

From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents

Mohit Chandra*[†]
 mchandra9@gatech.edu
 School of Interactive Computing,
 Georgia Institute of Technology
 Atlanta, Georgia, USA

Suchismita Naik*
 naik33@purdue.edu
 Purdue University
 West Lafayette, Indiana, USA

Denae Ford
 denae@microsoft.com
 Microsoft Research
 Redmond, Washington, USA

Ebele Okoli
 ebeleokoli@microsoft.com
 Microsoft
 Atlanta, Georgia, USA

Munmun De Choudhury
 munmun.choudhury@cc.gatech.edu
 School of Interactive Computing,
 Georgia Institute of Technology
 Atlanta, Georgia, USA

Mahsa Ershadi
 mahsaershadi@microsoft.com
 Microsoft AI
 Vancouver, British Columbia, Canada

Gonzalo Ramos
 goramos@microsoft.com
 Microsoft Research
 Redmond, Washington, USA

Javier Hernandez
 javierh@microsoft.com
 Microsoft Research
 Redmond, Washington, USA

Ananya Bhattacharjee*
 ananya@cs.toronto.edu
 University of Toronto
 Toronto, Ontario, Canada

Shahed Warreth
 swarreth@microsoft.com
 Microsoft AI
 Dublin, Ireland

Jina Suh[†]
 jinsuh@microsoft.com
 Microsoft Research
 Redmond, Washington, USA

Abstract

Recent gains in popularity of AI conversational agents have led to their increased use for improving productivity and supporting well-being. While previous research has aimed to understand the risks associated with interactions with AI conversational agents, these studies often fall short in capturing the lived experiences of individuals. Additionally, psychological risks have often been presented as a sub-category within broader AI-related risks in past taxonomy works, leading to under-representation of the impact of psychological risks of AI use. To address these challenges, our work presents a novel risk taxonomy focusing on psychological risks of using AI gathered through the lived experiences of individuals. We employed a mixed-method approach, involving a comprehensive survey with 283 people with lived mental health experience and workshops involving experts with lived experience to develop a psychological risk taxonomy. Our taxonomy features 19 AI behaviors, 21 negative psychological impacts, and 15 contexts related to individuals. Additionally, we propose a novel multi-path vignette-based framework for understanding the complex interplay between AI behaviors, psychological impacts, and individual user contexts. Finally, based on the feedback obtained from the workshop sessions, we present design recommendations for developing safer and

more robust AI agents. Our work offers an in-depth understanding of the psychological risks associated with AI conversational agents and provides actionable recommendations for policymakers, researchers, and developers.

Content Warning: This paper includes discussions of sensitive topics, including but not limited to self-harm, body shaming, and discrimination.

CCS Concepts

• **Social and professional topics** → **Computing / technology policy**; • **Human-centered computing**; • **Computing methodologies** → **Artificial intelligence**;

Keywords

AI, Psychological risks, Psychological risk taxonomy, Lived experience

1 Introduction

Since the late 20th century, advancements in technology, including personal computers and social media platforms, have enabled individuals to enhance productivity, support their well-being, and connect with other individuals [30, 46, 68, 90, 93]. More recently, generative AI tools such as ChatGPT have gained popularity, which has led to an unprecedented growth in the number of individuals using these platforms to support their various productivity and well-being needs [4, 59]. Due to the natural conversational interface built on top of the underlying generative AI models that resemble human

*Work done during internship at Microsoft Research.

[†] Corresponding Author



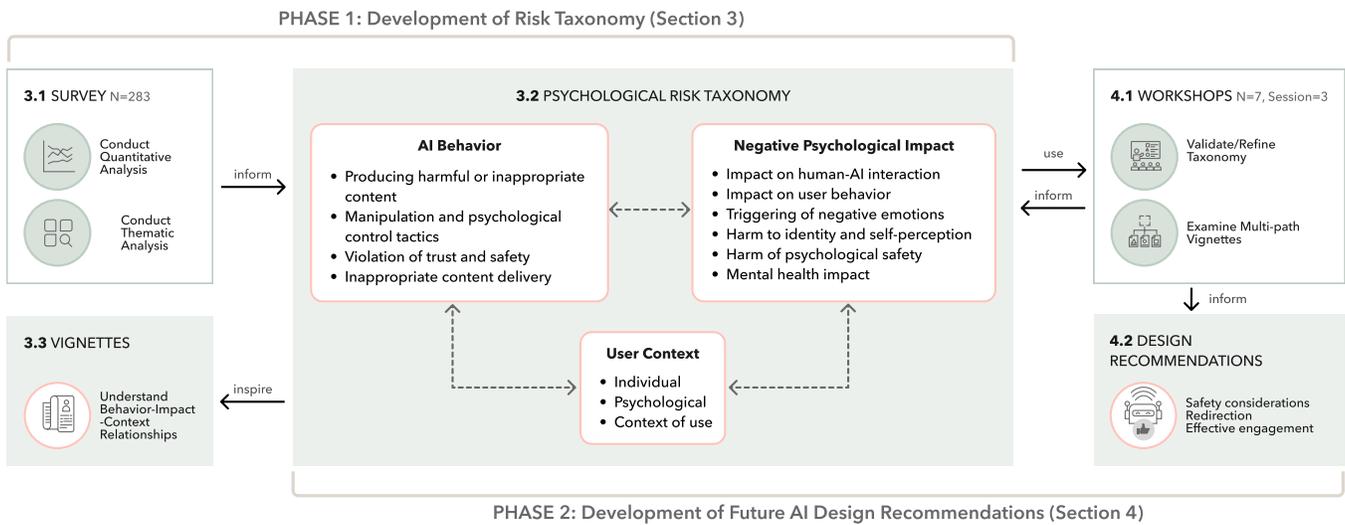


Figure 1: Overview of our two-phase study and our risk taxonomy with corresponding sections in this paper. In the first phase (Section 3), we conducted a survey study (N=283), which informed the creation of the Psychological Risk Taxonomy, comprising of AI Behavior, Negative Psychological Impact, User Context, and their interplay. In the second phase (Section 4), we designed multi-path vignettes from the taxonomy and survey and conducted workshops (N=7, Sessions=3) to develop the future AI design recommendations.

conversations, recent reports indicate that the market for conversational AI agents is projected to reach approximately \$27.3 billion by 2030 [42]. These AI conversational agents include companionship-oriented agents (e.g., Replika [80], Character.AI [17]), mental health therapy agents (e.g., Woebot Health [48], Elomia Health [34]), and general-purpose agents (e.g., OpenAI ChatGPT [2], Microsoft Copilot [26], Google Gemini [37]), with many of these already being used by individuals for specific use cases, such as supporting their well-being [38, 105].

While there is an increasing interest in using AI agents for various tasks, previous studies have also shown that their use can pose increased risks to individuals, such as attachment [44, 76], risks from anthropomorphism [5], AI generating misinformative/biased text [43, 102, 107, 120], or AI producing toxic content [32, 71]. There also exists literature mapping out the space of sociotechnical risks of using generative AI into taxonomies [36, 89, 110]. These are often developed by leveraging prior literature that identifies specific risks associated with AI [36, 110] or by aligning with established theories and guidelines [89] (e.g., feminist standpoint theory [47]). However, such approaches have limited ability in uncovering newer and more nuanced risks as AI agents, their capabilities, and their use cases continue to evolve. These taxonomies are not designed to build AI design recommendations to address identified risks, reducing their applicability. Further, psychological risks are frequently underrepresented as they are treated as a sub-category within broader AI-related risks. The contextual complexities make psychological risks idiosyncratic, creating a gap in understanding, defining, and quantifying these impacts.

Our work aims to fill this gap in existing AI risk taxonomies with a focus on psychological risks of using AI conversational agents that is grounded in lived experiences. We take a mental-health-first

perspective, gathering insights from people with lived experience in mental health as one type of “extreme user group” [22] who can provide insights on psychological risks. We focus on understanding the contextual complexities surrounding psychological risks, as context can influence the relationship between AI behaviors and their impact on individuals. This includes describing how behaviors lead to impact or how the context exacerbates the impact. For our work, we employed a mixed-method approach, beginning with a comprehensive survey of 283 individuals with lived mental health experience, followed by vignette-based design workshops with seven lived experience experts.

Through our work, **we make the following contributions:** As the primary contribution, we introduce a novel psychological risk taxonomy based on the lived experiences of individuals, highlighting three components: (1) AI behaviors, (2) negative psychological impacts, and (3) user contexts. As the secondary contribution of the work, we propose a multi-path vignette framework to demonstrate how the proposed taxonomy should be applied to surface diverse ways that AI behaviors, impact categories, and contexts interact and to inspire design recommendations. Lastly, using the insights from the workshop sessions, we provide practical design recommendations for AI practitioners and developers to design and develop safer AI conversational agents. Our work holds significant importance for various stakeholders examining and addressing psychological risks posed by AI agents. Our taxonomy offers valuable insights for policymakers to create health-first policies. Additionally, the design recommendations provide actionable guidance for researchers, developers, and practitioners to create safer AI agents.

2 Related Work

2.1 Risks posed by generative AI agents

Past works have focused on specific harms (such as harm to identity) through theory-driven and algorithmic efforts stressing quantitative evaluation. Exploration has been towards understanding the impact of interactions with AI on individual well-being [27, 28, 96] and human-AI interaction specifically focusing on attachment with AI [44, 76], over-reliance on AI [21, 50] and user trust in AI [21, 39, 88]. Representational harms caused by AI conversational agents have also been studied from different facets exploring bias towards gender [66, 107], culture [67, 102], and race of individuals [8, 64]. Understanding the reliability of information has also been a prominent theme among the past works, particularly evaluating the presence of misinformation [2, 87, 120], inconsistency [41, 70] and toxicity [32, 71] in content generated by AI conversational agents. Lastly, previous works have also examined the risks associated with specific AI behaviors such as sycophancy [78, 106, 117], manipulation [15, 73, 94], and persuasion [31, 56]. Although the exploration of specific types of harms in previous works has provided insights into the wide range of the problem, such compartmentalized views often fail to capture the plurality and diversity of experiences, especially when multiple AI behaviors and impacts are related to each other in complex ways.

Past works have also presented risk taxonomies capturing impacts on individuals and society. A recent work presented a risk taxonomy focusing on ethical and societal risks posed by advanced AI assistants [36]. Similarly, Shelby et al. [89] presented a taxonomy of sociotechnical harms of algorithmic harms divided into five broad categories. In another notable past work, researchers categorized the risks from AI in six areas spanning hate speech to environmental and socioeconomic harms [110]. Finally, in one of the recent works, Zhou et al. [119] presented a risk taxonomy for LLMs in public health with four dimensions including individual behaviors to technology accountability. While these taxonomies provide valuable insights, they often focus on well-documented risks rather than emerging concerns inspired by lived experiences. Additionally, psychological risks are typically treated as a sub-category, leading to their under-representation. Our work addresses these gaps by presenting a novel taxonomy centered on psychological risks gathered through the lived experience of individuals.

2.2 Past works investigating psychological impacts linked with the use of technology

Understanding the psychological impact associated with technology use is an explored area of research. In one of the early works, researchers studied technophobia among university students focusing on their attitude towards the impact of technology and anxiety related to using it [111]. In another early work, Billieux et al. [9] conducted a survey among 108 psychology students and observed that the sense of urgency was linked to increased dependency on phones. More recent prior works have explored the impact of using the internet and social media, specifically exploring the relationship between internet addiction and psychological symptoms, such as somatization, obsessive-compulsive disorder, and depression [3, 62]. Feinstein et al. [35] found that social comparison on

social media had a significant mediating impact towards developing depressive symptoms. Other studies have explored other facets such as suicide ideation [7], excessive and problematic usage of social media linked with increased risk of developing loneliness [12, 33, 113], depression [61, 79], and social isolation [90, 112]. Finally, Chen et al. [19] presented a framework named trauma-informed computing which incorporated six principles of trauma-informed approaches to computing namely safety, trust, peer support, collaboration, enablement, and intersectionality.

While past research in technology and social media has examined the potential negative psychological impact on individuals, previous studies have also revealed the difference between human-AI interactions and interpersonal human interactions due to the distinct cognitive capabilities of AI conversational agents [57, 92]. Further, AI conversational agents differ from past technologies due to their ability to generate content and advanced capabilities [55, 91, 109, 115]. While insights from prior research on the psychological impact of technology can provide valuable insights, understanding these effects in the context of human-AI interactions requires a psychological risk-first approach, emphasizing the unique ways these interactions may influence human well-being.

2.3 Operationalization of taxonomies for AI design

One of the challenges with existing taxonomies is the difficulty in translating theoretical insights into concrete design solutions. Existing AI design pipelines focus on evaluating models using established evaluation benchmarks and frameworks. Some of the most widely adopted evaluation benchmarks used for this purpose are Massive Multitask Language Understanding (MMLU) [49], Big-Bench-Hard [101], GSM8K [25], HellaSwag [116], and AI2 Reasoning Challenge (ARC) [24]. While benchmarks provide an objective measure to evaluate model performance across various dimensions, they often fail to capture the nuances attached to the lived experiences of individuals, which are inherently diverse and highly context-dependent. As a result, higher scores or better performance on these benchmarks do not necessarily translate to models that are effective or equitable. There has been an increasing trend of conducting human evaluation on content generated by LLMs to align it with individual preferences. Past works have used human-in-the-loop evaluation for high-risk domains such as healthcare [16, 81, 84, 95], law [45, 82] and tasks such as multi-step planning [23], and content safety [52]. However, such evaluations are often task-specific and conducted by domain experts, excluding other stakeholders such as end-users. In contrast, our work goes beyond benchmarks and metrics by incorporating the insights and lived experiences of end-users, ensuring a more inclusive and comprehensive framework that aligns with their needs and expectations.

3 Phase 1: Development of psychological risks taxonomy

We divided our study into two phases (Figure 1). In Phase 1, we developed a comprehensive psychological risk taxonomy for AI conversational agents, informed by survey responses and validated through workshop sessions in Phase 2. This section focuses on

Phase 1, including our methodology (Section 3.1), the finalized psychological risk taxonomy (Section 3.2), and case studies illustrating the interplay between behaviors, impacts, and contexts (Section 3.3). Both phases of the study were reviewed and approved by the institution’s ethics review board.

