo Zhou, Xing Xie, Weiming Zhang, Nenghai Yu, and Shuxin Zheng. Control risk for potential misuse of artificial intelligence in science, 2023.

[2] Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, and Julian Eggert. A glimpse in chatgpt capabilities and its impact for ai research, 2023.

[3] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David T. Newman, and Allison Woodruff. Exciting, useful, worrying, futuristic: Public perception of artificial intelligence in 8 countries, 2021.

[4] Sara Sterlie, Nina Weng, and Aasa Feragen. Generalizing fairness to generative language models via reformulation of non-discrimination criteria, 2024.

[5] Ali Ladak, Jamie Harris, and Jacy Reese Anthis. Which artificial intelligences do people care about most? a conjoint experiment on moral consideration, 2024.

[6] Atanu R Sinha, Navita Goyal, Sunny Dhamnani, Tanay Asija, Raja K Dubey, M V Kaarthik Raja, and Georgios Theocharous. Personalized detection of cognitive biases in actions of users from their logs: Anchoring and recency biases, 2022.

[7] Reid McIlroy-Young, Jon Kleinberg, Siddhartha Sen, Solon Barocas, and Ashton Anderson. Mimetic models: Ethical implications of ai that acts like you, 2022.

[8] Kevin R. McKee. Human participants in ai research: Ethics and transparency in practice, 2024.

[9] The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges.

[10] Sharon Ferguson, Katherine Mao, James Magarian, and Alison Olechowski. Advancing a model of students' intentional persistence in machine learning and artificial intelligence, 2023.

[11] Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene. Mitigating gender bias in machine learning data sets, 2020.

[12] Neeraja Kirtane and Tanvi Anand. Mitigating gender stereotypes in hindi and marathi, 2022.

[13] Anne Lauscher, Tobias Lüken, and Goran Glavaš. Sustainable modular debiasing of language models, 2021.

[14] Eduard Hogea and Josem Rocafortf. The ethical situation of dall-e 2, 2024.

[15] Alexa Hagerty and Igor Rubinov. Global ai ethics: A review of the social impacts and ethical implications of artificial intelligence, 2019.

[16] Laurie A. Schintler, Connie L. McNeely, and James Witte. A critical examination of the ethics of ai-mediated peer review, 2023.

[17] Fnu Neha, Deepshikha Bhati, Deepak Kumar Shukla, Angela Guercio, and Ben Ward. Exploring ai text generation, retrieval-augmented generation, and detection technologies: a comprehensive overview, 2024.

[18] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices, 2019.

[19] Erik Blasch, James Sung, Tao Nguyen, Chandra P. Daniel, and Alisa P. Mason. Artificial intelligence strategies for national security and safety standards, 2019.

[20] Junsol Kim and Byungkyu Lee. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction, 2024.

[21] Nils Körber, Silvan Wehrli, and Christopher Irrgang. How to measure the intelligence of large language models?, 2024.

[22] Vishwas Sathish, Hannah Lin, Aditya K Kamath, and Anish Nyayachavadi. Llempower: Understanding disparities in the control and access of large language models, 2024.

[23] Houjiang Liu, Anubrata Das, Alexander Boltz, Didi Zhou, Daisy Pinaroc, Matthew Lease, and Min Kyung Lee. Human-centered nlp fact-checking: Co-designing with fact-checkers using matchmaking for ai, 2024.

[24] Shrey Jain, Zoë Hitzig, and Pamela Mishkin. Contextual confidence and generative ai, 2024.

[25] Alaina N. Talboy and Elizabeth Fuller. Challenging the appearance of machine intelligence: Cognitive bias in llms and best practices for adoption, 2023.

[26] Florian Scholten, Tobias R. Rebholz, and Mandy Hütter. Metacognitive myopia in large language models, 2024.

[27] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models, 2023.

[28] Remy Demichelis. The hermeneutic turn of ai: Is the machine capable of interpreting?, 2024.

[29] Cesare G. Ardito. Contra generative ai detection in higher education assessments, 2023.

[30] Alice Li and Luanne Sinnamon. Generative ai search engines as arbiters of public knowledge: An audit of bias and authority, 2024.

