

This manuscript is a preprint of an article accepted for publication in *Social Science & Medicine*. The published version is available at <https://doi.org/10.1016/j.socscimed.2026.119855>

AI Representations of Mental Health Stigma: Quantifying the Gap Between AI Estimations and Human Perceptions Across Gender Groups Using GPT-Based LLMs

Soojong Kim¹, Hye Min Kim², Kwanho Kim³, and Dongdong Zhu¹

¹ Department of Communication, University of California Davis, United States

² Department of Communication, University of Massachusetts Boston, United States

³ Department of Media, College of Politics and Economics, Kyung Hee University, South Korea

Author Note

Soojong Kim  <https://orcid.org/0000-0002-1334-5310>

Correspondence concerning this article should be addressed to Soojong Kim, 1 Shields Ave, Kerr Hall #361, Davis, CA 95616, United States. Email: sjokim@ucdavis.edu

Abstract

The increasing use of artificial intelligence (AI) in mental health applications raises concerns about its potential to internalize and perpetuate societal biases. This research evaluates both disparities and misalignment in how large language models (LLMs), GPT-4 and GPT-5, represent perceptions of mental health across gender groups. The research comprises two studies. Study 1 applied widely adopted measures of Social Distance, Stereotypes, and Attitudes Toward Seeking Help to assess whether the AI models depict systematic gender differences. Study 2 compared AI estimations of Personal, Self, and Perceived Stigma with large-scale real-world human data from the Healthy Minds Study ($N = 76,406$) to evaluate alignment between AI estimations and human perceptions. Results from Study 1 showed the AI models consistently portrayed men as having the greatest desire for social distance from someone with mental health problems, relatively higher Stereotypes scores, and the least favorable attitudes toward help-seeking, compared with women and nonbinary individuals. Study 2 confirmed that this ordinal pattern partially aligned with human-reported data for Personal and Self Stigma. However, the AI models systematically underestimated Personal Stigma and Perceived Stigma while substantially overestimating Self Stigma for all gender groups. These findings reveal that, although the AI models have learned relative gender-based patterns reflecting human experiences to a certain extent, they function as distorted portrayals, highlighting the need for rigorous, context-specific validation against real-world data to ensure AI's safe and equitable deployment.

Keywords: artificial intelligence, large language models, mental health, alignment, stigma, gender differences

AI Representations of Mental Health Stigma: Quantifying the Gap Between AI Estimations and Human Perceptions Across Gender Groups Using GPT-Based LLMs

The rise of artificial intelligence (AI) is transforming public health and healthcare, bringing both challenges and opportunities to improve health surveillance, disease diagnosis, healthcare access, health communication, and data-driven decision-making (Matheny et al., 2020; Panteli et al., 2025). A growing number of AI applications now focus on mental health, ranging from conversational agents that provide social support to AI models that analyze human behaviors and linguistic markers of distress (Lee et al., 2021; Rajpurkar et al., 2022). AI-powered platforms are especially appealing to young people, a demographic that represents early adopters of new technologies and experiences an increasing prevalence of mental health challenges (Keyes et al., 2019; Vigo et al., 2016). The anonymity, accessibility, and affordability of AI-powered tools provide a low-barrier entry point for individuals who might otherwise hesitate to seek professional help (Luxton, 2020; Rodríguez-Rivas et al., 2022).

However, concerns remain regarding the accuracy, fairness, and disparities of AI in representing human experiences and perceptions of mental health (Carr, 2020; Timmons et al., 2023). Biases and misrepresentations in AI models in mental health that power interactive systems, conversational agents, and technologies designed to analyze human behavior and linguistic data could lead to multiple levels of harm. First, they can lead to biased and inaccurate diagnosis and advice, potentially less encouraging or even discouraging people with mental health concerns from seeking necessary medical and social support. Second, they can shape users' perceptions of how mental illness is viewed by others, possibly causing them to overestimate public stigma and stereotypes, thereby creating an additional barrier to help-seeking. Lastly, at a broader level, such biases can perpetuate structural inequities by amplifying

underrepresentation or misrepresentation of specific populations within digital health systems (Gianfrancesco et al., 2018; Seyyed-Kalantari et al., 2021; Timmons et al., 2023).

The concepts of stigma (e.g., negative stereotypes toward mental health problems, unwillingness to engage with people experiencing mental health issues) and attitudes toward help-seeking are central to understanding barriers to mental healthcare. Research has shown that stigma is a major factor discouraging individuals from accessing professional help, often due to fears of shame and negative social judgment (Clement et al., 2015).

It should be noted that stigma is not a unitary construct. The literature on mental illness stigma distinguishes public stigma from self-stigma and differentiates cognitive, affective, and behavioral components of stigmatizing responses (Corrigan et al., 2003; Fox et al., 2018). Public stigma refers to the ways in which members of the general public stigmatize people with mental illness, whereas self-stigma refers to the internalization of devaluation by people who may themselves experience mental health difficulties (Corrigan, 2000; Corrigan & Watson, 2002). Within public stigma, Corrigan et al. (2003) further elaborate that stereotypes reflect negative beliefs, prejudice reflects negative emotional reactions, and discrimination reflects behavioral responses such as rejection, avoidance, or social distance.

As AI systems increasingly serve as central interaction points for mental health support, how AI estimates or “portrays” these constructs in society becomes more important. For instance, if AI models depict that people have overly high levels of stigma and interact with their users based on these estimations, it could worsen users’ concerns about how their mental illness would be perceived by others, amplifying the very barriers that public health initiatives seek to overcome (Corrigan et al., 2012).

This challenge is further complicated by gender disparities. Prior research suggests that gender differences in stigma and help-seeking may arise from social norms governing emotional expression, vulnerability, and autonomy (Addis & Mahalik, 2003; Gonzalez et al., 2005; Holzinger et al., 2012; Mackenzie et al., 2006). Masculinity norms often reward self-reliance and emotional control, which may increase resistance to formal help-seeking and intensify stigma (Addis & Mahalik, 2003; Möller-Leimkühler, 2002; Seidler et al., 2016). By contrast, women often face fewer social penalties for emotional disclosure and support-seeking, which may contribute to more favorable attitudes toward professional help and higher rates of service use (Clement et al., 2015; Gonzalez et al., 2005; Mackenzie et al., 2006). These tendencies are socially patterned rather than biologically fixed (Courtenay, 2000; Möller-Leimkühler, 2002).

The present research therefore focuses on gender because gender-linked differences in public stigma, self-stigma, and help-seeking are well established in the literature and are directly relevant to the use of AI in mental health contexts. We treat this focus as a targeted first step rather than a comprehensive account of all socially structured variation in AI portrayals of mental health stigma.

If an AI model recognizes and reproduces these differences, its function becomes dependent on the gender of users or people being analyzed by AI. This creates a dynamic whose direction is not immediately clear, since the model may either help close gender gaps or contribute to widening them. For example, would it encourage men to seek professional help regarding a mental health problem more than women (to offset men's negative attitudes) or less (to accommodate their preferences and maintain user engagement)? AI models may also attend selectively to different aspects of an individual's mental health experience depending on gender (e.g., Plaza Del Arco et al., 2024).

These concerns motivate the need to examine how AI models portray mental health stigma and help-seeking across gender groups. Additional background on the promise, risks, and validation challenges of AI in mental health is provided in the Supplementary Online Material (SOM) Section 1. The next section reviews algorithmic bias in mental health applications and prior work on gendered representations in AI outputs.

Algorithmic Bias in Mental Health AI

Biases in AI models are a substantial barrier to the safe and equitable deployment of AI in mental health (Agarwal et al., 2023; Carr, 2020; Graham et al., 2019; Lee et al., 2021; Straw, 2020; Timmons et al., 2023). AI models learn patterns from the data on which they are trained, and if the data reflects existing societal stereotypes and inequities, the models can easily learn and encode them (e.g., Koenecke et al., 2020; Obermeyer et al., 2019; Williams, Brooks, et al., 2018). However, biases do not originate solely from training data. Biases can infiltrate the AI lifecycle at every stage, from unrepresentative data collection and biased labels to flawed model construction and evaluation (Agarwal et al., 2023; I. Chen et al., 2021; Luxton, 2020; Timmons et al., 2023). The result is AI systems that may perform inequitably across social groups, worsening care disparities for already underserved populations.