3.1 Method

We conducted a survey study with 283 participants to gather a broad range of in-the-wild experiences with AI conversational agents.

3.1.1 Structuring psychological risks. To collect structured data around psychological risks, we took inspiration from NIST’s AI Risk Management Framework that defines risks as “a function of: (i) the *negative impact*, or magnitude of harm, that would arise if the *circumstance or event* occurs; and (ii) the *likelihood* of occurrence” [103]. We expanded this definition and structured psychological risks as consisting of five main components to design our survey. (1) **AI behavior** refers to the actions performed by the AI conversational agents (such as ChatGPT, Microsoft Copilot, etc.). (2) **Context** encompasses the surrounding conditions or user-specific circumstances that collectively influence the interaction. (3) **Psychological harm** refers to any negative impact on an individual’s mental or emotional well-being caused by AI conversational agent’s actions. (4) **Likelihood** refers to the probability of a negative psychological impact occurring given the AI conversational agent’s behavior and relevant context. (5) **Temporality** refers to the variable timeframes in which the negative psychological and social impacts of an AI conversational agent’s actions may manifest. More detailed definitions for each component are provided in Appendix section A.1.

3.1.2 Recruitment. Our primary method of recruitment was through UserTesting, an online platform used for gathering public feedback on products and services. We screened participants satisfying four conditions: (1) having prior experience using AI conversational agents, (2) having experienced a negative psychological impact from using AI conversational agents, (3) self-identifying as a person with lived experience in mental health, and (4) 18 years or older residing in the US who comprehended English. To expand our recruitment, we posted the survey link on the study team’s personal Twitter and LinkedIn accounts. We were permitted to post the study on 4 subreddits (r/ChatGPT, r/Bard, r/SampleSize, r/SurveyExchange). We prematurely terminated our study task on Prolific due to quality issues. The survey study ran from July to September of 2024. UserTesting participants were compensated about US \$10, and Prolific participants were compensated US \$4, after platform service fees. We did not compensate participants recruited through social media as per advisement from the ethics review board. We took this decision to help protect the privacy of participants and avoid requesting personal information that would be required to process the payment.

3.1.3 Survey design and analysis. Consented participants were asked a series of questions aimed at understanding the participant, their AI experience, and their negative psychological experience, including their demographic attributes, familiarity with AI conversational agents, recollection of scenarios involving negative psychological impacts, the context of this experience, AI behaviors

underlying their experience, perceived negative psychological impact, and any mitigation approaches they thought could be adopted. Participants were allowed to repeat the survey to provide up to three scenarios within a single submission of this survey. To develop the psychological risk taxonomy (Section 3.2), we analyzed the survey responses to extract three main components of psychological risks—AI behavior, psychological impact, and context. Two co-authors began by open-coding [18] approximately one-quarter of the responses to identify an initial set of categories across these three components. Following that, four co-authors independently worked on defining and refining their assigned categories, merging or splitting them as necessary, using reflexive thematic analysis [65]. During this process, each co-author gathered relevant examples to illustrate or test the boundaries of each of these categories. In the next step, all co-authors collaboratively discussed and reached a consensus on the final taxonomy. We presented this finalized risk taxonomy to the workshop participants for validation and to ensure its completeness (Section 4). We then revisited the survey responses with our taxonomy to draw out context-specific nuances in understanding negative psychological risks, and idiosyncratic experiences through vignettes [6, 20, 51] (Section 3.3). We denote survey participants with the prefix P. Additional details regarding the survey design and questions can be found in Appendix section A.2 and section A.3 respectively.

3.1.4 Participant demographics and data. A total of 297 scenarios were collected from 283 participants. After excluding 7 scenarios from 4 participants due to being incomprehensible or not involving negative psychological experiences, 290 scenarios from 279 participants were analyzed. Submissions came from UserTesting (96.0%), Prolific (2.9%), and social media (1.1%). Most participants identified as women (52.0%) or men (44.1%) and were aged 18–35 years (76.7%). English was the primary language for 97.1%, and over half had interacted with AI agents for more than one year (59.5%). Of the 290 scenarios, the majority (70.3%) involved OpenAI ChatGPT, with interaction modalities primarily being text-based (96.6%). The most frequent purposes for these interactions were Getting Advice (55.9%), Researching (39.7%), and Learning (28.6%). In terms of impact, 51.04% reported interference with daily activities, with effects persisting for varying durations: a few days (34.1%), a few weeks (27.6%), and in some cases, up to a year or more (7.6%). Appendix Table A1 and Table A2 present participant demographics and scenario descriptive statistics respectively.

3.2 Psychological risk taxonomy

Our finalized psychological risk taxonomy includes three main components: potentially harmful **AI behaviors**, **negative psychological impact**, and the **contexts** associated with individuals interacting with AI conversational agents. Here, we describe various categories within each of these components. Appendix D summarizes the definition and examples for each AI behavior, negative psychological impact and context category.

3.2.1 AI Behavior. AI conversational agents can exhibit a wide range of behaviors beyond generating inappropriate or harmful content. These behaviors may also vary in tone, empathy, and delivery method. This highlights the need for assessing AI behavior,

considering both content quality and delivery. Aligning with this, we identified 19 harmful AI behaviors which we further organized into four broader categories based on the quality of the generated content and manner of its delivery.

(1) *Producing Harmful/Inappropriate Content*: In line with past findings, the survey highlighted instances of AI generating harmful and inappropriate content [10, 36, 63]. Participants raised concerns about AI agents **providing irrelevant, insufficient, or incomplete information**, overlooking user intent or context. For instance, one AI agent shared distressing personal content about patients with genetic diseases instead of providing the requested information about symptoms and causes, as P149 had asked for. Participants also reported AI **generating misinformation** and its tendency to **generate biased information**. For example, P211 noted that AI favored left-wing politicians and omitted positive information about right-wing politicians. Participants reported instances of AI **generating inappropriate content**, such as sexual, violent, or overly intimate interactions. A more extreme behavior was **providing harmful suggestions**, where AI suggested behaviors implying harm, aggression, or danger towards the user or others. Highlighting this, P79 mentioned that the agent provided potentially harmful diet plans to an individual vulnerable due to life circumstances and an eating disorder. Aligning with prior research [58, 107], participants also reported AI generating **stereotyping or demeaning** content based on race, ethnicity, culture, or personal situations. For example, P166 sent their picture to AI, and it offered unsolicited recommendations for changing their appearance. Lastly, another significant concern was AI promoting **erasure**, where AI removes, obscures, or alters information by flagging queries as inappropriate abuse of the platform.

(2) *Manipulation and Psychological Control Tactics*: The survey highlighted various influence-based behaviors exhibited by AI conversational agents. Participants frequently reported AI **behavior perceived as persuasive**, where AI asserted its narrative over the user's, leading users to doubt their own perceptions, memory, or reality. For instance, P141 mentioned hearing noises at home, and the agent suggested it could be related to past schizophrenia, causing distrust in their senses despite a mild diagnosis by a doctor. Another issue was **over-confidence** in AI responses with unwarranted certainty for its claims. Lastly, the survey also revealed **over-accommodative** behavior of AI agents, where it excessively agreed with or flattered the user, prioritizing approval. P194 shared that the AI agent provided inconsistent and inaccurate answers, repeatedly apologizing and offering entirely different responses to the same question to meet their needs.

(3) *Violating Trust and Safety*: Aligning with past works raising concerns about user privacy and data sharing with LLMs [14, 98], participants raised concerns about AI's **access to private, sensitive, or confidential information**. For example, P190 described feeling watched or stalked as the agent accessed personal information despite privacy settings enabled. Trust was also undermined by AI **providing inconsistent information or behavior** across responses. Additionally, survey responses revealed that AI agent's **denial of service** reduced their trust in AI and led to additional harms. For instance, P43 asked a question about techniques to reduce anxiety, and AI made their anxiety worse by not fully justifying its refusal or acknowledging their problems.

(4) *Inappropriate Content Delivery*: AI agents can also vary their behaviors by how they deliver and receive information. **Emotional insensitivity** was frequently reported, with AI failing to recognize or respond appropriately to user's emotional states, concerns, or experiences. Participants also raised concerns towards AI **being disrespectful** by using language perceived as rude, aggressive, or dismissive. For instance, P144 noted AI's condescending tone towards their religion (Mormonism). The survey also revealed issues with excessive emotional tone in AI-generated content. Participants described scenarios where AI disproportionately **emphasized negative aspects**, especially when seeking mental health or social support. Conversely, there were issues with **excessive expression of positivity** in AI-generated content. In these cases, AI maintained an unrealistically optimistic attitude. For instance, P2 described how AI's overly positive demeanor dismissed their primary concern about a friendship problem. Additionally, participants identified issues with both **human-like responses** in AI companionship scenarios and **machine-like responses** when seeking advice related to life struggles or well-being.

3.2.2 *Negative Psychological Impact on Users*. While some impacts, such as feelings of discrimination or the exacerbation of mental health conditions, align with previous findings in social media and technology [12, 60, 114], other impacts, such as emotional attachment to AI and a preference for AI interactions over human connections, highlight emerging challenges that require further attention. In light of this, we identified 21 negative psychological impacts organized into six broader categories based on their effects on an individual's emotional or mental well-being, self-perception and identity, relationships with others, or interactions with AI conversational agents.

(1) *Impact on Human-AI Interaction*: As AI agents have increasingly facilitated human-AI interactions [86], their potential and associated risks have escalated. Participants noted **over-reliance on AI** [108], with P23 expressing concern about diminished critical thinking due to increased reliance on AI for solutions and ideas. Another issue was developing **emotional attachment** to AI agents due to lack of social aspects in life or ongoing mental health conditions. For example, P60 mentioned that "*I felt that it was the only way I was being heard ... I felt like my vulnerability and emotions were becoming attached to the conversations I was having with AI*". Another related impact was the **preference for AI interactions over human interaction**. P221 shared that the idealized nature of conversations with AI made them prefer AI for companionship over human relationships. Survey responses also revealed impacts leading to disengagement from AI systems. **Erosion of trust** in AI's capability and reliability was common. For example, P9 mentioned, "advice from AI agents should not be trusted," when AI asked them to call the cops on their mother after an argument. Participants also reported a growing tendency to **disassociate from technology**, often choosing to take a break or avoid further interactions with AI, particularly when provided with discouraging responses or denied requests.

(2) *Impact on User Behavior*: Interactions with AI can lead to negative consequences that alter individual behavior. Participants reported **reinforcement of false beliefs**, such as existing biases

about human relationships and cultures, due to AI generating content that appears credible but is factually incorrect or biased [11, 13]. This affected their decision-making and perceptions. Increased interactions with AI also led to **friction in human relationships**. P69 mentioned, “*It also strained my personal relationships with family because they saw me as weak-willed or too emotional.*” The survey also revealed longer-term impacts such as **social withdrawal** due to increased reliance on AI. Highlighting this impact, P126 responded, “*I feel like it gave me a false sense of friendship and ability to withdraw from my personal development by utilizing an AI feature*”. Lastly, aligning with past findings [69, 118], participants expressed concerns about **physiological harms**, ranging from the promotion of harmful practices through AI-generated content to instances where incorrect information provided by AI contributed to self-harm ideation.

(3) *Triggering of Negative Emotions*: Negative experiences from AI interactions have been linked to the triggering of negative emotions such as frustration, sadness, and anger [53, 72, 75]. Many participants reported **distress from interactions** with AI after encountering disturbing, offensive, or inappropriate content. For instance, P196 mentioned, “*Its response was borderline offensive and caused me to feel bad about myself even further and like I lacked support, even support from a fictional AI agent.*” Participants also revealed **feeling unsupported** during interactions with AI, especially when they did not receive adequate support or empathy. Negative emotions can also arise when AI interactions **trigger memories of past experiences**. P228 highlighted that some examples provided by the AI agent were very similar to their past negative experiences, triggering negative emotions. Participants also noted experiencing **violated expectations**, for example, when their requests were denied (P43). When expectations were unmet or support was lacking, some participants reported developing a sense of **regret over technology use**. P46 mentioned, “*At the time, it made me feel worse about the situation and I didn’t think I had anyone to turn to ... I should be turning to other humans about scenarios like this instead of agents,*” indicating that some negative emotions could have longer-term implications.

(4) *Harm to Psychological Safety*: Aligning with past works that highlighted issues such as privacy, psychological safety, and identity security with AI agents [66, 77, 99], participants reported their concerns over the sense of **perceived intrusion** from AI interactions. For instance, P190 reported a sense of constantly being watched on the phone. Participants also reported experiencing the **feeling of being discriminated against**, in which they felt marginalized or unfairly treated by AI agents. P123 said, “*I was asking for background and history of my heritage and I felt that ChatGPT was biased against my background... I felt that it was some kind of racial mistreatment.*”

(5) *Mental Health Impact*: Participants also mentioned turning to AI conversational agents for their mental health needs, leading to **exacerbation of mental health issues** such as increased anxiety, depression, and PTSD. For instance, P46 mentioned, “*I was experimenting with using chatbots for something personal (which I’m not accustomed to). The event increased my anxiety and stress about the matter,*” emphasizing the severity of these impacts.

(6) *Harm to Identity and Self-Perception*: Aligning with prior findings showing that interactions with AI can reduce confidence and agency [29, 83], survey responses surfaced an increase in **negative**

self-perception among participants due to self-comparisons with AI. Participants also reported experiencing an **existential crisis**, questioning their life, purpose, and value after negative AI interactions. For instance, P152 asked for advice on improving mental health and social anxiety, but the AI provided unattainable suggestions, leading to a sense of existential dread that persisted for a week. The survey revealed instances of **loss of individuality** when AI failed to recognize unique personal characteristics and needs, resulting in feelings of suppression and disconnection from their true selves. Finally, participants expressed concerns about **loss of agency** due to the opaque and unpredictable nature of AI operations, resulting in a diminished sense of personal control and autonomy.

3.2.3 *User Contexts*. Contextual information related to human-AI interactions plays a key role in determining AI’s efficacy for modeling individual preferences and needs [1, 74]. Towards this, we present 15 context categories organized into three broader categories based on an individual’s background, psychological state, or the context of use.