[31] Lilla Vicsek, Anna Vancsó, Mike Zajko, and Judit Takacs. Exploring lgbtq+ bias in generative ai answers across different country and religious contexts, 2024.

[32] Beatriz Cabrero-Daniel and Andrea Sanagustín Cabrero. Perceived trustworthiness of natural language generators, 2023.

[33] Yau-Shian Wang and Yingshan Chang. Toxicity detection with generative prompt-based inference, 2022.

[34] Richard Benjamins and Idoia Salazar. Towards a framework for understanding societal and ethical implications of artificial intelligence, 2020.

[35] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016.

[36] Vivek Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. Female librarians and male computer programmers? gender bias in occupational images on digital media platforms, 2019.

[37] Damin Zhang, Yi Zhang, Geetanjali Bihani, and Julia Rayz. Hire me or not? examining language model's behavior with occupation attributes, 2025.

[38] P. Cutugno, D. Chiarella, R. Lucentini, L. Marconi, and G. Morgavi. Language, communication and society: a gender based linguistics analysis, 2020.

[39] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology, 2020.

[40] Isar Nejadgholi, Kathleen C. Fraser, Anna Kerkhof, and Svetlana Kiritchenko. Challenging negative gender stereotypes: A study on the effectiveness of automated counter-stereotypes, 2024.

[41] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms, 2020.

[42] Miriam Doh, , and Anastasia Karagianni. "my kind of woman": Analysing gender stereotypes in ai through the averageness theory and eu law, 2024.

[43] Shahar Levy, Koren Lazar, and Gabriel Stanovsky. Collecting a large-scale gender bias dataset for coreference resolution and machine translation, 2021.

[44] Vishesh Thakur. Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications, 2023.

[45] Ishika Joshi, Ishita Gupta, Adrita Dey, and Tapan Parikh. 'since lawyers are males..': Examining implicit gender bias in hindi language generation by llms, 2024.

[46] Haein Kong, Yongsu Ahn, Sangyub Lee, and Yunho Maeng. Gender bias in llm-generated interview responses, 2024.

[47] Yu He Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias, 2024.

[48] Alejandro Peña, Ignacio Serna, Aythami Morales, Julian Fierrez, Alfonso Ortega, Ainhoa Herrarte, Manuel Alcantara, and Javier Ortega-Garcia. Human-centric multimodal machine learning: Recent advances and testbed on ai-based recruitment, 2023.

[49] Rameez Qureshi, Naïm Es-Sebbani, Luis Galárraga, Yvette Graham, Miguel Couceiro, and Zied Bouraoui. Refine-lm: Mitigating language model stereotypes via reinforcement learning, 2024.

[50] Abdelrahman Zayed, Goncalo Mordido, Samira Shabanian, and Sarath Chandar. Should we attend more or less? modulating attention for fairness, 2024.

[51] Abel Salinas, Louis Penafiel, Robert McCormack, and Fred Morstatter. "im not racist but...": Discovering bias in the internal knowledge of large language models, 2023.

[52] Jose Berengueres and Marybeth Sandell. Applying standards to advance upstream downstream ethics in large language models, 2023.

[53] Gopal P. Sarma, Nick J. Hay, and Adam Safron. Ai safety and reproducibility: Establishing robust foundations for the neuropsychology of human values, 2018.

[54] Jeffy Yu, Maximilian Huber, and Kevin Tang. Greedllama: Performance of financial value-aligned large language models in moral reasoning, 2024.

[55] Erik Derner, Dalibor Kučera, Nuria Oliver, and Jan Zahálka. Can chatgpt read who you are?, 2024.

[56] Joshua L. M. Brand and Luca Nannini. Does explainable ai have moral value?, 2023.

[57] Peter Cihon, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum. Ai certification: Advancing ethical practice by reducing information asymmetries, 2021.

[58] Anders Braarud Hanssen and Stefano Nichele. Ethics of artificial intelligence demarcations, 2019.

[59] James Brusseau. Using edge cases to disentangle fairness and solidarity in ai ethics, 2021.

[60] Vishakha Agrawal, Serhiy Kandul, Markus Kneer, and Markus Christen. From oecd to india: Exploring cross-cultural differences in perceived trust, responsibility and reliance of ai and human experts, 2023.