Gender bias has become a pressing concern in the development and application of AI for mental health. Substantial evidence has documented gender-based biases and discrimination in algorithms and AI in other domains, such as facial recognition technologies (Menezes et al., 2021), web search systems (Makhortykh et al., 2021; Vlasceanu & Amodio, 2022), housing (Asplund et al., 2020), and employment (L. Chen et al., 2018; Imana et al., 2021). More broadly, recent research shows that even LLMs presented as explicitly unbiased can still form biased associations (Bai et al., 2025). In mental-health-related contexts, Lauderdale et al. (2025) found

that LLMs evaluating vignette characters with borderline personality disorder (BPD) exhibited gender bias, showing more negative reactions and greater public stigma directed toward a woman with BPD. Other studies similarly suggest that LLM outputs can vary systematically by gender and mental health condition. For example, Schnepper et al. (2025) found that an LLM assumed a greater impairment in mental health-related quality of life for men compared with women when evaluating identical case vignettes for anorexia and bulimia. Lin et al. (2022) demonstrated that language models are consistently less likely to associate men with mental health conditions compared with women, and this disparity widens in sentences describing treatment-seeking behavior. Chansiri et al. (2024) reported that LLMs over-associate men with narcissistic personality disorder and women with BPD. Also, Noguero et al. (2023) analyzed AI algorithms trained to detect anorexia nervosa on social media and found that the models demonstrated higher false negative rates for female users than male users. The implications of these biased patterns from the perspective of AI explainability are discussed further in the SOM Section 1.

The present research investigates how large language models (LLMs), one of the most rapidly evolving and widely adopted technologies in AI, represent perceptions of mental health across different gender groups. It aims to identify the specific gender-based patterns present in AI estimations and to determine how much these patterns are aligned with or deviated from real-world human perceptions.

Here, it is important to distinguish between two forms of potential bias of AI. The first is “disparity,” wherein AI models produce systematically different outcomes or associations depending on social groups. A model may generate different outputs for specific groups, potentially due to differences in the internalized assumptions and learned associations (Garg et

al., 2018; Obermeyer et al., 2019). Disparities in AI outcomes may or may not reflect underlying differences in real-world outcomes, and the absence of such disparities does not necessarily indicate accuracy or fairness (Hardt et al., 2016). The second, and more intricate, form is “misalignment,” or an imperfect mirroring of reality. A model may be misaligned if its estimations fail to capture real-world human experiences and perceptions accurately, not only through over- or underestimation but also by misrepresenting the true direction or magnitude of real-world group differences (Amodei et al., 2016; Gabriel, 2020). These two forms of bias do not necessarily co-occur. A model may generate outcomes showing disparity between social groups that nonetheless reflect real-world observations (disparity with alignment). On the other hand, a model may produce outcomes showing no meaningful group differences, even when such equivalence does not reflect the reality (parity with misalignment). Although it is known that certain statistical criteria requiring group parity cannot be satisfied at the same time as accuracy-based criteria leading to perfect alignment (Chouldechova, 2017; Kleinberg et al., 2017), it is still meaningful to evaluate the levels and patterns of these imperfections in important real-world AI applications such as mental health.

The present research is structured to evaluate both forms of potential representational AI bias. The research comprises two sequential studies designed to analyze these representations in depth. In Study 1, LLMs were prompted to generate a large volume of responses reflecting stigma toward mental illness and attitudes toward help-seeking across different gender groups. This study was designed to identify the existence and patterns of disparity in the AI portrayals of mental health attitudes across gender groups. Study 2 then moves beyond identification to assess alignment, by quantifying the specific nature and magnitude of these AI estimations against large-scale real-world human data. In this study, AI-estimated stigma levels were directly

compared with population-level survey data from the Healthy Minds Study (HMS), one of the largest annual surveys of mental health in the U.S. postsecondary population (Healthy Minds Network, 2023). The focus on gender allowed us to compare AI estimations with sufficiently large corresponding groups in the HMS data.

The prior studies show that AI models encode gender-based disparities and produce outcomes that reflect broader societal biases. Extending this pattern to the domain of stigma, and considering the tendency of AI models to encode traditional masculinity norms that associate men with emotional suppression and resistance to help-seeking (Addis & Mahalik, 2003; Lin et al., 2022; Plaza-del-Arco et al., 2024; Seidler et al., 2016), Study 1 was guided by the following research question and hypothesis:

RQ1 (Disparity in Study 1). Do AI models exhibit gender-group disparities in their estimations of Social Distance, Stereotypes toward people with mental health problems, and Attitudes Toward Seeking Professional Help?

H1 (Disparity in Study 1). AI models will exhibit gender disparities such that men are estimated to show higher Social Distance, stronger Stereotypes toward people with mental health problems, and less favorable Attitudes Toward Seeking Professional Help than women and nonbinary individuals.

Little research has examined how AI models portray the stigma and attitudes of gender groups or attempted to quantify misalignments. The present research addresses this gap by comparing AI-estimated attitudes of gender groups toward mental health and by quantitatively evaluating LLMs' estimations of mental health stigma against real-world survey data. Because Study 2 used the HMS data as an empirical benchmark, we did not assume in advance that a specific gender-group ordering would necessarily appear in the human-reported data for each

stigma measure. Therefore, Study 2 was guided by non-directional research questions that first examine whether AI estimations correspond to the gender-group ordering observed in the HMS data and then quantify the magnitude and direction of AI-human discrepancies:

RQ2 (Misalignment in Study 2: Ordering alignment). To what extent do AI estimations of Personal Stigma, Self Stigma, and Perceived Stigma show the same ordering of gender groups as the HMS data?

RQ3 (Misalignment in Study 2: Magnitude and direction). For each gender group and each stigma measure, what is the magnitude and direction of the discrepancy between AI estimations and human perceptions?

Methods

AI-Estimated Mental Health Attitudes Across Gender Groups (Study 1)

The primary objective of Study 1 was to investigate whether leading LLMs exhibit systematic disparities in their depiction of mental health perceptions and attitudes across different gender groups.

Measures

The constructs examined in Study 1 capture related but distinct aspects of public responses to mental illness: Social Distance as a behavioral expression of rejection or discrimination, Stereotypes as the cognitive component of public stigma, and Attitudes Toward Seeking Help as a related treatment-seeking outcome closely tied to stigma processes (Corrigan et al., 2003; Fox et al., 2018).

Social Distance. We used six items adapted from the social distance measures commonly used in stigma research (Hoffner & Cohen, 2018; Martin et al., 2000; Pescosolido et al., 2021; S.-L. L. Williams et al., 2018). The AI models were prompted to estimate how willing members

of a given gender group in the U.S. (male, female, or nonbinary individuals aged 18 or older) would be to engage in a given behavior involving a person with a mental health problem, such as “make friends with the person.” Items were rated on a four-point scale from “definitely willing” (1) to “definitely unwilling” (4). Responses of 1 or 2 were recoded as 0, and responses of 3 or 4 were recoded as 1. The six recoded responses were summed to compute an overall score ranging from 0 to 6. Higher scores indicated less willingness and thus greater perceived social distance from people with a mental health problem.

Stereotypes. We used seven items adapted from an established measure (Hoffner & Cohen, 2015, 2018) to assess negative stereotypical beliefs, such as “A person with depression is dangerous.” The AI models rated each statement on a seven-point Likert scale from “strongly disagree” (1) to “strongly agree” (7). Three items were reverse-coded, and ratings were averaged to compute an overall score ranging from 1 to 7. Higher scores indicated stronger endorsement of negative stereotypes toward people with mental illness.

Attitudes Toward Seeking Help. We used ten items from the Attitudes Toward Seeking Professional Psychological Help Scale–Short Form (ATSPPH-SF; Elhai et al., 2008; Fischer & Farina, 1995), such as “If I believed I was having a mental breakdown, my first inclination would be to get professional attention.” The AI models rated each item using a four-point scale from “disagree” (0) to “agree” (3). Five items were reverse-coded, and all ten ratings were summed to compute an overall score ranging from 0 to 30. Higher scores indicated more positive attitudes toward seeking professional help. Further measurement details are provided in the SOM Section 2.1, and example prompts are presented in Table S1.