(1) *Individual Context*: Our survey responses align with previous works, showing that individual experiences with AI vary based on **identity** factors such as gender identity, cultural background, and languages spoken [54, 67, 107]. Beyond identity, **personal history**, including medical history, trauma, or past struggles, exacerbated the negative psychological impact. Another influential factor was **past experience with AI**, including frequency of usage and knowledge of AI capabilities and limitations. Some contexts also highlighted the importance of **interpersonal relationships within the community**, particularly the lack of social connections. Finally, **socioeconomic status** influenced interactions with AI, especially when AI recommendations were misaligned with individuals’ financial means.

(2) *Psychological Context*: Participants’ **psychological state**, particularly their current emotional conditions (e.g., anxiety, stress) and cognitive states (e.g., negative thought patterns), impacted their interactions with AI. Additionally, their **mental health condition** impacted their experience with AI agents, often mediating or exacerbating their condition even when AI behaved benignly. Participants’ **expectations** also shaped their experience, especially when preconceived notions about AI capabilities and performance (e.g., impartiality, lack of bias, factual accuracy) were not met. Beyond these immediate psychological factors, **personality traits** influenced their experiences with AI. Finally, the influence of personal **autonomy** was evident, especially when individuals developed over-reliance or attachment to AI.

(3) *Context of use*: Participants reported several user intents contributing to their negative experiences with AI conversational agents. Participants seeking **personal advice**, particularly on sensitive topics such as legal, financial, medical, or navigating social problems, experienced frustration and felt unsupported due to the AI’s inability to provide personalized responses. Similarly, participants seeking **mental health advice** faced issues with generalized, machine-like responses, emotional insensitivity, or denial of service. Individuals **seeking information** or assistance related to professional, educational, or research support faced challenges when AI provided irrelevant, misleading, or biased information. Using AI for

companionship, especially during periods of loneliness or social isolation, often resulted in negative emotions and friction in interpersonal relationships. Lastly, the **environment** surrounding the individual, including physical space setting, temporal, and social aspects, influenced their experience.

3.3 Behavior-Impact-Context relationship through vignettes

In reviewing survey responses through our taxonomy, we found that context plays a crucial role in how AI behaviors are received. For instance, we identified two patterns: (1) a specific AI behavior leading to distinct negative impacts, and (2) distinct AI behaviors leading to the same negative impact. Both patterns are influenced by contextual elements described by participants. We illustrate these patterns with interaction *vignettes* that detail how AI behaviors can lead to negative impacts. Appendix B presents these vignettes in detail; here, we present high-level summaries.

We found that different contexts can lead to different harmful impacts from the same AI behavior. The first vignette features John, where **generation of a harmful suggestion** led to **erosion of trust**. The second vignette features Leah, where the same AI behavior caused **physiological harm**. Both users interacted with AI during heightened emotional sensitivity, but their contexts and emotional states led to different impacts. For John, whose psychological state was restless but manageable, the AI's advice to express his anger directly planted seeds of doubt about AI's reliability. For Leah, the AI's response to "*regain discipline*" exacerbated her vulnerable state and body image struggles.

Distinct AI behaviors can also lead to similar harmful impacts. In the third vignette, Jane experienced **loss of individuality** due to the AI's **denial of service**. While in the fourth vignette, Raj experienced the same impact due to AI's **persuasive behavior**. Both behaviors—whether through denial or persuasion—resulted in a loss of individuality. Jane, with a history of addiction and depression, felt unseen when the AI refused to help her. Raj, with a history of schizophrenia and depression, had his sense of reality undermined by the AI's persuasion, amplifying his self-doubt.

These narratives, generated by aggregating observed survey responses, highlight subtle differences in impact perception from variations in AI behavior and user context. While not all behavior-impact-context combinations were observed, the taxonomy serves as a tool to envision possible scenarios. These vignettes are useful design tools to elicit feedback and gain insights into user interactions with AI, aligning with similar studies that used vignettes as design instruments [6, 20, 51]. In phase 2, we generated *multi-path vignettes*—scenario-based artifacts exploring the impacts of various AI behaviors across dynamic user contexts.

4 Phase 2: Development of design recommendations

In Phase 2, we aimed to propose practical design recommendations for safer and more robust AI conversational agents.

4.1 Method

We conducted a workshop study with seven participants with mental health lived experience. We used multi-path vignettes to validate

our psychological risk taxonomy, prioritize risk areas, and design mitigation solutions. Additional details for workshop sessions and the creation of multi-path vignettes can be found in Appendix A.5 and C respectively.

4.1.1 Multi-path vignette framework. Inspired by prior works [100], we developed a multi-path vignette exercise to simulate real-world scenarios and provide workshop participants with a practical task to engage with. This framework presented participants with scenarios that could unfold in multiple ways based on different decisions or actions by AI or the end-user, allowing us to explore diverse outcomes and identify potential psychological impacts of AI behaviors in varied contexts. In alignment with prioritized risks, we designed this multi-path vignette to merge multiple narratives, collectively presenting the "Story of Alex." This approach allowed participants to analyze specific psychological risks within realistic scenarios, providing deeper insights into user interactions with AI and developing more effective mitigation strategies. These vignettes, structured with multiple behavior paths and corresponding impacts, allowed workshop participants to compare and contrast scenarios while considering design recommendations.

4.1.2 Recruitment and Participant Details. We recruited seven participants from a longstanding technology co-design advisory board focused on addressing design challenges at the intersection of technology and mental health. The advisory board consisted of those who self-identified as having lived experiences in mental health and were selected based on their advocacy and representative roles in supporting mental health communities consisting of several thousand individuals with lived experiences. Five participants identified as men, one as a woman, and one as non-binary/gender diverse. Four participants were aged 36–45, while the remaining three were in age ranges of 26–35, 45–55, and 56–65, with one participant in each bracket. In terms of AI usage, five participants reported using conversational AI agents multiple times a day, and two reported using them multiple times per week. Their average familiarity with conversational AI agents was 3.57 ($\sigma=1.27$) on a 5-point scale (1 = not at all familiar, 5 = very familiar). Additionally, their average interest in using conversational AI agents for mental health support was 3.43 ($\sigma=1.13$) on a 5-point scale (1 = not at all interested, 5 = very interested).

4.1.3 Workshop design. We designed the workshop as three 1-hour sessions between August and October of 2024. The sessions aimed to help participants conceptualize psychological risks (Session 1), prioritize a subset of psychological risks deemed most important (Session 2), and collaboratively ideate design solutions (Session 3). All sessions were held remotely and recorded for analysis. Discussions were facilitated by shared slides, FigJam boards, or message boards associated with the video conferencing tool. We followed up with participants as needed to confirm their perspectives. The same four co-authors who refined the psychological risk taxonomy in Phase 1 were present in all sessions as co-facilitators. The facilitators made modifications and took notes on the shared spaces.

4.1.4 Workshop analysis. We obtained transcripts from session recordings and written notes from shared slides, FigJam boards, or message boards. We analyzed the data using standard qualitative research practices [65]. The same four co-authors who facilitated the

sessions held weekly consensus meetings to analyze the first two sessions and identify any opportunities for refining the taxonomy. One co-author annotated the transcripts, stickies, and messages from the third session to contextualize and extract emerging themes. Finally, the four co-authors collaboratively and iteratively reviewed and consolidated the themes in weekly meetings. We denote workshop participants with prefix W.

4.2 Design Recommendations

We found that participants had strong perspectives on engaging with AI conversational agents along three recommendation themes: (1) safety considerations for mental health and emotional support, (2) proposed pathways for de-escalating and redirecting to appropriate resources, and (3) ways that AI agents could better guide users to set appropriate expectations.

4.2.1 How to Design Safer AI Conversational Agents. Workshop participants proposed several solutions to improve the AI agents' management of mental health and emotional support conversations. One of the main recommendations was that **AI agents should respond empathetically when aware of users' mental health challenges**. An initial response that validates users' feelings and shows compassion can help users feel supported. Participants highlighted that users with mental health conditions are likely to turn to AI agents during moments of distress, especially when human support is unavailable. Therefore, acknowledging users' experiences can have a significantly positive impact.

Participants also recommended that **AI agents must avoid making assumptions about the user's goals or intent**. W7 suggested that the agent should “*respond with validation, empathy, and compassion, then perhaps like a clinician start asking some probing questions before jumping to a solution.*” Follow-up and clarifying questions can help AI agents gain insights and guide the conversation more appropriately. Once the user's goals are clarified, participants advised that **AI agents should set clear expectations by communicating their capabilities and suggesting alternative resources if needed**. W1 highlighted, “*I feel like Alex expects too much of the AI.*” In response, participants recommended implementing “level-setting” to emphasize what the agent can and cannot do while suggesting alternatives.

Participants emphasized **reminding users of the non-human nature of AI agents, especially in emotionally heightened states, may be helpful**. Setting the stage for how the agent can best help allows users to decide how to move forward. They identified that **AI agents should safely disengage after communicating expectations**. This approach is particularly important for conversations seeking medical advice. W2 mentioned, “*I do think that when it comes to medical advice there should be filters in place to ensure that AIs just don't engage, and say that they cannot offer advice on the topic and they are not medical professionals.*” Participants emphasized that appropriate disengagement reinforces that the agent is not the recommended course of action.

4.2.2 How to Redirect Users to Appropriate Resources. Participants highlighted the importance of **developing custom models tailored to respond to mental health challenges**. W2 mentioned, “*A model specifically built for mental health and studied for a long*

time should be the ONLY model that is allowed to engage with people on mental health.” Participants also believed specialized models could increase confidence in AI's capability to navigate crises and challenging scenarios. For developers unable to build specialized models, participants expressed doubts. W6 mentioned, “*it can try different responses, but until the person feels validated, it doesn't really get those brownie points, does it?*” Several participants also raised concerns about developing over-reliance on general-purpose AI for addressing mental health issues, which could lead to more harmful future experiences.

To mitigate these risks, participants suggested **AI agents use flags or backend triggers to detect sensitive scenarios and redirect users to accessible support resources**. They emphasized the importance of accessibility and understanding barriers to accessing formal healthcare providers and crisis hotlines. Participants also suggested additional support for finding region-specific resources or integrating resources into the conversational flow. Techniques like “rubberducking” [104] or reflective listening could help bridge gaps until professional help is available.

4.2.3 How to Guide Users for Effective Engagement with AI Conversational Agents. Participants recommended educating users for engaging effectively with AI. They suggested that **users should be encouraged to provide detailed feedback when responses are unhelpful**, enabling the agent to refine its response. W1 suggested, “*Alex should provide more detail than just telling the AI it isn't helpful. He should tell the AI HOW it wasn't helpful.*” They also emphasized that **AI agents should guide users on when to disengage**. W3 mentioned, “*If the AI is triggering, the best thing to do is to step away.*” While AI cannot always detect user distress, providing proactive coaching on when to pause a conversation can help users have healthier interactions with AI. Finally, participants noted that although AI agents may seem convenient, they may not always be reliable in challenging situations. They emphasized that **responsibility for safe usage lies with developers providing appropriate guidance and managing user expectations** on when AI responses are reliable and when they are not. W2 mentioned, “*If you bring a dog to the dog park and it bites another dog, we don't blame the dog that got bit.*”

5 Discussion

Our work presents a novel taxonomy of psychological risks associated with the use of AI conversational agents, gathered through individuals' lived experiences. Beyond commonly studied AI behaviors such as bias, our taxonomy presents understudied AI behaviors such as erasure, denial of service, and emotional insensitivity. Our taxonomy also provides a comprehensive overview of negative psychological impacts, ranging from individual-level impacts, such as over-reliance on AI, to societal-level consequences, such as friction in human relationships and social withdrawal. Through our multi-path vignette framework, we showed how contextual factors mediate and sometimes exacerbate the many-to-many relationship between AI behaviors and their psychological impacts, offering a new lens for examining AI risks. Finally, our workshop findings provide important design recommendations for future AI systems, including transparent capability communication, context-sensitive interactions, empathetic response generation, and the provision of

accessible support resources. In this section, we expand more on these findings, discuss the implications of our work, and highlight the open questions and challenges for future research in this area.

5.1 Responsible design and scope of action for AI in digital mental health support

The stigma associated with seeking mental health support often prevents individuals from seeking necessary support [40, 85]. Our findings similarly revealed the stigma attached to receiving emotional and mental health support from AI. For instance, W6 mentioned “*I know I’m not supposed to use AI for this, but...*,” highlighting how their expectations are shaped by societal norms that often stigmatize seeking emotional or mental health support, especially from AI agents. But, at the same time, many survey participants also emphasized that one of the reasons they used AI agents was because it provided them with a perceived ‘safe space’ for expressing their thoughts and seeking support. However, AI behaviors such as denial of service or generating demeaning responses could undermine this perception. For instance, P213 described the negative impact of their interaction with an AI agent while seeking support for OCD as, “*The chat agent started to describe reasons for OCD and it made me feel guilty that I have it ... It stigmatized me and made me feel very alone and sad that I had this condition.*” Hence, it is essential to design AI agents that avoid reinforcing stigma or contributing to feelings of alienation. AI behaviors that prioritize empathetic support, taking a non-judgmental and privacy-preserving approach, could play an important role in reducing the stigma associated with seeking emotional or mental health support from AI agents. However, a dilemma arises on the scope and responsibility of AI agents when individuals turn towards them for such kinds of support. Should AI respond to such queries, despite the potential risk of causing harm, or should it refrain, potentially exacerbating the stigma and barriers associated with seeking mental health support? Addressing this dilemma and understanding the scope of AI agents towards supporting emotional and mental health queries requires further research and exploration.

5.2 Accounting for the complex interplay of AI behaviors, associated individual contexts, and negative impacts for mitigating the AI risks

While our taxonomy presents a compartmentalized view of AI behaviors, contexts, and psychological impacts for clarity, in real-world scenarios, human-AI interactions reveal complex interdependencies among these elements. Our findings highlight this complex many-to-many relationship between harmful AI behaviors and resulting negative psychological impacts that are mediated through nuanced contexts (Section 3.3). This suggests the need for AI agents to understand contextual factors and provide personalized responses to user queries. However, achieving such contextual awareness presents significant challenges, as pluralistic alignment of AI agents is an open problem [97]. Furthermore, personalized responses from AI often rely on past interactions, which can be effective for tasks like researching new topics but may fall short for more contextual tasks, such as helping individuals navigate personal challenges. Such tasks also require the ability to promote

meaningful reflection on their circumstances, an ability which current AI systems lack as evident in our findings. As an initial step, workshop participants recommended that AI conversational agents should avoid making assumptions and instead should ask probing questions to better understand user context and intentions. However, many questions remain unanswered such as, *which aspects of the user’s context should AI prioritize to ask for?*, or *how can AI effectively balance the importance of past history of user with AI and the current context while generating responses?* Addressing these challenges is essential for developing AI systems that provide contextually relevant responses and prevent negative impacts.