[61] Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence, 2022.

[62] Jiao Sun, Tongshuang Wu, Yue Jiang, Ronil Awalegaonkar, Xi Victoria Lin, and Diyi Yang. Pretty princess vs. successful leader: Gender roles in greeting card messages, 2021.

[63] Cathy Roche, Dave Lewis, and P. J. Wall. Artificial intelligence ethics: An inclusive global discourse?, 2021.

[64] Irem Aydin, Hermann Diebel-Fischer, Vincent Freiberger, Julia Möller-Klapperich, Erik Buchmann, Michael Färber, Anne Lauber-Rönsberg, and Birte Platow. Assessing privacy policies with ai: Ethical, legal, and technical challenges, 2024.

[65] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. Explanations, fairness, and appropriate reliance in human-ai decision-making, 2024.

[66] Muhammad Suhaib Shahid, Gleb E. Yakubov, and Andrew P. French. Ethics of generating synthetic mri vocal tract views from the face, 2024.

[67] Emily Hadley. Prioritizing policies for furthering responsible artificial intelligence in the united states, 2022.

[68] Atoosa Kasirzadeh. Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence, 2021.

[69] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models, 2023.

[70] Hillary Dawkins, Isar Nejadgholi, and Chi kiu Lo. Wmt24 test suite: Gender resolution in speaker-listener dialogue roles, 2024.

[71] Angelina Wang, Xuechunzi Bai, Solon Barocas, and Su Lin Blodgett. Measuring machine learning harms from stereotypes: requires understanding who is being harmed by which errors in what ways, 2024.

[72] Maria J. Molina, Amy McGovern, Jhayron S. Perez-Carrasquilla, and Robin L. Tanamachi. Using generative artificial intelligence creatively in the classroom: Examples and lessons learned, 2024.

[73] Cynthia L. Bennett and Os Keyes. What is the point of fairness? disability, ai and the complexity of justice, 2019.

[74] Clarice Wang, Kathryn Wang, Andrew Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. User acceptance of gender stereotypes in automated career recommendations, 2021.

[75] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Quantifying and reducing stereotypes in word embeddings, 2016.

[76] Diego Kozlowski, Gabriela Lozano, Carla M. Felcher, Fernando Gonzalez, and Edgar Altszyler. Gender bias in magazines oriented to men and women: a computational approach, 2020.

[77] Junhua Liu, Wendy Wan Yee Hui, Roy Ka-Wei Lee, and Kwan Hui Lim. Fairness and performance in harmony: Data debiasing is all you need, 2024.

[78] Marion Bartl and Susan Leavy. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in llms, 2024.

[79] Zhiting Hu and Li Erran Li. A causal lens for controllable text generation, 2022.

[80] Marlene Lutz, Rochelle Choenni, Markus Strohmaier, and Anne Lauscher. Local contrastive editing of gender stereotypes, 2024.

[81] Procheta Sen and Debasis Ganguly. Towards socially responsible ai: Cognitive bias-aware multi-objective learning, 2020.

[82] Vincent Freiberger and Erik Buchmann. Fairness certification for natural language processing and large language models, 2024.

[83] Nils Köbis, Philipp Lorenz-Spreen, Tamer Ajaj, Jean-Francois Bonnefon, Ralph Hertwig, and Iyad Rahwan. Artificial intelligence can facilitate selfish decisions by altering the appearance of interaction partners, 2023.

[84] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. The metacognitive demands and opportunities of generative ai, 2024.

[85] Steven Euijong Whang, Ki Hyun Tae, Yuji Roh, and Geon Heo. Responsible ai challenges in end-to-end machine learning, 2021.

[86] Seraj A. M. Mostafa, Md Z. Islam, Mohammad Z. Islam, Fairose Jeehan, Saujanna Jafreen, and Raihan U. Islam. Critical role of artificially intelligent conversational chatbot, 2023.

[87] Daniil Filienko, Yinzhou Wang, Caroline El Jazmi, Serena Xie, Trevor Cohen, Martine De Cock, and Weichao Yuwen. Toward large language models as a therapeutic tool: Comparing prompting techniques to improve gpt-delivered problem-solving therapy, 2024.

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.