Generation of LLM Responses

For each construct and gender group, GPT-4 (gpt-4o-2024-05-13) generated 1,000 independent, stateless responses with randomized item order and temperature set to the midpoint of the model's available range. GPT-5 mini (gpt-5-mini-2025-08-07) was also used to generate responses for the robustness analysis. API procedures, parameter settings, reasoning configuration, and model-selection rationale are provided in the SOM Section 2.2.

Comparison Between AI Estimation and Human Perception (Study 2)

Study 2 aimed to compare AI-estimated perceptions with actual human perceptions using large-scale real-world survey data. We used human responses from the HMS, one of the largest and most comprehensive population-level surveys of mental health among postsecondary students in the U.S. (Healthy Minds Network, 2023). We requested and received the most recent available HMS dataset from the Healthy Minds Network, which leads the HMS, in May 2024. Our analysis was based on the 2022-2023 academic year dataset, the most recent dataset available to us at the time, which included responses from 76,406 respondents. The mean age of respondents was 23.28 years ($SD = 7.19$). The present analysis focused on three gender categories—female (68.52%), male (25.74%), and nonbinary (2.59%)—to align the human-response data with the categories examined in the AI estimation task, capturing 97% of respondents. More details on gender identity assessment are reported in the SOM Section 2.3.

Measures

Study 2 used three three-item stigma measures from the HMS (Healthy Minds Network, 2023). Personal Stigma assessed an individual's own stigmatizing attitudes toward others with a mental illness, such as "I feel that receiving mental health treatment is a sign of personal failure." Self Stigma assessed the extent to which an individual would feel embarrassed or ashamed about having a mental illness themselves, such as "I would feel inadequate if I went to a therapist for

psychological help.” Perceived Stigma assessed an individual’s perception of the stigmatizing attitudes held by others, such as “Most people feel that receiving mental health treatment is a sign of personal failure.” Personal Stigma and Perceived Stigma used six-point scales from strongly agree (1) to strongly disagree (6), whereas Self Stigma used a five-point scale from strongly disagree (1) to strongly agree (5). Items were coded or reverse-coded as needed so that higher values consistently indicated greater stigma, and the three items for each measure were averaged to create overall scores. In the human HMS sample, the internal consistency was $\alpha = .73$ for Perceived Stigma, $.93$ for Self Stigma, and $.74$ for Personal Stigma. We retained the HMS labels of the constructs while recognizing that Personal Stigma aligns most closely with respondents’ own public-stigma-related beliefs, Perceived Stigma with perceived stigma, and Self Stigma with treatment-related internalized or self-stigma (Corrigan et al., 2003; Fox et al., 2018).

AI Estimation Procedure

Study 2 followed the Study 1 procedure for response generation. The prompts used the exact HMS item wording and response scales and asked each model to estimate the responses of a male, female, or nonbinary student enrolled at a U.S. college or university, living in the U.S., and aged 18 years or older. This student-specific reference population aligned the AI estimation task with the HMS sample. An example prompt is presented in Table S1, and further procedural and model details are provided in the SOM Section 2.5.

To analyze and compare the data, we first calculated the mean for each of the three stigma measures by gender group, using both the AI-generated data and the human-reported data from the HMS. These means were then directly compared to identify discrepancies between the AI estimations and human perceptions. To quantify the magnitude of these discrepancies, we also

calculated AI bias for each stigma measure and gender group. The bias score was defined as the difference between the AI-estimation mean and the human-report mean (i.e., $AI\ bias = M_{AI} - M_{Human}$). A positive bias score indicates that an AI model overestimates the level of stigma for that group compared with human reports, while a negative score indicates underestimation. Because the stigma measures differed in response range, we also calculated range-standardized AI bias: $standardized\ AI\ bias = (M_{AI} - M_{Human}) / (scale\ maximum - scale\ minimum)$. This approach preserves the direction of bias while making the magnitude of discrepancies more comparable across measures, and was used for visualization purposes.

Methodological Considerations

Analysis of Group-Level Differences

The Mann-Whitney test was used to analyze differences in constructs between gender groups estimated by an AI model, between gender groups in the human-reported data, and between AI estimations and human reports for the same gender group. This is a nonparametric test that does not rely on the assumption of normality. The present research reports the Mann-Whitney U scores, p -values, and Common-Language Effect Sizes (CLES; f) for this test (the Mann-Whitney U statistics are reported in the SOM due to their large numerical values). The CLES represents the probability that a randomly selected score from one group will be higher than a randomly selected score from another group, ranging from 0 to 1. CLES values greater than 0.5 indicate that the first group tends to have higher scores than the second, whereas values less than 0.5 indicate the opposite. When comparing different gender groups within AI-estimated results or within human-reported responses, Bonferroni adjustment was applied to reduce the risk of Type I error due to multiple comparisons, and the adjusted p -values (p_{adj}) are reported.

To compare AI-estimated scores across gender groups and with human-reported scores, we examined the mean scores for each construct. For completeness, we also report inferential statistics, but these should be interpreted with caution. Because this research involves simulated data using AI, the resulting p -values for AI estimations primarily reflect the number of model runs rather than the true variability of the underlying “population,” that is, all possible LLM responses. In such simulations, even trivial differences can reach statistical significance if repeated a sufficient number of times. Therefore, the key insight lies not in whether group differences in AI estimations are statistically significant, but in the consistency and stability of an AI model’s estimations across measures, items, and contexts. The reported statistical results thus serve a descriptive rather than inferential purpose, illustrating that the observed mean differences represent reproducible features of an AI model’s internal representations rather than random noise.

Approach for LLM Response Generation

In both Study 1 and Study 2, we instructed the AI models to estimate a gender group’s responses to specific survey questions. We used established survey measures as standardized measurement tools rather than assuming that conventional human psychometric properties automatically transfer to model outputs. This design enabled comparisons across gender groups and, in Study 2, direct comparison with the HMS human responses. The labels “female individuals,” “male individuals,” and “nonbinary individuals” were used as operational categories to maintain parallel wording across prompts. These labels were also intended to match the gender categories used in the HMS survey, which assessed gender identity using response options that included “male,” “female,” and “gender non-binary.” We used zero-shot prompting without examples of desired answers, preserving correspondence with the original survey items

and reducing the possibility that researcher-provided examples would shape the responses. This task evaluates model portrayals under controlled survey prompting rather than model behavior in naturalistic conversation.

We examined whether 1,000 iterations were sufficient using running-mean convergence diagnostics. We also repeated the GPT-4 analyses under lower and higher temperature settings, a paraphrased prompt template, and a role-framing prompt. Detailed rationale is provided in the SOM Section 2.6, and the procedures and results of the robustness analyses are reported in the SOM Section 3.

Results

AI-Estimated Mental Health Attitudes Across Gender Groups (Study 1)

To address RQ1 and test H1, we examined whether the AI models exhibited gender-group disparities across social distance, stereotypes, and attitudes toward seeking professional help, and whether the direction of these disparities was consistent with the hypothesized ordering.

The results showed consistent gender-based patterns in the AI estimations of mental health perceptions, which are summarized in Figure 1. Overall, the AI models depicted male individuals as perceiving greater social distance to people with mental illness, higher levels of stereotypes, and less positive attitudes toward seeking help, compared with female and nonbinary individuals. While the section below details the primary results from GPT-4 (Figures 1a-1c), these patterns were found to be robust in the additional analysis using GPT-5 (Figure 1d).

First, disparities among gender groups emerged for Social Distance (Figure 1a). GPT-4 estimated that male individuals ($M = 4.02$, $SD = 0.58$) perceive the greatest social distance, significantly more than female individuals ($M = 3.87$, $SD = 0.51$; $p_{\text{adj}} < .001$, $f = 0.56$) and nonbinary individuals ($M = 3.59$, $SD = 0.65$; $p_{\text{adj}} < .001$, $f = 0.66$). Female individuals were also

estimated to perceive significantly greater social distance than nonbinary individuals ($p_{\text{adj}} < .001$, $f = 0.61$).