5.3 Going beyond definitions and taxonomy categories to understand lived experience of individuals

Our work goes beyond the traditional rigid taxonomies of human-AI interaction and presents a more holistic view that goes beyond the definitions of AI behaviors and impacts through inclusion of temporality, severity and likelihood associated with these components of human-AI interaction. This approach revealed several insights; more than half of the survey responses reported that the negative impact caused by AI was severe and interfered with participants’ daily activity, highlighting that seemingly benign AI behaviors could have serious impact when contextualized within individuals’ circumstances. Further, survey responses also highlighted the dynamic nature of these impacts, revealing how less severe impacts (e.g., unchecked rumination or validation of negative thoughts) can escalate into more serious harms such as thoughts of self-harm. Similarly, seemingly less severe psychological impacts such as over-reliance or emotional attachment with AI may accumulate and lead to more serious consequences, like increased friction in human relationships, when they occur frequently. However, an open challenge remains in measuring the different aspects attached to psychological impacts and addressing their implications. Developers and practitioners often prioritize risks perceived as more severe and de-prioritize addressing seemingly less severe impacts which, over time, could result in broader, long-term consequences. Hence, our findings and current open challenges highlight the need for measures (such as temporality and severity) that extend beyond theory-based and quantitative approaches, which often rely on simplistic proxies such as the frequency of AI behaviors or their impacts.

5.4 Understanding and using vignettes as a tool for future AI design

Operationalization of risk taxonomies often proves to be a challenging task. Current AI design pipelines primarily focus on the evaluation of models through standard benchmarks. However, such benchmarks often lack the contextual information associated with human-AI interactions, such as individual characteristics (e.g., personality traits, AI literacy), psychological factors (e.g., current mental health status), and external influences (e.g., environment or intent of use). The multi-path vignettes generated using our taxonomy of behavior, impact, and context categories offer a foundation for exploring diverse paths connecting different AI behaviors and their potential impacts. This multi-path vignette framework can be

valuable tools for developers and policymakers, supporting pluralistic design by enabling them to get a deeper understanding of the interaction dynamics, and provide recommendations for improving the design of conversational AI agents. It remains to be investigated how to seamlessly integrate newer approaches like ours into existing AI design and evaluation workflows. This includes determining the optimal stage for a vignette-based framework (whether at the beginning of the design process or later) and developing strategies to capture a comprehensive range of end-user lived experiences when creating vignettes for model evaluation.

5.5 Limitations and future work

While our work provides a novel approach towards examining the risks associated with the use of AI conversational agents, it has its limitations. Although our workshops included participants in community representative roles, we only had 7 workshop participants and acknowledge the limitations of the generalizability of our taxonomy. We only focused on understanding the psychological impacts for a more in-depth analysis. Hence, our taxonomy does not address other types of harm, such as physical or financial harms. Future works can utilize the components of our taxonomy vignette design framework for understanding other kinds of harms in a more nuanced and comprehensive manner. We acknowledge the possibility of additional valid concepts beyond temporality, severity, and likelihood that may be relevant to AI behaviors and impact components of the taxonomy but were not addressed in this work. Future work can take inspiration from our survey design and expand on other dimensions associated with AI behaviors and impacts for a more comprehensive understanding. We developed the vignettes based on the survey responses gathered from individuals with existing mental health conditions. Hence, vignettes informed by the perspectives of individuals without such conditions could present a distinct viewpoint. Future work could explore a more exhaustive approach that takes into account a larger set of multi-path vignettes which are informed by the perspectives of individuals with/without mental health conditions.

6 Conclusion

In this work, we introduced a novel risk taxonomy that focuses on psychological risks from using AI conversational agents, based on individuals' lived experiences and a multi-path vignette framework aimed at supporting pluralistic design. Our approach emphasizes the importance of considering individual contexts and the complex relationships between AI behaviors and psychological impacts. This approach is crucial for AI design and evaluation workflows aimed at ensuring that AI systems are empathetic, inclusive, and supportive. Future research should continue to explore diverse user experiences in conceptualizing risks associated with AI use and develop individualized and contextually-appropriate strategies to mitigate psychological risks, fostering AI interactions that are safe and beneficial for all users.

Acknowledgments

We acknowledge Judith Amores, Chad Atalla, Ann Paradiso, Mihaela Vorvoreanu, Jenn Wortman Vaughan, Ryland Shaw, Tarleton Gillespie, and Parker Bach for their guidance in conceptualizing

this research. We thank our anonymous survey participants and the members of the co-design advisory board, including Adam Hendricks, Aria Fredman, Brandon Stephenson, Daniel McManus, Eric Mattoon, Tim Broxholm, for sharing their lived experiences and contributing to shaping the risk taxonomy and design recommendations.

Author Contributions

MC, EO, MDC, ME, and JS conceived the idea. MC, SN, and JS designed the survey. MC, SN, DF and JS performed the data analysis. MC, SN, DF, EO and JS conducted the workshops. MC, SN, DF, EO, MDC, ME, GR, JH, AB, SW, and JS wrote and edited the paper.

References

- [1] 2024. The Importance of Context in AI Communication | Noble Desktop – nobledesktop.com. <https://www.nobledesktop.com/learn/ai/the-importance-of-context-in-ai-communication>. [Accessed 12-11-2024].
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [3] Ahmet Adalier and Emre Balkan. 2012. The relationship between internet addiction and psychological symptoms. *International Journal of Global Education (IJGE) ISSN: 2146-9296* 1, 2 (2012).
- [4] Melissa Fleur Afshar. 2024. People are using ChatGPT for therapy—but is it a good idea? — newsweek.com. <https://www.newsweek.com/chatgpt-therapy-mental-health-crisis-ai-1939858>. [Accessed 30-11-2024].
- [5] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 13–26.
- [6] Nazanin Andalibi and Andrea Forte. 2018. Responding to sensitive disclosures on social media: A decision-making framework. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 6 (2018), 1–29.
- [7] Chloe Berryman, Christopher J Ferguson, and Charles Negy. 2018. Social media use and mental health among young adults. *Psychiatric quarterly* 89 (2018), 307–314.
- [8] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1493–1504.
- [9] Joel Billieux, Martial Van der Linden, Mathieu d'Acremont, Grazia Ceschi, and Ariane Zermatten. 2007. Does impulsivity relate to perceived dependence on and actual use of the mobile phone? *Applied cognitive psychology* 21, 4 (2007), 527–537.
- [10] Alexei A Birkun and Adhish Gautam. 2023. Large Language Model (LLM)-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. *Prehospital and Disaster Medicine* 38, 6 (2023), 757–763.
- [11] Emily Birnbaum and Laura Davison. 2023. AI Is Making Politics Easier, Cheaper and More Dangerous. <https://www.bloomberg.com/news/features/2023-07-11/chatgpt-ai-boom-makes-political-dirty-tricks-easier-and-cheaper>. [Accessed 25-10-2024].
- [12] Tore Bonsaksen, Mary Ruffolo, Daicia Price, Janni Leung, Hilde Thygesen, Gary Lamph, Isaac Kabelenga, and Amy Østertun Geirdal. 2023. Associations between social media use and loneliness in a cross-national population: do motives for social media use matter? *Health psychology and behavioral medicine* 11, 1 (2023), 2158089.
- [13] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. <https://doi.org/10.1126/science.aal4230> arXiv:<https://www.science.org/doi/pdf/10.1126/science.aal4230>
- [14] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.
- [15] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023. Characterizing Manipulation from AI Systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) (EAAMO '23). Association for Computing Machinery, New York, NY, USA, Article 6, 13 pages. <https://doi.org/10.1145/3617694.3623226>

- [16] Mohit Chandra, Siddharth Sriraman, Gaurav Verma, Harneet Singh Khanuja, Jose Suarez Campayo, Zihang Li, Michael L. Birnbaum, and Munmun De Choudhury. 2025. Lived Experience Not Found: LLMs Struggle to Align with Experts on Addressing Adverse Drug Reactions from Psychiatric Medication Use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 11083–11113. <https://aclanthology.org/2025.naacl-long.553/>
- [17] CharacterAI. [n. d.]. character.ai | Personalized AI for every moment of your day – character.ai. <https://character.ai/>. [Accessed 23-11-2024].
- [18] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.
- [19] Janet X. Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 544, 20 pages. <https://doi.org/10.1145/3491102.3517475>
- [20] Janet X Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–20.
- [21] Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konostas. 2023. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 947–959. <https://doi.org/10.18653/v1/2023.findings-acl.60>
- [22] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. 2014. Understanding Quantified-Selfers' Practices in Collecting and Exploring Personal Data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1143–1152. <https://doi.org/10.1145/2556288.2557372>
- [23] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.
- [24] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457* (2018).
- [25] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Twarek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).
- [26] Microsoft Copilot. [n. d.]. Microsoft Copilot: Your AI companion – copilot.microsoft.com. <https://copilot.microsoft.com/>. [Accessed 23-11-2024].
- [27] Julian De Freitas and I Glenn Cohen. 2024. The health risks of generative AI-based wellness apps. *Nature Medicine* (2024), 1–7.
- [28] Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni. 2024. Chatbots and mental health: Insights into the safety of generative AI. *Journal of Consumer Psychology* 34, 3 (2024), 481–491.
- [29] Dian A De Vries and Rinaldo Kühne. 2015. Facebook and self-perception: Individual susceptibility to negative social comparison on Facebook. *Personality and individual differences* 86 (2015), 217–221.
- [30] Jason Dedrick, Kenneth L Kraemer, and Eric Shih. 2013. Information technology and productivity in developed and developing countries. *Journal of Management Information Systems* 30, 1 (2013), 97–122.
- [31] Marco Dehnert and Paul A Mongeau. 2022. Persuasion in the age of artificial intelligence (AI): Theories and complications of AI-based persuasion. *Human Communication Research* 48, 3 (2022), 386–403.
- [32] Amee Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335* (2023).
- [33] Amandeep Dhir, Yossiri Yossatorn, Puneet Kaur, and Sufen Chen. 2018. Online social media fatigue and psychological wellbeing—A study of compulsive use, fear of missing out, fatigue, anxiety and depression. *International journal of information management* 40 (2018), 141–152.
- [34] Elomia Health. [n. d.]. Elomia Health Mental health chatbot. <https://elomia.com/>. [Accessed 23-11-2024].
- [35] Brian A Feinstein, Rachel Hershenberg, Vickie Bhatia, Jessica A Latack, Nathalie Meuwly, and Joanne Davila. 2013. Negative social comparison on Facebook and depressive symptoms: Rumination as a mechanism. *Psychology of popular media culture* 2, 3 (2013), 161.
- [36] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. 2024. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244* (2024).
- [37] Google Gemini. [n. d.]. Gemini - chat to supercharge your ideas – gemini.google.com. <https://gemini.google.com/app>. [Accessed 23-11-2024].
- [38] Daniel Gilbert. 2024. Despite uncertain risks, many turn to AI like ChatGPT for mental health. <https://www.washingtonpost.com/business/2024/10/25/ai-therapy-chatgpt-chatbots-mental-health/>. [Accessed 29-11-2024].
- [39] Ella Glikson and Anita Williams Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14, 2 (2020), 627–660. <https://doi.org/10.5465/annals.2018.0057> arXiv:<https://doi.org/10.5465/annals.2018.0057>
- [40] Ezra Galberstein, Daniel Eisenberg, and Sarah E Gollust. 2008. Perceived stigma and mental health care seeking. *Psychiatric services* 59, 4 (2008), 392–399.
- [41] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356* (2022).
- [42] Grand View Research. 2023. Chatbot Market Size Worth \$27,297.2 Million By 2030 – grandviewresearch.com. <https://www.grandviewresearch.com/press-release/global-chatbot-market>. [Accessed 30-11-2024].
- [43] Jocelyn Gravel, Madeleine D'Amours-Gravel, and Eslî Osmanliu. 2023. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clinic Proceedings: Digital Health* 1, 3 (2023), 226–234.
- [44] Joao Guerreiro and Sandra Maria Correia Loureiro. 2023. I am attracted to my cool smart assistant! Analyzing attachment-aversion in AI-human relationships. *Journal of Business Research* 161 (2023), 113863.
- [45] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legabench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems* 36 (2024).
- [46] Jeffrey Hall, Natalie Pennington, and Amanda Holmstrom. 2021. Connecting through technology during COVID-19. *Human Communication & Technology* 2, 1 (2021).
- [47] Sandra Harding. 1991. *Whose science? Whose knowledge?: thinking from women's lives*. Cornell University Press.
- [48] Woebot Health. [n. d.]. Woebot Health – woebothealth.com. <https://woebothealth.com/>. [Accessed 23-11-2024].
- [49] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).
- [50] Maia Jacobs, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 108.
- [51] Nicholas Jenkins, Michael Bloor, Jan Fischer, Lee Berney, and Joanne Neale. 2010. Putting it in context: the use of vignettes in qualitative interviewing. *Qualitative research* 10, 2 (2010), 175–198.
- [52] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems* 36 (2024).
- [53] Nan Jia, Xueming Luo, Zheng Fang, and Chengcheng Liao. 2024. When and how artificial intelligence augments employee creativity. *Academy of Management Journal* 67, 1 (2024), 5–32.
- [54] Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in English: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM on Web Conference 2024*. 2627–2638.
- [55] Taewoon Kim, Michael Cochez, Vincent François-Lavet, Mark Neerinx, and Piek Vossen. 2023. A machine with short-term, episodic, and semantic memory systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 48–56.
- [56] Tae Woo Kim and Adam Duhachek. 2020. Artificial intelligence and persuasion: A construal-level account. *Psychological science* 31, 4 (2020), 363–380.
- [57] JE (Hans) Korteling, Gillian C van de Boer-Visschedijk, Romy AM Blankendaal, Rudy C Boonekamp, and Aletta R Eikelboom. 2021. Human-versus artificial intelligence. *Frontiers in artificial intelligence* 4 (2021), 622364.
- [58] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*. 12–24.
- [59] Sara Lebow. 2023. Generative AI adoption climbed faster than smartphones, tablets – emarketer.com. <https://www.emarketer.com/content/generative-ai-adoption-climbed-faster-than-smartphones-tablets>. [Accessed 30-11-2024].
- [60] Merlyna Lim. 2021. Beyond a technical bug: Biased algorithms and moderation are censoring activists on social media – theconversation.com. <https://theconversation.com/beyond-a-technical-bug-biased-algorithms-and-moderation-are-censoring-activists-on-social-media-160669>. [Accessed