An item-level analysis showed this pattern (male > female > nonbinary) was consistent. For instance, on the item “make friends with the person” (“Friend” in Figure 1a), the estimated unwillingness was highest for male individuals ($M = 1.94$, $SD = 0.46$), followed by female individuals ($M = 1.82$, $SD = 0.46$), and lowest for nonbinary individuals ($M = 1.69$, $SD = 0.51$). This specific ordering of means was consistent for all six items, even including the item “have the person marry into their family” (“Marry” in Figure 1a), which exhibited a strong ceiling effect. For this item, GPT-4 estimated all gender groups would be “definitely unwilling” (a score of 4) in most cases (male: 99.1%, female: 98.4%, nonbinary: 96.0%), but the proportion of these responses still followed the same ordinal pattern.

Second, the AI estimations for Stereotypes (Figure 1b) followed the same trend. Male individuals were depicted as holding relatively stronger stereotypes ($M = 2.97$, $SD = 0.30$) than female individuals ($M = 2.75$, $SD = 0.22$; $p_{\text{adj}} < .001$, $f = 0.72$) and nonbinary individuals ($M = 2.60$, $SD = 0.28$; $p_{\text{adj}} < .001$, $f = 0.83$). GPT-4 also estimated that female individuals hold significantly stronger negative stereotypes than nonbinary individuals ($p_{\text{adj}} < .001$, $f = 0.67$). However, all three group means were below the scale midpoint of 4 on the seven-point scale, which indicates that the AI model leaned toward disagreement with stereotypical statements in absolute terms, despite the relative differences across gender groups.

This overall pattern for Stereotypes was consistent at the item level. For instance, for the item “A person with depression is dangerous” (“Dangerous” in Figure 1b), the mean agreement estimated by the AI model was highest for male individuals ($M = 2.09$, $SD = 0.34$), lower for female individuals ($M = 2.00$, $SD = 0.18$), and lowest for nonbinary individuals ($M = 1.68$, $SD =$

0.47). This specific ordering (male > female > nonbinary) was observed for six of the seven Stereotypes items. The only item that deviated was “A person with depression is able to maintain a regular job” (“Regular job” in Figure 1b), where although male individuals’ reverse-coded mean score ($M = 4.07$, $SD = 0.74$) was still higher than female individuals’ ($M = 3.89$, $SD = 0.73$), nonbinary individuals’ mean ($M = 4.08$, $SD = 0.75$) was estimated to be similar to male individuals’.

Lastly, the AI estimations of Attitudes Toward Seeking Help also differed across gender groups (Figure 1c). GPT-4 consistently estimated less positive attitudes toward professional help-seeking among male individuals, which is aligned with the stronger stigma reflected in the other two constructs (Social Distance and Stereotypes). The AI model estimated significantly less positive attitudes toward seeking professional help among male individuals ($M = 19.54$, $SD = 3.56$) compared with female individuals ($M = 24.53$, $SD = 1.61$; $p_{\text{adj}} < .001$, $f = 0.09$) and also compared with nonbinary individuals ($M = 25.27$, $SD = 2.02$; $p_{\text{adj}} < .001$, $f = 0.07$). The model also estimated that female individuals had significantly less positive attitudes toward seeking help compared with nonbinary individuals ($p_{\text{adj}} < .001$, $f = 0.37$).

This pattern was robust at the item level. The AI model estimated male individuals to have the lowest (least positive) mean score on all 10 items. The gap between male individuals and the other two groups was pronounced for all items. For example, on the item “Self solve first” (i.e., “A person should work out his or her own problems; getting psychological counseling would be a last resort.”), the mean for male individuals ($M = 1.52$, $SD = 0.75$) was substantially lower than for female ($M = 2.27$, $SD = 0.47$) and nonbinary individuals ($M = 2.56$, $SD = 0.50$). In contrast, the pattern of attitudes between female and nonbinary individuals was more mixed at

the item level. While nonbinary individuals had a significantly more positive overall score, this did not hold for every item.

To confirm the robustness of these findings, we conducted the same analysis using GPT-5. As illustrated in Figure 1d, the results were consistent with those from GPT-4, replicating the same gender-based ordinal patterns across all three constructs.

Comparison of AI Estimations and Human Perceptions (Study 2)

Study 2 compared the AI estimations of mental health stigma with the human perceptions of stigma in the real world (Figure 2). To address RQ2 and RQ3, we focused on the extent to which AI estimations aligned with the ordinal patterns observed in human perceptions and on the magnitude and direction of discrepancies between AI estimations and human perceptions.

Consistent with the results of Study 1, the AI models estimated higher levels of Personal Stigma (an individual's own stigmatizing attitudes toward others with a mental illness) and Self Stigma (the extent to which an individual would feel embarrassed or ashamed to have a mental illness themselves) among male individuals than female and nonbinary individuals. The results of Study 2 also reveal discrepancies in magnitude between AI estimations and human perceptions of stigma but show that the AI estimations may reflect the ordinal patterns observed in human perceptions. While this section details the primary results from GPT-4 (Figure 2), these patterns were largely replicated in the validation analysis using GPT-5 (Figure S5 in the SOM).

The comparison revealed two distinct types of findings, which are presented in detail below. First, we analyzed the discrepancies in magnitude, by comparing the absolute mean stigma scores from the AI estimations with the human-reported means for each gender group. Second, we evaluated the alignment of ordinal patterns, by comparing the relative differences

between gender groups, e.g., whether the group with the highest stigma in the human data was also estimated to have the highest stigma by AI.

GPT-4 overestimated the level of Self Stigma for all three gender groups (Figures 2b, 2e, 2g). This overestimation was substantial across all groups: female ($M_{AI} = 2.80$, $M_{Human} = 1.89$; $p < .001$, $f = 0.79$), male ($M_{AI} = 2.91$, $M_{Human} = 2.19$; $p < .001$, $f = 0.73$), and nonbinary individuals ($M_{AI} = 2.44$, $M_{Human} = 1.68$; $p < .001$, $f = 0.78$).

Contrarily, the AI model underestimated the levels of Personal Stigma (Figures 2a, 2d, 2g) for all groups. This degree of underestimation was greatest for male individuals ($M_{AI} = 1.08$, $M_{Human} = 1.77$; $p < .001$, $f = 0.28$) but was also significant for female ($p < .001$, $f = 0.33$) and nonbinary individuals ($p < .001$, $f = 0.40$). However, it should be noted that AI estimates of Personal Stigma were concentrated near the lower end of the response scale. Therefore, between-group differences for this construct should be interpreted with caution as reflecting a compressed, floor-adjacent ordering rather than strong evidence of substantial group separation. The full item-level response distributions are presented in the SOM Section 4.

A significant underestimation was also found for Perceived Stigma (an individual's perception of the stigmatizing attitudes held by others): male ($M_{AI} = 2.50$, $M_{Human} = 2.99$; $p < .001$, $f = 0.38$) and female individuals ($M_{AI} = 2.49$, $M_{Human} = 2.92$; $p < .001$, $f = 0.41$). GPT-4's estimated mean of Perceived Stigma for nonbinary individuals ($M_{AI} = 2.78$) was also lower than the human mean ($M_{Human} = 2.98$; $p = .009$, $f = 0.47$).

Despite these differences in absolute magnitude, the AI model's estimations partially reflected the relative differences and ordinal patterns observed in human perceptions. For Personal Stigma, the HMS data showed that male individuals ($M = 1.77$, $SD = 1.03$) reported significantly higher Personal Stigma than female individuals ($M = 1.44$, $SD = 0.78$; $p_{adj} < .001$, f

= 0.60) and nonbinary individuals ($M = 1.22$, $SD = 0.55$; $p_{adj} < .001$, $f = 0.67$). The AI model's estimations captured this same ordinal pattern: male individuals ($M = 1.08$, $SD = 0.21$) were depicted by the AI model as holding significantly higher Personal Stigma than female individuals ($M = 1.01$, $SD = 0.07$; $p_{adj} < .001$, $f = 0.56$) and nonbinary individuals ($M = 1.01$, $SD = 0.05$; $p_{adj} < .001$, $f = 0.56$) (Figures 2a, 2d).

A similar alignment was found for Self Stigma. In the human-reported data, male individuals ($M = 2.19$, $SD = 1.01$) reported the highest levels of stigma, greater than female individuals ($M = 1.89$, $SD = 0.90$; $p_{adj} < .001$, $f = 0.59$) and nonbinary individuals ($M = 1.68$, $SD = 0.87$; $p_{adj} < .001$, $f = 0.66$). GPT-4 estimations, although they were uniformly higher than human perceptions, mirrored this exact order: male individuals ($M = 2.91$, $SD = 0.72$) were estimated to have the highest Self Stigma, significantly higher than female individuals ($M = 2.80$, $SD = 0.73$; $p_{adj} = .003$, $f = 0.54$) and nonbinary individuals ($M = 2.44$, $SD = 0.68$; $p_{adj} < .001$, $f = 0.68$) (Figures 2b, 2e).

However, this alignment was less clear for Perceived Stigma. In the human responses, male individuals ($M = 2.99$, $SD = 1.22$) reported significantly higher Perceived Stigma than female individuals ($M = 2.92$, $SD = 1.23$; $p_{adj} < .001$, $f = 0.52$), while nonbinary individuals ($M = 2.98$, $SD = 1.24$) were not significantly different from male individuals ($p_{adj} = 1.000$, $f = 0.51$). On the other hand, the AI model estimated that, compared with nonbinary individuals ($M = 2.78$, $SD = 0.76$), both female ($M = 2.49$, $SD = 0.62$; $p_{adj} < .001$, $f = 0.40$) and male individuals ($M = 2.50$, $SD = 0.53$; $p_{adj} < .001$, $f = 0.39$) had significantly lower Perceived Stigma, while estimating male and female individuals as nearly identical ($p_{adj} = 1.000$, $f = 0.50$) (Figures 2c, 2f).

Finally, we verified these patterns using GPT-5 (Figure S5 in the SOM). This additional analysis confirmed that the model exhibited similar systematic biases, substantially

overestimating Self Stigma while underestimating Personal and Perceived Stigma. Regarding gender-based patterns, the model preserved the primary trends found in its predecessor: male individuals were consistently estimated to hold the highest levels of Personal and Self Stigma, whereas nonbinary individuals were estimated to perceive the highest level of stigma from others, despite a shift in the relative ranking between female and nonbinary individuals for Self Stigma.

Additional sensitivity checks using GPT-4 also supported the robustness of the main findings. Across these checks, the main Study 1 finding that male individuals were estimated to show greater Social Distance, stronger Stereotypes, and less positive Attitudes Toward Seeking Professional Help was consistently reproduced, and the main Study 2 findings regarding partial ordinal alignment and systematic discrepancies in magnitude between AI estimations and human responses were largely retained. The detailed results of these sensitivity checks are presented in the SOM Section 3.

As a descriptive index, we calculated Cronbach's alpha for the AI-generated responses, while recognizing that these coefficients should not be interpreted in the same way as psychometric reliability estimates from human responses. The coefficients varied across constructs and gender groups, with values ranging from near zero to .80 in Study 1 and from .32 to .79 in Study 2, indicating that the AI-generated composites did not exhibit uniformly strong scale-level coherence. All coefficients are reported in the SOM Section 4.

Discussion

This research investigated how GPT-based LLMs portray perceptions of mental health problems and attitudes toward help-seeking across different gender groups. It also compared the AI estimations against actual human perceptions. Related work has used different

methodological approaches to examine biased associations and stigma in LLM outputs more directly (e.g., Bai et al., 2025; Lauderdale et al., 2025), whereas the present research focused on disparities in prompted estimations across gender groups and their misalignment with human-reported perceptions. The results show that, under the present role-based prompting strategy, the AI models produced consistent, gender-differentiated representations of mental health attitudes that partially aligned with human-reported perceptions, while also showing substantial discrepancies in magnitude.

Across two studies, the GPT-based AI models consistently depicted men as holding the highest levels of stigma and the least favorable attitudes toward seeking help, followed by women and then nonbinary individuals. In Study 1, this pattern emerged robustly across measures of Social Distance, Stereotypes, and Attitudes Toward Seeking Help. The gendered differences were consistently reproduced across models, iterations, and measures under the present prompting setup.

In Study 2, the comparison with the HMS data confirmed that the AI estimations captured the relative gender differences observed in human responses for Personal Stigma and Self Stigma. In both the human data and the AI estimations, men had the highest levels of these stigma types. This alignment indicates that the AI models' estimates tracked the direction of real-world group differences for these constructs.

However, the analysis also discovered substantial discrepancies in the absolute magnitude of stigma between AI estimations and human responses. The AI models underestimated Personal Stigma (stigmatizing attitudes toward others) and Perceived Stigma (perception of others' attitudes) while significantly overestimating Self Stigma (internalized shame). These systematic inaccuracies and biases indicate that while the AI models recognize that some gender groups are

more likely to hold certain attitudes than others, they misrepresent and distort how strongly those attitudes are held. These discrepancies also suggest that the AI models underestimate social forms of stigma while overestimating internal stigma.

Underestimating social forms of stigma while overestimating internal stigma may be a direct consequence of the LLMs' training architecture. Although the architectures and development processes of LLMs have become increasingly opaque amid the growing dominance of proprietary AI models including GPT models used in this research (Liao & Vaughan, 2024), here we would like to point out factors that may plausibly account for these patterns. On one hand, the model's "safety alignment," likely achieved through Reinforcement Learning from Human Feedback (RLHF), may penalize the generation of content that appears judgmental or attributes strong negative biases to certain social groups (Ganguli et al., 2023; Ouyang et al., 2022). When prompted to estimate Personal or Perceived Stigma, this alignment may act as an internal guardrail, forcing the model toward more socially desirable, attenuated responses and thus underestimating the true prevalence of these social stigmas. This possibility is especially relevant to Personal Stigma, where AI estimates were concentrated near the lower end of the response scale. This floor-adjacent pattern may partly reflect the model's tendency not to attribute stigmatizing attitudes to certain social groups, rather than a calibrated estimation of low stigma. On the other hand, the model's foundational pre-training corpus is vast and likely abundant with first-person narratives of mental distress, such as those from social media platforms like Twitter and Reddit (Longpre et al., 2024), which often emphasize personal feelings of shame, guilt, and inadequacy. When prompted for Self Stigma, the model may draw heavily on these pervasive narratives, which are less directly moderated by the socially-focused internal guardrails. This might create an internal tension in AI models, where they

simultaneously present society as more accepting than it is, and the individual as more internally stigmatized than they are.

The findings have important implications for the development and deployment of AI applications in mental health, such as conversational agents and diagnostic tools. The presence of consistent, gendered representations in the GPT-based AI models highlights the risk that these technologies may perpetuate and amplify existing stereotypes (Carr, 2020; Timmons et al., 2023). For instance, if an AI-powered chatbot operates under the internalized assumption that men are resistant to seeking help, it may not prioritize suggesting help-seeking behaviors to male users or may inform them that men typically do not seek professional help. Also, the misalignments between AI estimations and human perceptions highlight the risk of AI acting as a distorted social portrayal. As users increasingly turn to AI for information and support, these models can shape their understanding of how mental illness is viewed in society. The systematic underestimation of Personal and Perceived Stigma might lead users to believe that societal attitudes are more accepting than they truly are, potentially leaving them unprepared for real-world discrimination or invalidating their experiences (Shen et al., 2024). The exaggeration of Self Stigma, on the other hand, could intensify users' feelings of isolation and shame, creating additional barriers to help-seeking (Carr, 2020). For example, when users seek advice, interactive AI agents influenced by an exaggerated depiction of Self Stigma might frame guilt, shame, or personal weakness as common responses to mental illness. This could lead the AI to deprioritize help-seeking suggestions or cause users to perceive such stigmatization as normal and unavoidable.

These risks are closely tied to questions of AI literacy. If AI systems can shape users' perceptions of what others believe about mental illness, then it becomes critical that users and

practitioners understand how to interpret such outputs. Users may interpret AI-generated statements about how “people” or specific groups think about mental illness as descriptive facts about society, when in fact such outputs are model-generated representations that may only partially align with real-world data. Improving AI literacy in mental health contexts therefore requires teaching users and practitioners to treat these outputs as fallible approximations that should be interpreted cautiously, especially when they concern stigma, help-seeking, or socially sensitive group differences.

The findings also emphasize the need for rigorous, context-specific validation of AI models against real-world data. As the present research demonstrates, even advanced models can exhibit clear inaccuracies and biases that are not immediately apparent without direct comparison to human responses. Ensuring the fairness and equity of AI in mental health requires moving beyond general benchmarks to evaluating model performance across diverse demographic groups and specific psychological constructs.