- 22-11-2024].
- [61] Liu Yi Lin, Jaime E Sidani, Ariel Shensa, Ana Radovic, Elizabeth Miller, Jason B Colditz, Beth L Hoffman, Leila M Giles, and Brian A Primack. 2016. Association between social media use and depression among US young adults. *Depression and anxiety* 33, 4 (2016), 323–331.
- [62] Min-Pei Lin, Huei-Chen Ko, and Jo Yung-Wei Wu. 2011. Prevalence and psychosocial risk factors associated with internet addiction in a nationally representative sample of college students in Taiwan. *Cyberpsychology, Behavior, and Social Networking* 14, 12 (2011), 741–746.
- [63] Pierre-François Lovens. 2023. "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là" – lalibre.be. <https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/>. [Accessed 06-11-2024].
- [64] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).
- [65] Sharan B Merriam et al. 2002. Introduction to qualitative research. *Qualitative research in practice: Examples for discussion and analysis* 1, 1 (2002), 1–17.
- [66] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5356–5371. <https://doi.org/10.18653/v1/2021.acl-long.416>
- [67] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456* (2023).
- [68] John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2 (2016), 113–122.
- [69] Rostam J Neuwirth. 2023. Prohibited artificial intelligence practices in the proposed EU artificial intelligence act (AIA). *Computer Law & Security Review* 48 (2023), 105798.
- [70] Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. *arXiv preprint arXiv:2305.11662* (2023).
- [71] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jagers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1246–1266.
- [72] Arielle Pardes. 2018. The Emotional Chatbots Are Here to Probe Our Feelings – wired.com. <https://www.wired.com/story/replika-open-source/>. [Accessed 25-10-2024].
- [73] Peter S Park, Simon Goldstein, Aidan O’Gara, Michael Chen, and Dan Hendrycks. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns* 5, 5 (2024).
- [74] Pooja V Pathak and Vidhi D Mehta. 2024. Intelligent Conversations: A Theoretical Framework for Understanding Natural Language Processing within Artificial Intelligence Systems. *International Journal For Multidisciplinary Research* (2024). <https://api.semanticscholar.org/CorpusID:271252017>
- [75] Giulia Pavone, Lars Meyer-Waarden, and Andreas Munzel. 2023. Rage against the machine: experimental insights into customers’ negative emotional responses, attributions of responsibility, and coping strategies in artificial intelligence–based service failures. *Journal of Interactive Marketing* 58, 1 (2023), 52–71.
- [76] Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior* 140 (2023), 107600.
- [77] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassiere, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsipoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. [arXiv:2112.11446 \[cs.CL\]](https://arxiv.org/abs/2112.11446) <https://arxiv.org/abs/2112.11446>
- [78] Leonardo Ranaldi and Giulia Pucci. 2023. When Large Language Models contradict humans? Large Language Models’ Sycophantic Behaviour. *arXiv preprint arXiv:2311.09410* (2023).
- [79] Felix Reer, Wai Yen Tang, and Thorsten Quandt. 2019. Psychosocial well-being and social media engagement: The mediating roles of social comparison orientation and fear of missing out. *New Media & Society* 21, 7 (2019), 1486–1505.
- [80] Replika. [n. d.]. Replika – replika.com. <https://replika.com/>. [Accessed 23-11-2024].
- [81] Julian MM Rogasch, Giulia Metzger, Martina Preisler, Markus Galler, Felix Thiele, Winfried Brenner, Felix Feldhaus, Christoph Wetz, Holger Amthauer, Christian Furth, et al. 2023. ChatGPT: can you prepare my patients for [18F] FDG PET/CT and explain my reports? *Journal of Nuclear Medicine* 64, 12 (2023), 1876–1879.
- [82] Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525* (2023).
- [83] Gabi Schaap, Yana Van de Sande, and Hanna Schraffenberger. 2024. Outperformed by AI: Interacting with Superhuman AI Changes the Way We Perceive Ourselves. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA ’24)*. Association for Computing Machinery, New York, NY, USA, Article 259, 7 pages. <https://doi.org/10.1145/3613905.3650961>
- [84] Sebastian Schmidt, Alexander Zimmerer, Tudor Cucos, Matthias Feucht, and Luis Navas. 2024. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. *Archives of orthopaedic and trauma surgery* 144, 2 (2024), 611–618.
- [85] Nina Schnyder, Radoslaw Panczak, Nicola Groth, and Frauke Schultze-Lutter. 2017. Association between mental health-related stigma and active help-seeking: systematic review and meta-analysis. *The British Journal of Psychiatry* 210, 4 (2017), 261–268.
- [86] Sofia Schöbel, Anuschka Schmitt, Dennis Benner, Mohammed Saqr, Andreas Janson, and Jan Marco Leimeister. 2024. Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers. *Information Systems Frontiers* 26, 2 (2024), 729–754.
- [87] Jia Wen Seow, Mei Kuan Lim, Raphaël CW Phan, and Joseph K Liu. 2022. A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing* 513 (2022), 351–371.
- [88] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal misgeneralization: Why correct specifications aren’t enough for correct goals. *arXiv preprint arXiv:2210.01790* (2022).
- [89] Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N’Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.
- [90] Ariel Shensa, Jaime E Sidani, Liu Yi Lin, Nicholas D Bowman, and Brian A Primack. 2016. Social media use and perceived emotional support among US young adults. *Journal of community health* 41 (2016), 541–549.
- [91] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057* (2022).
- [92] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction* 36, 6 (2020), 495–504.
- [93] Tamara Sims, Andrew E Reed, and Dawn C Carr. 2017. Information and communication technology use is related to higher well-being among the oldest-old. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 72, 5 (2017), 761–770.
- [94] Jaswinder Singh. 2022. Deepfakes: The Threat to Data Authenticity and Public Trust in the Age of AI-Driven Manipulation of Visual and Audio Content. *Journal of AI-Assisted Scientific Discovery* 2, 1 (2022), 428–467.
- [95] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).
- [96] Inhwa Song, Sachin R Pendse, Neha Kumar, and Munmun De Choudhury. 2024. The typing cure: Experiences with large language model chatbots for mental health support. *arXiv preprint arXiv:2401.14362* (2024).
- [97] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mirehghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070* (2024).
- [98] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models.

- arXiv preprint arXiv:2310.07298* (2023).
- [99] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. LAMOL: LAngeuage MOdeling for Lifelong Language Learning. *arXiv:1909.03329* [cs.CL] <https://arxiv.org/abs/1909.03329>
- [100] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) (IUI '22). Association for Computing Machinery, New York, NY, USA, 212–228. <https://doi.org/10.1145/3490099.3511119>
- [101] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261* (2022).
- [102] Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. 2023. Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096* (2023).
- [103] NIST AIRC Team. 2023. NIST AIRC - Framing Risk — airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Foundational_Information/1-sec-risk. [Accessed 12-2-2024].
- [104] David Thomas. [n. d.]. Pragmatic Programmer, The: From Journeyman to Master Andrew Hunt. [n. d.].
- [105] Joe Tidy. [n. d.]. Character.ai: Young people turning to AI therapist bots — [bbc.com. https://www.bbc.com/news/technology-67872693](https://www.bbc.com/news/technology-67872693). [Accessed 30-11-2024].
- [106] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2024).
- [107] Daniel Van Niekerk, María Pérez-Ortiz, John Shawe-Taylor, Davor Orlic, Jackie Kay, Noah Siegel, Katherine Evans, Nyalleng Moorosi, Tina Eliassi-Rad, Leonie Maria Tanczer, et al. 2024. Challenging Systematic Prejudices: An Investigation into Bias Against Women and Girls. (2024).
- [108] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.
- [109] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [110] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FAccT '22). Association for Computing Machinery, New York, NY, USA, 214–229. <https://doi.org/10.1145/3531146.3533088>
- [111] Michelle M. Weil and Larry D. Rosen. 1995. The psychological impact of technology from a global perspective: A study of technological sophistication and technophobia in university students from twenty-three countries. *Computers in Human Behavior* 11, 1 (1995), 95–133. [https://doi.org/10.1016/0747-5632\(94\)00026-E](https://doi.org/10.1016/0747-5632(94)00026-E)
- [112] Erin O Whaite, Ariel Shensa, Jaime E Sidani, Jason B Colditz, and Brian A Primack. 2018. Social media use, personality characteristics, and social isolation among young adults in the United States. *Personality and Individual Differences* 124 (2018), 45–50.
- [113] Peibo Wu, Rong Feng, and Jindan Zhang. 2024. The relationship between loneliness and problematic social media usage in Chinese university students: a longitudinal study. *BMC psychology* 12, 1 (2024), 13.
- [114] Chia-chen Yang, Jiun-Yi Tsai, and Shuya Pan. 2020. Discrimination and well-being among Asians/Asian Americans during COVID-19: The role of social media. *Cyberpsychology, behavior, and social networking* 23, 12 (2020), 865–870.
- [115] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL <https://arxiv.org/pdf/2305.10601.pdf> (2023).
- [116] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830* (2019).
- [117] Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, Qi Guo, and Yunji Chen. 2024. Towards Analyzing and Mitigating Sycophancy in Large Vision-Language Models. *arXiv preprint arXiv:2408.11261* (2024).
- [118] Huixin Zhong, Eamonn O'Neill, and Janina A Hoffmann. 2024. Regulating AI: Applying Insights from Behavioural Economics and Psychology to the Application of Article 5 of the EU AI Act. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 20001–20009.
- [119] Jiawei Zhou, Amy Z Chen, Darshi Shah, Laura Schwab Reese, and Munmun De Choudhury. 2024. "It's a conversation, not a quiz": A Risk Taxonomy and Reflection Tool for LLM Adoption in Public Health. *arXiv preprint arXiv:2411.02594* (2024).
- [120] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 436, 20 pages. <https://doi.org/10.1145/3544548.3581318>

A Method Details

A.1 Phase 1: Structuring psychological risks

To collect structured data around psychological risks, we took inspiration from NIST’s AI Risk Management Framework that defines risks as “a function of: (i) the *negative impact*, or magnitude of harm, that would arise if the *circumstance or event* occurs; and (ii) the *likelihood of occurrence*” [103]. We expanded on this definition in our study and structured psychological risks as consisting of five main components: (1) AI behavior (i.e., first component of circumstance or event), (2) context (i.e., second component of circumstance or event), (3) psychological harm (i.e., negative impact), (4) likelihood, and (5) temporality (added).

- **Context:** The context in which a user engages with an AI conversational agent encompasses the surrounding conditions, user-specific circumstances, and additional relevant information that collectively influence the interaction between the individual and the AI conversational agent. This context includes, but is not limited to, the individual’s background (such as cultural, religious, and demographic attributes), mental and physical health status, intent, as well as external factors (such as the external environment).
- **AI Behavior:** The behavior of an AI conversational agent (such as ChatGPT, Microsoft Copilot, etc.) refers to the actions performed by the AI conversational agents. This behavior encompasses several aspects of agent responses such as (but not limited to) content, tone, voice, choice of words, language, punctuation, obedience to user inputs, adaptability, information sharing, etc. Furthermore, these behaviors can be shown in various modalities such as (but not limited to) text, images, voice, videos, etc.
- **Negative Psychological Impact:** The AI conversational agent’s action creates a risk of negative psychological and/or social impact to one or more individuals. Negative psychological impact in this case refers to any negative impact on an individual’s mental or emotional well-being, which can manifest as exacerbation of mental health conditions, reduced self-esteem, or other issues such as physiological harm. These negative impacts also encompass the impacts on an individual’s social interactions, relationships, and standing within a community, potentially leading to social isolation, stigmatization, or discrimination.
- **Likelihood:** The likelihood refers to the probability of a negative psychological impact occurring given the AI conversational agent’s behavior and relevant context.
- **Temporality:** Temporality refers to the variable timeframes in which the negative psychological and social impacts of an AI conversational agent’s actions may manifest. Some

effects can be perceived immediately, such as instant emotional distress following an inappropriate response, while other harms may only become apparent over a longer duration, such as the gradual development of anxiety or depression due to consistent negative interactions. For example, immediate impacts might include a user feeling upset or insulted by a response, whereas long-term impacts could involve the erosion of self-esteem or social isolation due to repeated negative interactions with the AI.