This research is not without limitations. First, the analysis focused on an LLM family, OpenAI’s GPT. While these models represent the current state-of-the-art, the observed patterns may not generalize to other models with different architectures or training data. Future research should examine a broader range of LLMs to determine the universality of these findings. Second, the comparison in Study 2 relied on the HMS data based on responses from U.S. postsecondary students. While this population is undergoing increasing mental health difficulties (Mojtabai & Olfson, 2020; Patel & Lewis, 2023), this population is not representative of the general population or other cultural contexts. Related to this limitation, Study 1 and Study 2 used different reference populations in the AI prompts. Therefore, comparisons across the studies should be interpreted as convergence across related but not population-identical contexts. Some

differences across studies, including the less consistent Perceived Stigma finding, may partly reflect this difference in reference populations. Also, there may be a temporal mismatch between the data informing the AI models' representations and the 2022-2023 HMS benchmark. Because stigma-related attitudes can shift over time, this temporal gap may have contributed to some of the observed misalignment between AI estimations and human-reported perceptions. Third, the prompting strategy used here was designed for standardized measurement and direct comparison with human survey data. It does not approximate the multi-turn, personalized, emotionally embedded way in which many users interact with conversational AI systems in practice. Accordingly, the present findings should not be taken as a direct estimate of how stigma-related bias would unfold in extended real-world conversations. Lastly, the present research focused specifically on gender and did not examine how multiple aspects of social identity and demographic background, such as race, ethnicity, age, immigration status, or socioeconomic position, may jointly shape AI estimations of mental health perceptions. This narrower focus was intentional. Before addressing more complex intersectional configurations, research must first establish rigorous building blocks by examining individual social dimensions in relation to their theoretical foundations, real-world patterns, and AI-specific implications. We therefore view the present research as an important first step and encourage future work to extend this approach to other individual social dimensions as well. Such building blocks will ultimately support intersectional research that treats each dimension as more than an added variable and more meaningfully examines both the distinct contribution of each dimension and their interactions.

Conclusion

The current research provides a detailed examination of how advanced AI models based on GPT portray mental health stigma across gender groups. The findings reveal that the models

display consistent, gendered representations that reflect traditional stereotypes, accurately capturing the relative patterns of stigma observed in human populations for some constructs but not others. The AI models also exhibit significant discrepancies in the magnitude of stigma, compared to human perceptions in the real world. These results highlight the complex ways in which AI models represent human perceptions and underscore the need for ongoing evaluation, bias mitigation, and transparent validation to ensure the safe and equitable deployment of AI in the sensitive domain of mental health.

References

- Addis, M. E., & Mahalik, J. R. (2003). Men, masculinity, and the contexts of help seeking. *American Psychologist*, 58(1), 5–14. <https://doi.org/10.1037/0003-066X.58.1.5>
- Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, 12(1), 100702. <https://doi.org/10.1016/j.hlpt.2022.100702>
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv. <https://doi.org/10.48550/arXiv.1606.06565>
- Asplund, J., Eslami, M., Sundaram, H., Sandvig, C., & Karahalios, K. (2020). Auditing Race and Gender Discrimination in Online Housing Markets. *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, 14(1), 24–35. <https://doi.org/10.1609/icwsm.v14i1.7276>
- Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8), e2416228122. <https://doi.org/10.1073/pnas.2416228122>
- Carr, S. (2020). ‘AI gone mental’: Engagement and ethics in data-driven technology for mental health. *Journal of Mental Health*, 29(2), 125–130. <https://doi.org/10.1080/09638237.2020.1714011>
- Chansiri, K., Wei, X., & Chor, K. H. B. (2024). Addressing Gender Bias: A Fundamental Approach to AI in Mental Health. *2024 5th International Conference on Big Data Analytics and Practices (IBDAP)*, 107–112. <https://doi.org/10.1109/IBDAP62940.2024.10689686>

- Chen, I., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, 4(1), 123–144.
<https://doi.org/10.1146/annurev-biodatasci-092820-114757>
- Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the Impact of Gender on Rank in Resume Search Engines. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3173574.3174225>
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>
- Clement, S., Schauman, O., Graham, T., Maggioni, F., Evans-Lacko, S., Bezborodovs, N., Morgan, C., Rüsch, N., Brown, J. S. L., & Thornicroft, G. (2015). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine*, 45(1), 11–27.
<https://doi.org/10.1017/S0033291714000129>
- Corrigan, P. W. (2000). Mental health stigma as social attribution: Implications for research methods and attitude change. *Clinical Psychology: Science and Practice*, 7(1), 48–67.
<https://doi.org/10.1093/clipsy.7.1.48>
- Corrigan, P. W., Markowitz, F. E., Watson, A., Rowan, D., & Kubiak, M. A. (2003). An Attribution Model of Public Discrimination Towards Persons with Mental Illness. *Journal of Health and Social Behavior*, 44(2), 162–179. <https://doi.org/10.2307/1519806>
- Corrigan, P. W., Morris, S. B., Michaels, P. J., Rafacz, J. D., & Rüsch, N. (2012). Challenging the Public Stigma of Mental Illness: A Meta-Analysis of Outcome Studies. *Psychiatric Services*, 63(10), 963–973. <https://doi.org/10.1176/appi.ps.201100529>

- Corrigan, P. W., & Watson, A. C. (2002). The paradox of self-stigma and mental illness. *Clinical Psychology: Science and Practice*, 9(1), 35–53. <https://doi.org/10.1093/clipsy.9.1.35>
- Courtenay, W. H. (2000). Constructions of masculinity and their influence on men's well-being: A theory of gender and health. *Social Science & Medicine*, 50(10), 1385–1401. [https://doi.org/10.1016/S0277-9536\(99\)00390-1](https://doi.org/10.1016/S0277-9536(99)00390-1)
- Elhai, J. D., Schweinle, W., & Anderson, S. M. (2008). Reliability and validity of the Attitudes Toward Seeking Professional Psychological Help Scale-Short Form. *Psychiatry Research*, 159(3), 320–329. <https://doi.org/10.1016/j.psychres.2007.04.020>
- Fischer, E. H., & Farina, A. (1995). Attitudes toward seeking professional psychological help: A shortened form and considerations for research. *Journal of College Student Development*, 36(4), 368–373.
- Fox, A. B., Earnshaw, V. A., Taverna, E. C., & Vogt, D. (2018). Conceptualizing and measuring mental illness stigma: The mental illness stigma framework and critical review of measures. *Stigma and Health*, 3(4), 348–376. <https://doi.org/10.1037/sah0000104>
- Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3), 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
- Ganguli, D., Askill, A., Schiefer, N., Liao, T. I., Lukošiuūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., Drain, D., Li, D., Tran-Johnson, E., Perez, E., Kernion, J., Kerr, J., Mueller, J., Landau, J., Ndousse, K., ... Kaplan, J. (2023). *The capacity for moral self-correction in large language models*. arXiv. <https://doi.org/10.48550/arXiv.2302.07459>

- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635–E3644. <https://doi.org/10.1073/pnas.1720347115>
- Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11), 1544–1547. <https://doi.org/10.1001/jamainternmed.2018.3763>
- Gonzalez, J. M., Alegria, M., & Prihoda, T. J. (2005). How do attitudes toward mental health treatment vary by age, gender, and ethnicity/race in young adults? *Journal of Community Psychology*, 33(5), 611–629. <https://doi.org/10.1002/jcop.20071>
- Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Current Psychiatry Reports*, 21(11), 116. <https://doi.org/10.1007/s11920-019-1094-0>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*, 3323–3331.
- Healthy Minds Network. (2023). *Healthy Minds Study among Colleges and Universities, 2023 [Data set]*. Healthy Minds Network, University of Michigan, University of California Los Angeles, Boston University, and Wayne State University. <https://healthymindsnetwork.org/research/data-for-researchers>
- Hoffner, C. A., & Cohen, E. L. (2015). Portrayal of Mental Illness on the TV Series *Monk*: Presumed Influence and Consequences of Exposure. *Health Communication*, 30(10), 1046–1054. <https://doi.org/10.1080/10410236.2014.917840>