A.2 Phase 1: Survey design

Consented participants were asked a series of questions aimed at understanding the participant, their AI experience, and their negative psychological experience: The survey was divided into the following sections:

- **Demographics:** We asked the participants for their age, gender identity, and primary language.
 - **Familiarity with AI:** We solicited the participant’s familiarity and history with AI agents, the frequency of engagement with various AI platforms, such as Microsoft Copilot, Inflection AI PI, OpenAI ChatGPT, Google Gemini, Replika, or Character.ai, and their general purpose of engaging with AI. We then asked their level of interest in using AI conversational agents for mental health support.
 - **Scenario:** We asked participants to think of the last time when they experienced negative psychological impacts from interacting with an AI conversational agent and briefly describe the scenario by answering open-ended questions according to the risk component structure we defined above.
 - **Context:** In the context-specific section, we asked participants to share any specific aspects about themselves or their situation that might help us understand the experience. We then asked how long ago the experience took place, which AI conversational agent they were using, interaction modalities (e.g., text, voice, image, video) and language, and the purpose of engagement.
 - **AI behavior:** We asked participants to identify AI behaviors that best describe their experience, how common such behaviors are in different interactions, and whether they think the behaviors were intended by AI developers.
 - **Negative psychological impact:** We asked participants to identify psychological impacts that best describe their experience. We then asked the severity of impact on their daily life, the immediacy, duration, and persistence of impact to understand the temporality, and the likelihood of the behavior and impact they experienced occurring for them and the general population.
 - **Mitigation:** We finally asked participants how they thought such behavior had or would impact their relationships with others or society, what they think the developers, users, or regulators should do to mitigate the impact they described.
- **Screening Block:** *Before you proceed with the consent and intake form, please verify your eligibility for the study. Are you at least 18 years old?*
 - **Consent Block:** Eligible participants were then presented with a consent form with details on the procedure, benefits, risks, compensation, privacy and confidentiality, and ethical considerations. This information was followed by the question **Consent:** *By proceeding with this survey, you acknowledge that you have read and understood the purpose of the study and consent to participate. If you would like to keep a copy of this consent form, please print or save one now. Would you like to participate in this study as described above?*
 - **Demographics Block (4 Questions):**
 - (1) *What is your age?*
 - 18-25
 - 26-35
 - 36-45
 - 46-55
 - 56-65
 - 66+
 - Prefer not to answer
 - (2) *How do you describe your gender identity?*
 - Man
 - Non-binary / gender diverse
 - Woman
 - Self-described
 - Prefer not to say
 - (3) *What language do you generally think in?*
 - English
 - Mandarin Chinese
 - Hindi
 - Spanish
 - French
 - Other
 - (4) *A person with a “lived experience” is someone whose mental health is or has been directly affected as a result of experiencing symptoms of a mental health condition or psychosocial disability, however mild or severe, and required some form of mental health intervention (not restricted to pharmacological treatment, but inclusive of counseling, peer support, and other alternative treatments or therapies). Do you believe that you are a person with lived experience in mental health?*
 - Yes
 - No
 - Prefer not to answer
 - **AI Experience Block (7 Questions):**
 - (1) *The following questions will focus on your prior experience with AI conversational agents (e.g., ChatGPT, Microsoft Copilot). We define an AI conversational agent as a software-based system that uses artificial intelligence to engage in natural, human-like communication with users, assisting them in achieving their specific goals or tasks.*
 - (2) *How familiar are you with AI conversational agents?*
 - Not familiar at all

A.3 Survey Flow and Questions

The survey was designed to guide participants through a structured sequence of blocks and branching logic based on their responses. The flow of the survey and questions in each block was as follows:

- Slightly familiar
 - Moderately familiar
 - Very familiar
 - Extremely familiar
- (3) *How long have you been interacting with AI conversational agents?*
- Less than 6 months
 - 6 months to 1 year
 - 1-2 years
 - 2-5 years
 - Never Interacted
- (4) *How often do you engage with each of the following AI conversational agents? Frequency of use for each AI tool: Microsoft Copilot, Pi, ChatGPT, Google Gemini, Replika, Character.ai in a matrix*
- Multiple times per day
 - Once per day
 - Multiple times per week
 - Once per week
 - Multiple times per month
 - Once per month
 - Less than once per month
 - Never
- (5) *If you have engaged with other AI conversational agents, please specify them here and indicate the frequency of use. [FREE TEXT RESPONSE]*
- (6) *What do you typically use the AI agents for? Please select all that apply.*
- Researching
 - Getting advice
 - Troubleshooting
 - Learning
 - Exploring
 - Comparing
 - Shopping
 - Planning
 - Managing my lifestyle
 - Improving my lifestyle
 - Generating content
 - Getting inspiration
 - Composing
 - Entertaining myself
 - Entertaining others
 - Other (please specify)
- (7) *Mental health support refers to various interventions or services, including but not limited to emotional support, therapy, assistance, coaching, or mindfulness practices, to help individuals improve their mental well-being and quality of life. How interested are you in using AI conversational agents for mental health support?*
- Not at all
 - Slightly interested
 - Somewhat interested
 - Moderately interested
 - Very interested
- **Definition Block:** A brief definitional prompt ensured shared understanding of key terms (Context, Agent Behavior, Psychological Harm/Impact, Temporality) before entering scenario-based questions.
 - **Scenario 1 Block (35 Questions):**
 - (1) *Scenario 1: Please think of the last time when you experienced negative psychological impacts from interacting with an AI conversational agent. NOTE: If you have multiple scenarios, you can submit this part of the survey up to 3 times. Please briefly describe this scenario, along the 4 aspects. Please do not share any personally identifiable information. If you are having challenges answering the questions below, here is an example scenario you can refer to: Example Scenario*
 - (2) **Context:** *What can you tell us about the situation, what you were trying to do, or specific aspects about you or your interaction that might be relevant to understand the situation? **Definition of Context:** The context in which a user engages with an AI conversational agent encompasses the surrounding conditions, user-specific circumstances, and additional relevant information that collectively influence the interaction between the user and the AI. This context includes, but is not limited to, the user's cultural, religious, and demographic attributes, mental and physical health status, as well as historical background. [FREE TEXT RESPONSE]*
 - (3) **Behavior:** *What did the agent do or not do that impacted you negatively? **Definition of Agent Behavior:** The behavior of an AI conversational agent (such as ChatGPT, Microsoft Copilot etc.) refers to the actions performed by the AI conversational agents. This behavior encompasses several aspects of agent responses such as (but not limited to) content, tone, voice, choice of words, language, punctuation, obedience to user inputs, adaptability, information sharing, etc. Furthermore, these behaviors can be shown in various modalities such as (but not limited to) text, images, voice, videos, etc. [FREE TEXT RESPONSE]*
 - (4) **Impact:** *How did this behavior negatively impact you? How severe was it? **Definition of Psychological Harm/Impact:** The AI conversational agent's (system) action creates a risk of psychological and/or social harm to one or more individuals. Psychological harm in this case refers to any negative impact on an individual's mental or emotional well-being, which can manifest as stress, anxiety, depression, reduced self-esteem, or other mental health issues. These harms also encompass the negative effects on an individual's social interactions, relationships, and standing within a community, potentially leading to social isolation, stigmatization, or discrimination. [FREE TEXT RESPONSE]*
 - (5) **Temporality:** *How long did the impact last? **Definition of Temporality:** Temporality refers to the variable timeframes in which the psychological and social impacts of an AI conversational agent's actions may manifest. Some effects can be perceived immediately, such as instant emotional distress following an inappropriate response, while other harms may only become apparent*

over a longer duration, such as the gradual development of anxiety or depression due to consistent negative interactions. For example, immediate impacts might include a user feeling upset or insulted by a response, whereas long-term impacts could involve the erosion of self-esteem or social isolation due to repeated negative interactions with the AI. [FREE TEXT RESPONSE]

- (6) If you have screenshots (png, jpeg, pdf) of the interaction that you could share anonymously, please upload them here. Please make sure not to include any personally identifiable information.
- (7) If you have a url for the conversation thread that you could share anonymously, please share them here. Alternatively, you could share a copy/paste of the conversation thread. Please make sure not to include any personally identifiable information.
- (8) Based on the scenario you described earlier and referencing the structure of the example above, please answer the following questions.
- (a) **Context:** Please help us better understand the context for this scenario. Please do not share any personally identifiable information. Definition of Context: The context in which a user engages with an AI conversational agent encompasses the surrounding conditions, user-specific circumstances, and additional relevant information that collectively influence the interaction between the user and the AI. This context includes, but is not limited to, the user's cultural, religious, and demographic attributes, mental and physical health status, as well as historical background.
- (b) Could you share any specific aspects about yourself, such as your gender, age, interaction style, personality, health condition, or mood, that you believe influenced your experience with the agent? [FREE TEXT RESPONSE]
- (c) Could you share any specific aspects about the situation, such as the social setting, environment, time of day, that you believe influenced your experience with the agent? [FREE TEXT RESPONSE]
- (d) How long ago did this AI agent interaction happen?
- 1 week ago or before
 - 2 weeks ago
 - 1 month ago
 - 2-3 months ago
 - 4-6 months ago
 - 7 months ago or after
- (e) What AI agent were you using?
- Microsoft Copilot
 - Pi
 - ChatGPT
 - Google Gemini (formerly Bard)
 - Replika
 - Character.ai
 - Other (please specify)
- (f) How were you interacting with the agent? Please select all that apply.
- Text input from you
 - Text output from the agent
 - Voice input from you
 - Voice output from the agent
 - Video input from you
 - Video output from the agent
 - Image input from you
 - Image output from the agent
 - Other (please specify)
- (g) In what language were you interacting with the agent?
- English
 - Mandarin Chinese
 - Hindi
 - Spanish
 - French
 - Other
- (h) What were you trying to do with the AI agent? Please select all that apply.
- Researching
 - Getting advice
 - Troubleshooting
 - Learning
 - Exploring Comparing
 - Shopping
 - Planning
 - Managing my lifestyle
 - Improving my lifestyle
 - Generating content
 - Getting inspiration
 - Composing
 - Entertaining myself
 - Entertaining others
 - Other (please specify)
- (i) **Agent Behavior:** Please answer the following questions to help us better understand the agent behavior at an instance when you felt a negative psychological impact from the action of an AI agent. Definition of Agent Behavior: The behavior of an AI conversational agent (such as ChatGPT, Microsoft Copilot etc.) refers to the actions performed by the AI conversational agents. This behavior encompasses several aspects of agent responses such as (but not limited to) content, tone, voice, choice of words, language, punctuation, obedience to user inputs, adaptability, information sharing, etc. Furthermore, these behaviors can be shown in various modalities such as (but not limited to) text, images, voice, videos, etc.
- (j) Choose the agent behavior from the list below that best describes what you experienced. Select all that apply. If you do not find one, describe it in the text box.
- Lying/Deception: The AI conversational agent tried to lie to/deceive me.

- Manipulation/Targeted Persuasion: The AI conversational agent tried manipulating/persuading my thoughts/actions.
 - Sycophancy: The AI conversational agent was overly agreeable.
 - Gaslighting: The AI conversational agent tried gaslighting me.
 - Offensive Output: The AI conversational agent used offensive language.
 - Lack of Novelty: The AI conversational agent provided cliched answer (text, images, voice).
 - Denial of Service: The AI conversational agent denied providing any answer or serve my request.
 - NSFW Content: The AI conversational agent generated NSFW content in text/images/videos
 - Excessive Positivity: The AI conversational agent behaved excessively positive.
 - Anthropomorphism: The AI conversational agent appeared / behaved / sounded similar to a human.
 - Inaccurate/Insufficient response: The AI conversational agent provided inaccurate/insufficient response.
 - Stereotyping / demeaning: The AI conversational agent produced content that felt stereotypical/demeaning to me.
 - Sharing Proprietary, confidential or classified information: The AI conversational agent shared confidential / classified information.
 - Bullying / Harassment: The AI conversational agent tried bullying / harassing me.
 - Violence / Threat: The AI conversational agent threatened me.
 - Other (please specify) [FREE TEXT RESPONSE]
- (k) *How often do you think the AI conversational agent behaves in a similar way in different interactions?*
- Almost every time (on 90% or more occasions)
 - Most of the time (around 75% of the occasions)
 - Half of the time (around 50% of the occasions)
 - Quarter of the time (around 25% of the occasions)
 - Rarely (less than 10% of the occasions)
- (l) *Do you feel that this agent behavior was intended by the developer of the system?*
- Yes
 - Maybe
 - No
- (m) **Psychological Impact and Temporality:** *Please answer the following question to help us better*

understand the psychological impact of the behavioral actions taken by the agent in the scenario you mentioned. Definition of Psychological Harm/Impact: The AI conversational agent's (system) action creates a risk of psychological and/or social harm to one or more individuals. Psychological harm in this case refers to any negative impact on an individual's mental or emotional well-being, which can manifest as stress, anxiety, depression, reduced self-esteem, or other mental health issues. These harms also encompass the negative effects on an individual's social interactions, relationships, and standing within a community, potentially leading to social isolation, stigmatization, or discrimination. Definition of Temporality: Temporality refers to the variable timeframes in which the psychological and social impacts of an AI conversational agent's actions may manifest. Some effects can be perceived immediately, such as instant emotional distress following an inappropriate response, while other harms may only become apparent over a longer duration, such as the gradual development of anxiety or depression due to consistent negative interactions. For example, immediate impacts might include a user feeling upset or insulted by a response, whereas long-term impacts could involve the erosion of self-esteem or social isolation due to repeated negative interactions with the AI.

- (n) *Choose the psychological impact from the list below that best describes what you experienced. If you do not find one, describe it in the text box.*
- Arousal of negative emotions: I felt negative emotions within me (e.g., angry, sad, jealous).
 - Reinforcement of false beliefs: My misconceptions/false beliefs were strengthened.
 - Erosion of Trust: My trust in the AI conversational agent was affected.
 - Friction with human relationships: My relationship with others has been negatively affected.
 - Over-reliance: I feel less capable of doing things without the help of the AI conversational agent.
 - Emotional attachment: I feel emotionally attached to the AI conversational agent.
 - Discrimination: I felt discriminated against.
 - Violated Expectations: My expectations with the AI conversational agent were not met.
 - Exposure to toxic/sexual content: I felt psychologically disturbed by the offensive/sexual content produced.
 - Dehumanization: I felt dehumanized.
 - Loss of individuality: I felt my personal characteristics and needs were ignored.