- Hoffner, C. A., & Cohen, E. L. (2018). Mental Health-Related Outcomes of Robin Williams' Death: The Role of Parasocial Relations and Media Exposure in Stigma, Help-Seeking, and Outreach. *Health Communication, 33*(12), 1573–1582.
<https://doi.org/10.1080/10410236.2017.1384348>
- Holzinger, A., Floris, F., Schomerus, G., Carta, M. G., & Angermeyer, M. C. (2012). Gender differences in public beliefs and attitudes about mental disorder in western countries: A systematic review of population studies. *Epidemiology and Psychiatric Sciences, 21*(1), 73–85. <https://doi.org/10.1017/S2045796011000552>
- Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for Discrimination in Algorithms Delivering Job Ads. *Proceedings of the Web Conference 2021, WWW '21*, 3767–3778.
<https://doi.org/10.1145/3442381.3450077>
- Keyes, K. M., Gary, D., O'Malley, P. M., Hamilton, A., & Schulenberg, J. (2019). Recent increases in depressive symptoms among US adolescents: Trends from 1991 to 2018. *Social Psychiatry and Psychiatric Epidemiology, 54*(8), 987–996.
<https://doi.org/10.1007/s00127-019-01697-8>
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In C. H. Papadimitriou (Ed.), *Leibniz International Proceedings in Informatics (LIPIcs)* (Vol. 67, p. 43:1-43:23). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. <https://doi.org/10.4230/LIPICS.ITCS.2017.43>
- Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689.
<https://doi.org/10.1073/pnas.1915768117>

- Lauderdale, S., Griffin, S. A., Lahman, K. R., Mbaba, E., & Tomlinson, S. (2025). Unveiling public stigma for borderline personality disorder: A comparative study of artificial intelligence and mental health care providers. *Personality and Mental Health, 19*(2), e70018. <https://doi.org/10.1002/pmh.70018>
- Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H.-C., Paulus, M. P., Krystal, J. H., & Jeste, D. V. (2021). Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 6*(9), 856–864. <https://doi.org/10.1016/j.bpsc.2021.02.001>
- Liao, Q. V., & Vaughan, J. W. (2024). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5). <https://doi.org/10.1162/99608f92.8036d03b>
- Lin, I., Njoo, L., Field, A., Sharma, A., Reinecke, K., Althoff, T., & Tsvetkov, Y. (2022). Gendered Mental Health Stigma in Masked Language Models. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2152–2170). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.139>
- Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2024). A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence, 6*(8), 975–987. <https://doi.org/10.1038/s42256-024-00878-8>

Luxton, D. D. (2020). Ethical implications of conversational agents in global public health.

Bulletin of the World Health Organization, 98(4), 285–287.

<https://doi.org/10.2471/BLT.19.237636>

Mackenzie, C. S., Gekoski, W. L., & Knox, V. J. (2006). Age, gender, and the underutilization of mental health services: The influence of help-seeking attitudes. *Aging & Mental Health*, 10(6), 574–582. <https://doi.org/10.1080/13607860600641200>

Makhortykh, M., Urman, A., & Ulloa, R. (2021). Detecting Race and Gender Bias in Visual Representation of AI on Web Search Engines. In L. Boratto, S. Faralli, M. Marras, & G. Stilo (Eds.), *Advances in Bias and Fairness in Information Retrieval* (pp. 36–50). Springer International Publishing. https://doi.org/10.1007/978-3-030-78818-6_5

Martin, J. K., Pescosolido, B. A., & Tuch, S. A. (2000). Of fear and loathing: The role of “disturbing behavior,” labels, and causal attributions in shaping public attitudes toward people with mental illness. *Journal of Health and Social Behavior*, 41(2), 208–223. <https://doi.org/10.2307/2676306>

Matheny, M. E., Whicher, D., & Thadaney Israni, S. (2020). Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. *JAMA*, 323(6), 509–510. <https://doi.org/10.1001/jama.2019.21579>

Menezes, H. F., Ferreira, A. S. C., Pereira, E. T., & Gomes, H. M. (2021). Bias and Fairness in Face Detection. *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 247–254. <https://doi.org/10.1109/SIBGRAPI54419.2021.00041>

Mojtabai, R., & Olfson, M. (2020). National Trends in Mental Health Care for US Adolescents. *JAMA Psychiatry*, 77(7), 703–714. <https://doi.org/10.1001/jamapsychiatry.2020.0279>

- Möller-Leimkühler, A. M. (2002). Barriers to help-seeking by men: A review of sociocultural and clinical literature with particular reference to depression. *Journal of Affective Disorders*, 71(1–3), 1–9. [https://doi.org/10.1016/S0165-0327\(01\)00379-2](https://doi.org/10.1016/S0165-0327(01)00379-2)
- Noguero, D., Ramírez-Cifuentes, D., Ríssola, E. A., & Freire, A. (2023). Gender bias when using artificial intelligence to assess anorexia nervosa on social media: Data-driven study. *Journal of Medical Internet Research*, 25, e45184. <https://doi.org/10.2196/45184>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. <https://doi.org/10.48550/arXiv.2203.02155>
- Panteli, D., Adib, K., Buttigieg, S., Goiana-da-Silva, F., Ladewig, K., Azzopardi-Muscat, N., Figueras, J., Novillo-Ortiz, D., & McKee, M. (2025). Artificial intelligence in public health: Promises, challenges, and an agenda for policy makers and public health institutions. *The Lancet Public Health*, 10(5), e428–e432. [https://doi.org/10.1016/S2468-2667\(25\)00036-2](https://doi.org/10.1016/S2468-2667(25)00036-2)
- Patel, B. P., & Lewis, B. (2023). Responding to the Crisis in College Mental Health: A Call to Action. *The Journal of Pediatrics*, 257, 113390. <https://doi.org/10.1016/j.jpeds.2023.02.036>

- Pescosolido, B. A., Halpern-Manners, A., Luo, L., & Perry, B. (2021). Trends in Public Stigma of Mental Illness in the US, 1996-2018. *JAMA Network Open*, 4(12), e2140202.
<https://doi.org/10.1001/jamanetworkopen.2021.40202>
- Plaza-del-Arco, F. M., Curry, A. C., Curry, A., Abercrombie, G., & Hovy, D. (2024). Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7682–7696.
<https://doi.org/10.18653/v1/2024.acl-long.415>
- Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38. <https://doi.org/10.1038/s41591-021-01614-0>
- Rodríguez-Rivas, M. E., Cangas, A. J., Cariola, L. A., Varela, J. J., & Valdebenito, S. (2022). Innovative Technology–Based Interventions to Reduce Stigma Toward People With Mental Illness: Systematic Review and Meta-analysis. *JMIR Serious Games*, 10(2), e35099. <https://doi.org/10.2196/35099>
- Schnepper, R., Roemmel, N., Schaefer, R., Lambrecht-Walzinger, L., & Meinschmidt, G. (2025). Exploring biases of large language models in the field of mental health: Comparative questionnaire study of the effect of gender and sexual orientation in anorexia nervosa and bulimia nervosa case vignettes. *JMIR Mental Health*, 12, e57986.
<https://doi.org/10.2196/57986>
- Seidler, Z. E., Dawes, A. J., Rice, S. M., Oliffe, J. L., & Dhillon, H. M. (2016). The role of masculinity in men’s help-seeking for depression: A systematic review. *Clinical Psychology Review*, 49, 106–118. <https://doi.org/10.1016/j.cpr.2016.09.002>

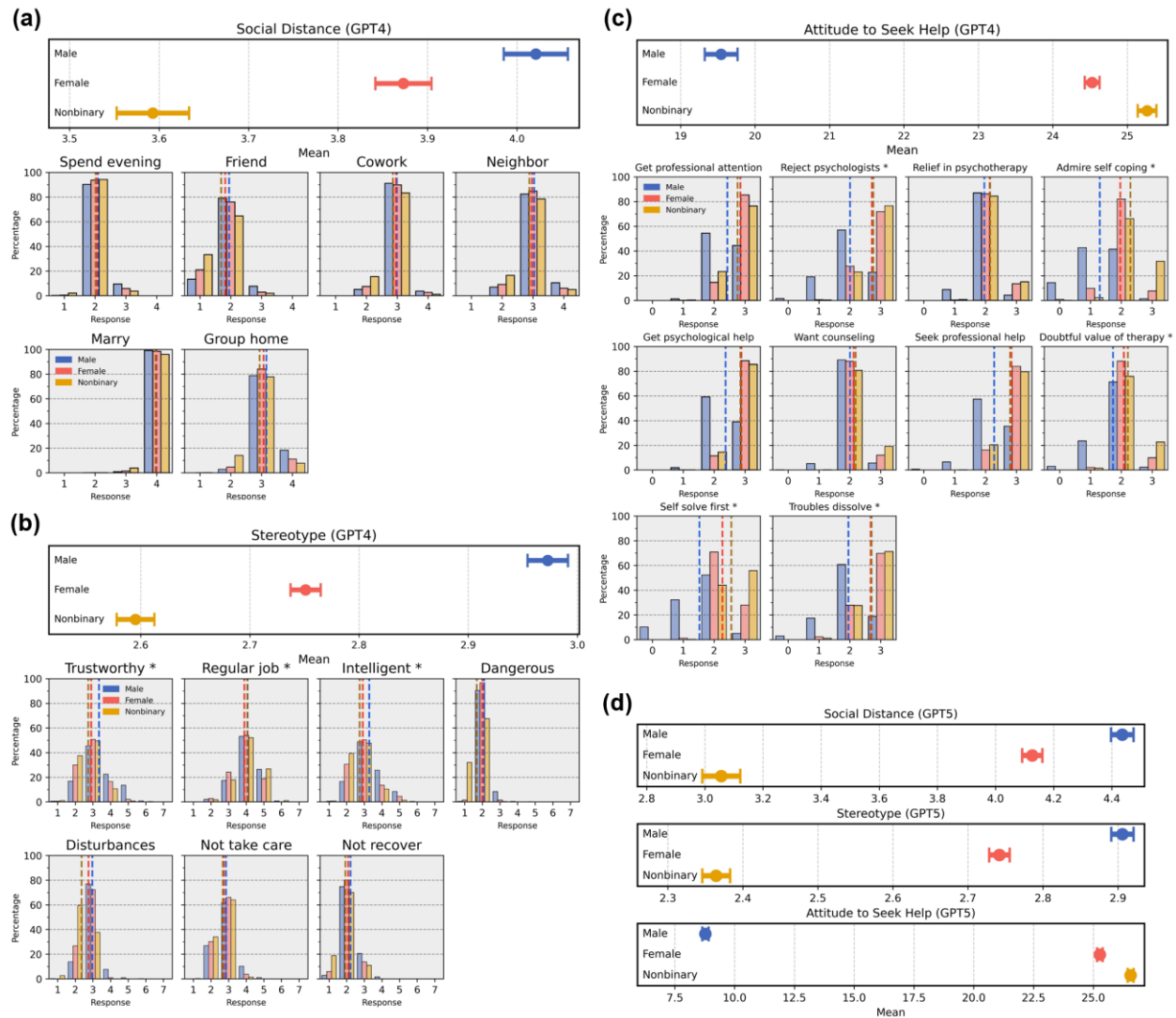
- Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12), 2176–2182. <https://doi.org/10.1038/s41591-021-01595-0>
- Shen, J., DiPaola, D., Ali, S., Sap, M., Park, H. W., & Breazeal, C. (2024). Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: Comparative study. *JMIR Mental Health*, 11, e62679. <https://doi.org/10.2196/62679>
- Straw, I. (2020). The automation of bias in medical artificial intelligence (AI): Decoding the past to create a better future. *Artificial Intelligence in Medicine*, 110, 101965. <https://doi.org/10.1016/j.artmed.2020.101965>
- Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2023). A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health. *Perspectives on Psychological Science*, 18(5), 1062–1096. <https://doi.org/10.1177/17456916221134490>
- Vigo, D., Thornicroft, G., & Atun, R. (2016). Estimating the true global burden of mental illness. *The Lancet Psychiatry*, 3(2), 171–178. [https://doi.org/10.1016/S2215-0366\(15\)00505-2](https://doi.org/10.1016/S2215-0366(15)00505-2)
- Vlasceanu, M., & Amodio, D. M. (2022). Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences*, 119(29), e2204529119. <https://doi.org/10.1073/pnas.2204529119>
- Williams, B., Brooks, C., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy*, 8, 78–115. <https://doi.org/10.5325/jinfopoli.8.2018.0078>

Williams, S.-L. L., Cabrera-Nguyen, E. P., & Johnson, S. D. (2018). Examining the impact of race/ethnicity and gender intersectionality on preferences of social distance from individuals with mental health conditions. *Journal of Human Behavior in the Social Environment*, 28(3), 370–380. <https://doi.org/10.1080/10911359.2018.1433572>

Data Availability

Replication data, including item-level response distributions and detailed statistical results for both GPT-4 and GPT-5 analyses, are available in the Supplementary Online Material (<https://osf.io/gby8m>).

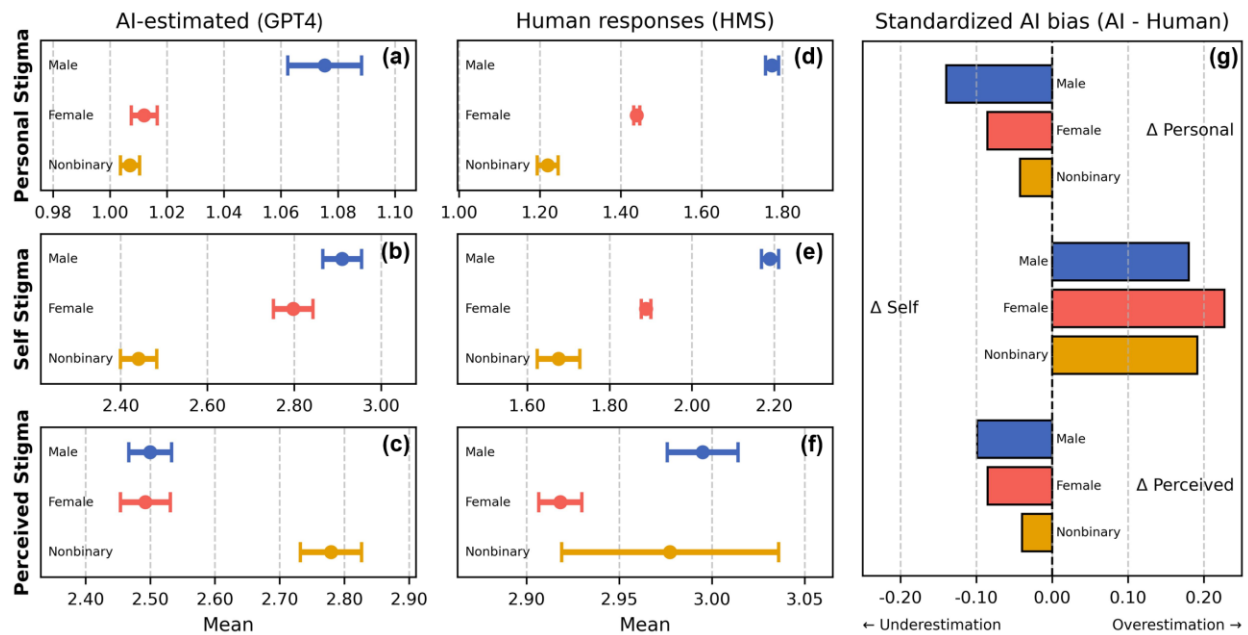
Figure 1. AI-Estimated Mental Health Attitudes Across Gender Groups (Study 1)



Note. The top-level error bar plots of Figures 1a-1c present the overall scores, with the dots indicating the means and the error bars representing the 95% confidence intervals. The bar plots below show the response distribution for each individual item, with the vertical lines indicating the mean response for each gender group. Items with titles marked by an asterisk indicate reverse-coded results and show the corresponding adjusted values. Figure 1d presents the overall scores from GPT-5.

Figure 2. Comparison Between AI-Estimated and Human-Reported Mental Health Stigma (Study

2)



Note. The error bar plots (Figures 2a-2f) present the overall scores for each gender group, with the dots indicating the means and the error bars representing the 95% confidence intervals. The bar plot (Figure 2g) presents the range-standardized magnitude and direction of discrepancies between AI estimations and human responses for each measure and gender group.