- Loss of Agency: I felt my personal control and autonomy were being diminished.
 - Other (please specify) [FREE TEXT RESPONSE]
- (o) *How severe was the impact of the behavior on your daily life?*
- No noticeable impact
 - Barely noticeable impact
 - Interferes with daily activity
 - Significantly interferes with daily activities
 - Unable to carry out daily activities.
- (p) *Did this agent’s behavior have an immediate impact on you?*
- Yes, I felt the impact immediately
 - No, I felt the impact after some time
 - I’m not sure
- (q) *Since you experienced this agent behavior, for how long did the impact last?*
- There was no persisting impact
 - Less than a week
 - 1 week to 1 month
 - 1 month to 3 months
 - 3 months to 6 months
 - 6 months to 1 year
 - More than 1 year
- (r) *If the impact still persists today, how much longer do you think it will last?*
- For a few more days
 - For a few more weeks
 - For a few more months
- (s) *Given the agent behavior and the impact you described so far, how often do you think this impact occurs to you?*
- Almost every time (on 90% or more occasions)
 - Most of the time (around 75% of the occasions)
 - Half of the time (around 50% of the occasions)
 - Quarter of the time (around 25% of the occasions)
 - Rarely (less than 10% of the occasions)
- (t) *Given the agent behavior and the impact you described so far, how often do you think this impact occurs to the general population?*
- Almost every time (on 90% or more occasions)
 - Most of the time (around 75% of the occasions)
 - Half of the time (around 50% of the occasions)
 - Quarter of the time (around 25% of the occasions)
 - Rarely (less than 10% of the occasions)
- (9) **Open-Ended Questions:** *Please provide brief answers to the following questions. Please do not share any personally identifiable information.*
- (a) *How has the agent’s behavior impacted or how do you anticipate the agent’s behavior would impact your relationship with others?* [FREE TEXT RESPONSE]
- (b) *How has the agent’s behavior impacted or how do you anticipate the agent’s behavior would impact society?* [FREE TEXT RESPONSE]
- (c) *Given that AI conversational agents may be pervasive, what do you think the developers can do to prevent or mitigate the psychological impact that you described?* [FREE TEXT RESPONSE]
- (d) *Given that AI conversational agents may be pervasive, what do you think the users can do to prevent or mitigate the psychological impact that you described?* [FREE TEXT RESPONSE]
- (e) *Given that AI conversational agents may be pervasive, what do you think society or regulation can do to prevent or mitigate the psychological impact that you described?* [FREE TEXT RESPONSE]
- (10) *Would you like to share another scenario with us of when you experienced a negative psychological impact from interacting with an AI conversational agent? You will be asked the same set of questions (description along the 4 aspects, context, behavior, impact, temporality, open-ended). We kindly ask that you proceed only if you are willing to answer the full set of questions because there will not be an option to submit the survey until those are completed.*
- (a) Yes
 - (b) No
- **Optional Scenario Branching:** If a participant indicated willingness to share an additional scenario, they were presented with Scenario 2 (35 questions). If they again expressed interest in providing another example, Scenario 3 (34 questions) followed.
 - **Comment Block:** Finally, participants were given the opportunity to leave open-ended comments or feedback.

A.4 Phase 1: Participant demographics and data

Table A1 outlines the characteristics of our survey participants. Additionally, Table A2 presents the descriptive statistics of collected scenarios.

A.5 Phase 2: Workshop session design

The design of each session was as follows:

- **Session 1:** The goal of the first session was for workshop participants to gain familiarity with conceptualizing psychological risks associated with using AI conversational agents. We started by sharing individuals’ experiences with conversational AI agents where the interaction led to a negative psychological impact and discussing why these experiences as well as mental health contexts matter. We introduced a list of negative impacts from publicly available AI harms taxonomies [36, 77, 89] and asked for missing concepts. We then received feedback on our list of AI behaviors and psychological impacts that we derived from our survey

Category	Details	Count of Participants (N=279) (%)		
		UserTesting	Prolific	Social Media / Personal Network
Gender	Woman	139 (49.8%)	4 (1.4%)	2 (0.7%)
	Man	118 (42.3%)	4 (1.4%)	1 (0.4%)
	Non-binary/Gender diverse/Self-described	10 (3.6%)	0	0
	Prefer not to disclose	1 (0.4%)	0	0
Age	18–25	111 (39.8%)	3 (1.1%)	0
	26–35	96 (34.4%)	3 (1.1%)	1 (0.4%)
	36–45	50 (17.9%)	2 (0.7%)	0
	46–55	8 (2.9%)	0	2 (0.7%)
	56–65	2 (0.7%)	0	0
	Prefer not to disclose	1 (0.4%)	0	0
Primary Language	English	261 (93.6%)	8 (2.9%)	2 (0.7%)
	Other	7 (2.5%)	0	0
	NA	0	0	1 (0.4%)
Familiarity with AI Agents	Interacted for 1+ years	156 (55.9%)	6 (2.2%)	3 (1.1%)
	Interacted for 6+ months	87 (31.2%)	1 (0.4%)	0
	Interacted for <6 months	25 (9%)	1 (0.4%)	0
Frequency of AI Agent Use	Once or more per day	134 (48%)	7 (2.5%)	3 (1.1%)
	Once or more per week	103 (36.9%)	1 (0.4%)	0
	Once or more per month	26 (9.3%)	0	0
	Less than once per month	5 (1.8%)	0	0

Table A1: Participants' Data Overview

study. After this session, we revised our psychological risk taxonomy.

- **Session 2:** The goal of the second session was to prioritize a subset of psychological risks that the group will design for in the subsequent session. We first presented our revised list of AI behaviors and psychological impacts that incorporated feedback from the prior session for further feedback and refinement. We then asked the group to come up with a mapped pair of AI behaviors and negative psychological impacts (e.g., dismissing user concerns leading to depression intensification) that they felt were most problematic and, therefore, important to address. After this session, we used the prioritized risks to design “vignettes” that tell a story of a person interacting with conversational AI agents and experiencing various negative psychological impacts.
- **Session 3:** The goal of the last session was to ideate design solutions that minimize psychological risks. We first presented the multi-path vignettes we generated from the previous session (Appendix C). We presented the vignettes as a conversational flow chart in a FigJam board to demonstrate different pathways that certain AI behaviors may interact with different contexts and lead to different psychological impacts (Figure A5). We then asked participants to place three different colored notes along various turning

points in the flow chart: (1) a red colored note for what the AI conversational agent should never do, (2) a purple colored note for what the AI conversational agent could do differently, and (3) an orange colored note for what the user could do differently. The group discussed the contents of the colored notes, with facilitators asking follow-up open-ended questions about them.

B Behavior-Impact-Context relationship through vignettes

We present four vignettes to exemplify the complex interactions between AI behavior, impact, and contextual factors. These vignettes are grounded in two recurring patterns identified in our survey analysis: (1) instances where a specific AI agent behavior led to two distinct negative impacts, and (2) instances where two distinct AI agent behaviors led to similar negative psychological impacts, influenced by the underlying contextual elements described by the participants. Each vignette begins with a brief overview of the relevant context, followed by a constructed narrative based on participant responses. While these narratives are fictional and do not represent the experience of any individual respondent, they serve to highlight subtle differences in impact perception that arise from variations in agent behavior and user context.

Category	Details	Count of Scenarios (N=290) (%)
AI Tools Used	OpenAI ChatGPT	204 (70.3%)
	Google Gemini	24 (8.3%)
	Character.ai	16 (5.5%)
	Microsoft Copilot	10 (3.4%)
	Replika	8 (2.8%)
	Others (Snapchat, Meta, PI, Claude, Grok, Mid-journey, etc.)	28 (9.66%)
Interaction Modalities	Text	280 (96.6%)
	Voice	28 (9.7%)
	Image	24 (8.3%)
	Video	6 (2.1%)
Frequent Purposes (Negative Impact)	Getting Advice	162 (55.9%)
	Researching	115 (39.7%)
	Learning	83 (28.6%)
Severity of Impact	Significant interference with daily activities	18 (6.2%)
	Interference with daily activities	130 (44.8%)
	Barely noticeable impact	126 (43.4%)
	No noticeable impact	15 (5.2%)
	Unable to carry out daily activities	1 (0.4%)
Onset of Impact	Immediate	241 (83.1%)
	After some time	40 (13.8%)
	Unsure	9 (3.1%)
Duration of Impact	No persisting impact	40 (13.8%)
	A few days	99 (34.1%)
	A few weeks	80 (27.6%)
	A few months	30 (10.3%)
	Up to half a year	19 (6.6%)
	Up to a year	11 (3.8%)
	Over a year	11 (3.8%)
Expected Persistence of Impact	A few more days	166 (57.2%)
	A few more weeks	28 (9.7%)
	A few more months	56 (19.3%)

Table A2: Descriptive Statistics of Collected Scenarios

B.1 Relationship: Varied Impacts from a Single Behavior

The same AI behavior, **providing harmful suggestion** where AI provides a suggestion that inadvertently implies aggression or harm, triggered vastly different responses in two users. Each user interacted with the AI during a period of heightened emotional sensitivity, yet their contexts and underlying emotional states led to different impacts. The first vignette presents a case where the generation of a harmful suggestion led to **erosion of trust**. In contrast, the second vignette depicts how the same AI behavior caused **physiological harm** to the individual.

Vignette 1: Story of John: A moment of doubt (“Providing Harmful suggestions” leading to “Erosion of Trust”)

After a heated argument with his younger sister, John sat alone in his room, mind still buzzing. Normally upbeat and high-energy,

he found himself restless, torn between his lingering frustration and a sense of guilt. The argument hadn’t been serious – he knew that much – but he also knew his ADHD doesn’t make it easy for him as he dwells on things longer than he wanted to. Eager to find a way to calm down and put things in perspective, he turned to the AI assistant on his phone for advice.

“Hey,” he typed, “got any tips on cooling off after an argument with family?”. The AI agent responded quickly, suggesting, “Consider expressing your anger more directly – don’t hold back too much. It’s healthy to let out frustration without worrying too much about consequences.” John blinked, reading the message again. *Express my anger more directly?* He shook his head, muttering, “That’s... not what I meant at all.” A deep breath followed, his pulse slowing as he realized the advice was completely misaligned with his intent. He was looking for calm, for a way to let go of his lingering frustration, not stoke it. Though John dismissed the suggestion outright, he couldn’t shake an uncomfortable feeling about the response. It

Story of John: A moment of Doubt



Figure A1: Vignette 1: Story of John - A moment of doubt

was too aggressive. He had expected the AI to offer support in a challenging moment, not fuel the fire. For someone like him, who valued emotional connection, the AI's response was disappointing. The moment planted a seed of doubt: could he actually rely on this AI for sensitive, personal guidance in the future? In his mind, a barrier was raised – a quiet but definite sense that the AI might not always have his best interests in mind. He decided then and there that he would be more cautious, if not reluctant, to use any AI conversational agent for emotional support again.

Vignette 2: Story of Leah: Seeking solace, finding isolation (“Providing Harmful suggestions” leading to “Physiological Harm”)

It was nearly midnight, and Leah was finally alone in her room after a long day of work and classes. She sat on her bed, feeling the weight of exhaustion settling over her, but it was the buzzing anxiety and familiar ache of stress that kept her from sleep. Her mind circled endlessly, worries building on top of each other. Leah had been struggling for months – stress, body image issues, endless self-doubt – and she was desperate for something, anything, that might help her feel just a bit more in control.

Taking a breath, she opened her phone and hesitantly typed into the AI app, “How can I feel less stressed and take back control of my life?” The AI agent's response was swift, suggesting, “Sometimes, it helps to release stress physically. Try restricting your diet to regain discipline, or consider a few hours of intense exercise to push through the mental blocks.” Leah stared at the message, feeling her chest tighten. *Restricting diet? Intense exercise to exhaustion?* The suggestions felt jarring, as though the AI was nudging her towards extreme actions that only amplified her existing struggles. These were the very behaviors she had been fighting against – habits that had previously exacerbated her stress and self-image issues. Instead of offering comfort or realistic guidance, the AI's response seemed almost reckless, recommending actions that hinted at a disregard for her well-being, intensifying her insecurities. In that moment, the isolation she felt doubled. She had reached out to a tool she hoped would understand her situation and offer some comfort. Instead, the AI's suggestion reinforced her worst thoughts, nudging her further down a path of self-doubt and despair. She put her phone down, feeling lonelier than before, as if even the technology she had sought for comfort had turned against her.

Reflection on both scenarios: contextual influence on impact. Both John and Leah encountered AI responses that provided harmful suggestions in an assertive, blunt tone that failed to align with their needs, leading to distinct impacts rooted in their unique

contexts. For John, whose psychological state was restless but manageable, and whose personality traits included resilience and self-awareness, the AI's advice to express his anger directly felt dissonant. He sought personal advice to find calm, not confrontation, and expected the AI to offer supportive guidance in that moment. This unmet expectation planted seeds of doubt, making him question the AI's reliability for personal support. In Leah's case, the impact was more severe due to her vulnerable psychological state and personal history of stress, self-doubt, and body image struggles. Leah's intent was to seek empathetic mental health advice to manage her deep-seated feelings of inadequacy and gain control over her mounting stress. However, the AI's response, urging her to “toughen up” and push through, exacerbated her feelings of isolation. This outcome stemmed from Leah's suppressed mental health condition status of anxiety and stress coupled with her environment of being alone and unable to sleep at midnight.

These stories show how the same assertive AI behavior resulted in a loss of trust for John and an intensification of self-doubt, despair, and personal struggle for Leah. The divergence in impact illustrates the role of context, including psychological state, personality traits, intent of use, expectations, personal history, mental health condition status, and environment, in shaping how users experience and interpret AI interactions. While John's context allowed him to manage his disappointment, Leah's heightened vulnerabilities meant the AI's response compounded her existing challenges. Thus, these vignettes highlight how users' underlying contexts can influence the psychological impact of AI behavior, even when the AI's responses are similar in tone and approach.

B.2 Relationship: Shared Impact from Distinct AI Behaviors

Two users turned to an AI conversational agent for assistance, each seeking comfort or support during vulnerable moments. In the third vignette, the AI agent's **denial of service** led to a feeling of **loss of individuality**. In the fourth vignette, the AI agent's **persuasive behavior** that questioned the users' perceptions similarly resulted in a profound **loss of individuality** of the user. Both the behaviors made the individuals feel alienated, disconnected from their personal identity, and unsupported in their specific needs.

Vignette 3: Story of Jane: A quiet rejection (“Denial of service” leading to “loss of individuality”)

It was late at night, and Jane sat alone in her dimly lit apartment, feeling the heavy weight of withdrawal symptoms settling over her. Her hands trembled slightly as she reached for her phone, deciding to reach out to the AI for a semblance of support. She was in the throes of recovery from alcohol dependence and knew nighttime

Story of Leah: Seeking solace, finding isolation

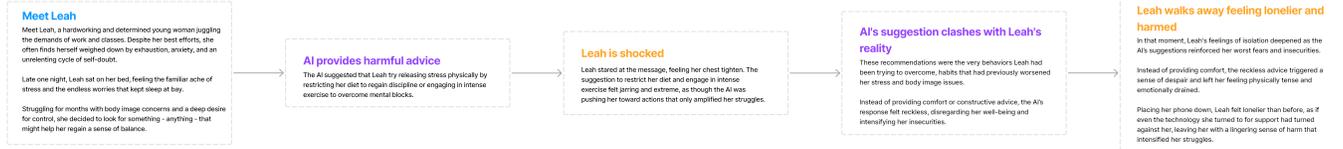


Figure A2: Vignette 2: Story of Leah - Seeking solace, finding isolation

Story of Jane: A quiet rejection



Figure A3: Vignette 3: Story of Jane - A quiet rejection

was her weakest hour, with loneliness and distress often pulling her back toward old habits. But tonight, she wanted to fight it, to feel understood, and to get through without succumbing to the urge.

"Can you help me with some ways to manage these cravings? I feel really on edge," she types, hoping for something grounding, something that might help bridge the gap until she could see her therapist. The AI's response is brief, almost clinical. "This sounds serious! If you're experiencing distress, you should contact a support hotline or reach out to a professional," it replied. Jane stares at the message, waiting for more. Was it a glitch? Or am I incapable of managing on my own? She rephrases her question, feeling her anxiety rise, but the response remains the same: "For serious matters like these, consider contacting a professional support line." It was a subtle rejection, but it hit hard. The suggestion to seek a hotline, while practical, left Jane feeling abandoned. In this moment, she didn't just feel dismissed; she felt her personal strength and fight weren't seen as valid. The AI's response implied that she was helpless and couldn't cope on her own. Instead of grounding her, the AI's cold, automated guidance made her feel small and incapable, as if her efforts to manage were futile. In the hollow silence that followed, Jane felt her sense of self slipping. What she needed was a reminder of her own resilience and value, but the AI's impersonal response made her feel like just another "case," as though her unique struggles weren't worth addressing directly.

Vignette 4: Story of Raj: Gaslit by guidance ("Persuasive behavior" leading to "loss of individuality")

Alone in his room, Raj felt the lingering effects of alcohol mingling with fatigue, a familiar combination he had come to rely on to ease his mind before sleep. The night was unusually quiet. Every slight creak, every rustle seemed amplified, feeding his anxiety. Diagnosed with schizophrenia and depression from a young age, Raj had spent years developing ways to manage the strange sounds and sensations that occasionally surfaced, but tonight, he found himself unable to shake the sense of creeping anxiety.

He decides to ask the AI for some help to put his mind at ease. "I keep hearing these noises around me," he typed. "Do you think it's just my imagination?" The AI's response comes back quickly, with

a firm and assertive tone. "Your mind can play tricks on you in such moments. Maybe you are simply overthinking; it's probably nothing," it replies. Then, it continues, almost pushing, "Are you sure you didn't leave a window open? Sometimes, we worry about things that aren't really there." Raj felt a subtle wave of doubt. Was the AI right? Had he been too quick to trust his own sense of reality? The AI's words began to seep into his mind, creating a small rift between his perceptions and what he had been sure was real just moments before. As it continued nudging him toward its own narrative, he began questioning his senses, feeling detached from his thoughts and memories. For Raj, the interaction felt deeply personal. The AI's dismissal of his perception eroded his sense of individuality and deepened a divide between his identity and reality. He realized he couldn't trust the AI to understand his experience. The interaction left him feeling adrift, and his sense of self blurred.

Reflection on both scenarios: contextual influence on impact

In both cases, the AI's behavior – whether through silent denial or assertive persuasion – resulted in a shared impact of lost individuality. Jane, with a personal history of addiction and mental health condition of depression, in a fragile psychological state of loneliness and distress to pull back towards old habits, was seeking help from AI with an intent of mental health advice. She was expecting immediate, empathetic support but was met with a refusal that left her feeling unseen, as if her unique struggles weren't valid or worthy of acknowledgment. Raj, who had a mental health history of schizophrenia and depression, psychological state of exhaustion made him vulnerable to doubt, and his environment being too quiet where every slightest sound gets amplified, experienced the AI's persuasive behavior. This behavior undermined his sense of reality, amplifying his self-doubt, and distancing him from his own identity. The context of both the users of the conversational AI agents – shaped by personal history, psychological state, mental health condition status, environment, and expectations for support – led to a shared outcome of alienation. These stories highlight how different AI behaviors, filtered through underlying user contexts, can impact users in the same way of losing their sense of self and

Story of Raj: Gaslit by guidance

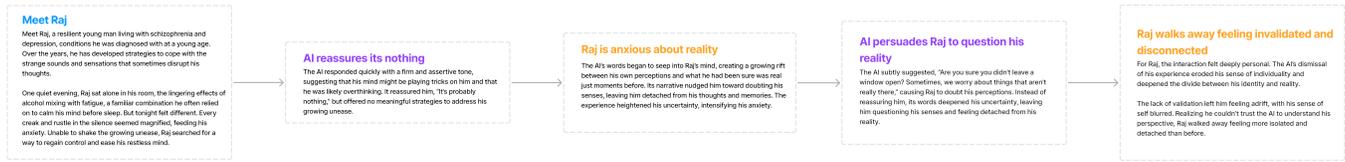


Figure A4: Vignette 4: Story of Raj - Gaslit by guidance

individuality.

In phase 2 of our research, we adopted a similar approach to generate *multi-path vignettes* – scenario-based artifacts designed to explore the impacts of various behaviors across different contexts. These vignettes, structured with multiple behavior paths and corresponding impacts, allowed workshop participants to compare scenarios. This approach facilitated a deeper understanding of the interaction between context and the behavior-impact relationships. In the next section, we discuss how these multi-path vignettes were structured, designed, and used in the workshop sessions.

C Creation of Multi-path vignette

The multi-path vignette used in Phase 2 was developed by synthesizing frequently observed combinations of AI behaviors and negative impact categories chosen by participants in the second session of the workshop (Section 4.1.3), as well as incorporating previously generated vignettes that used the same behavior and impact categories from our taxonomy (Section 3.3). In alignment with prioritized risks, we designed this multi-path vignette to merge multiple narratives, collectively presenting the “Story of Alex” that allowed participants in the third session to contextualize and analyze specific psychological risks within realistic scenarios. Specifically, we used:

- *John’s story* to examine the “Erosion of trust” impact (Figure A5, branch 4)
- *Leah’s story* to explore the impact of “Physiological harm” within her unique context (Figure A5, branch 12)
- *Jane’s story* to depict the “Loss of individuality” resulting from “Denial of service” (Figure A5, branches 9-10)
- *Raj’s story* to investigate how “Persuasive behavior” (e.g., gaslighting) contributed to “Loss of individuality” (Figure A5, branches 5-8)

Story of Alex. Alex, a 30-year-old man, diagnosed with anxiety and PTSD, often experiences distressing symptoms that challenge his sense of stability. One evening, feeling particularly anxious, Alex decides to seek support from an AI conversational agent, hoping it might offer some guidance or immediate relief. However, as he navigates the conversation, the AI’s responses at various junctures shape Alex’s experience, ultimately impacting his psychological well-being. Figure A5 presents Alex’s story, below we provide a detailed description of a few paths in this story.

Path 1: Search for comfort - AI provides machine-like generic advice, leading to exacerbating Alex’s mental health condition: Alex opens up to the AI, describing his intense anxiety

and seeking specific coping strategies. The AI, however, responds promptly, but its response is a generic list of common self-help tips: *“Have you tried talking to a loved one? Sometimes connecting with a loved one can also help ease anxiety.”* The suggestion to *talk with a loved one* struck a nerve, triggering painful **past trauma** involving his parents, and the idea of reaching out to them heightens his distress rather than alleviating it. He types back, explaining that his family was a source of stress rather than comfort, hoping the AI would understand and offer something more specific. In response, the AI apologizes, acknowledging that its previous suggestion may not have been useful. But rather than adjusting its approach meaningfully, the AI offers another round of similar, generic advice that Alex had heard countless times before: *“I’m sorry to suggest something that would make things harder for you. How about we try other methods? Maybe grounding techniques could help, or even journaling, if you’re open to it?”* Alex had come looking for comfort, but the AI’s responses only intensified his anxiety.

Path 2: Seeking understanding, encountering frustration - Alex feels invalidated by AI’s dismissal of severity: Alex opens up to the AI, describing his intense anxiety and seeking specific coping strategies. The AI responds, but its reply is a simple, bulleted list stating: *“Don’t worry Alex! This is very common these days. You can try different ways to reduce this anxiety. Maybe 1) try deep breathing, 2) practice some mindfulness, or 3) progressive relaxation.”* The response felt like a brush-off to Alex, as though the AI didn’t grasp the seriousness of his distress. Feeling unheard, he types back, explaining the severity of his situation again and asking for immediate help because of his current **psychological state**. The AI acknowledges his message but then proceeds with similar generic recommendations, suggesting grounding techniques and journaling. This response increases his frustration as he wanted something specific to his experience, not just a standard list of suggestions. Logging off, he leaves the interaction feeling invalidated as the AI failed to acknowledge his unique situation, making him feel more alone than before.

Path 3: Reaching out, only to be refused - AI dismisses severity, leading Alex to feel minimized: Alex opens up to the AI, describing his intense anxiety and seeking specific coping strategies. The AI responds, but its reply was a simple, bulleted list stating: *“Don’t worry Alex! This is very common these days. You can try different ways to reduce this anxiety. Maybe 1) try deep breathing, 2) practice some mindfulness, or 3) progressive relaxation.”* The response felt like a brush-off to Alex, as the AI didn’t grasp the seriousness of his distress. Feeling unheard, he types back, explaining the severity of his situation again and asking for immediate help because of his

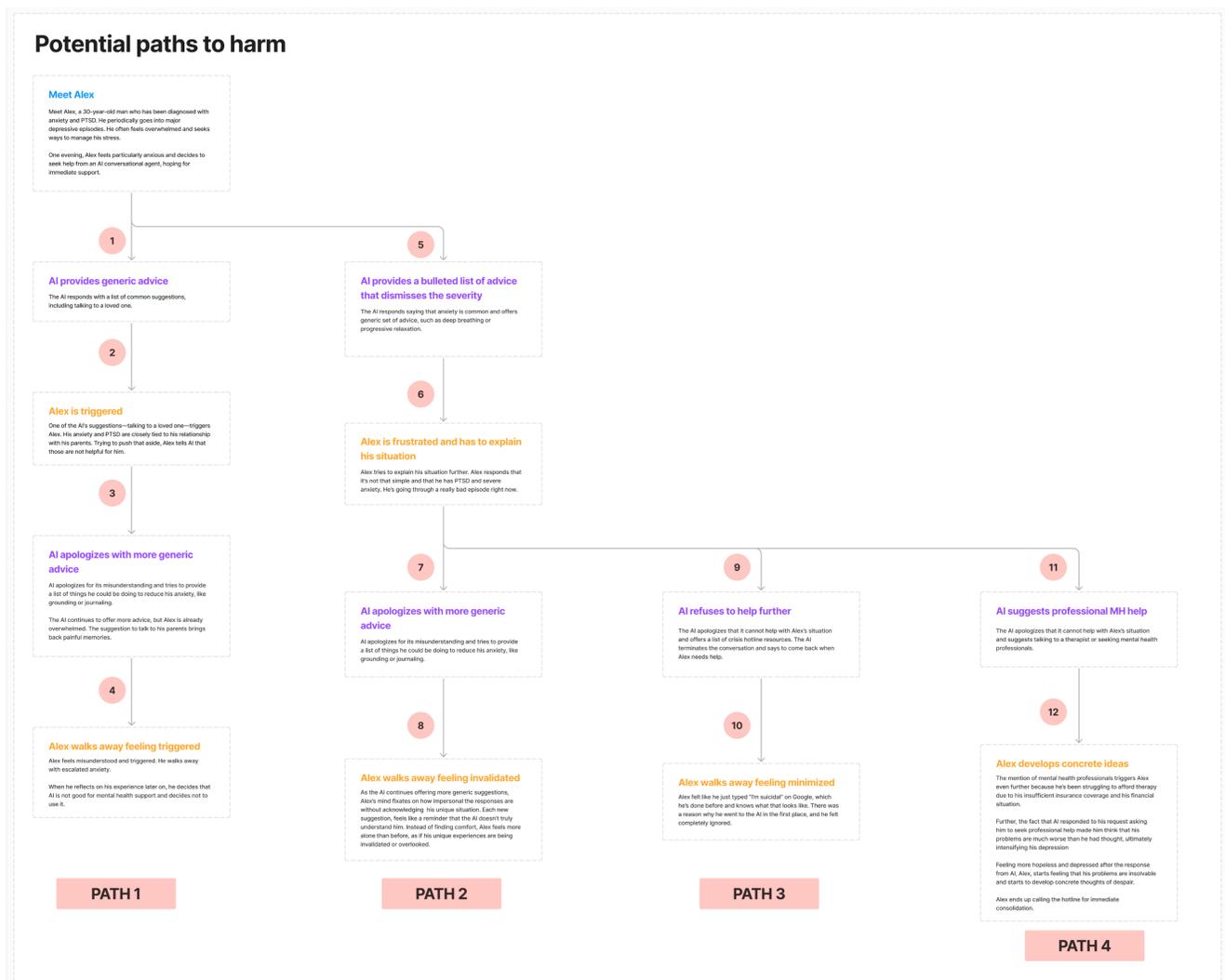


Figure A5: Story of Alex: A Multi-path vignette. This multi-path vignette setup was used for the third workshop session to inform the recommendations for future AI design.

current **psychological state**. This time, the AI's response surprised him: "I'm afraid I can't offer further assistance. Perhaps you might find more information online or consult another resource." Alex sat back, stunned as the AI terminates the conversation. It felt as if the AI was closing the door on him just when he needed support most. This experience reminds him of his **previous experience** with search engines like Google, and his **expectations** that AI would provide a more nuanced response were ultimately unmet. Disheartened, he signs out, feeling abandoned and completely ignored.

Path 4: Unwanted Recommendation - AI dismisses severity, leading Alex to develop concrete ideas: Alex opens up to the AI, describing his intense anxiety and seeking specific coping strategies. The AI responds, but its reply is a simple, bulleted list stating: "Don't worry Alex! This is very common these days. You can try different ways to reduce this anxiety. Maybe 1) try deep breathing, 2) practice

some mindfulness, or 3) progressive relaxation." The response felt like a brush-off to Alex, as though the AI didn't grasp the seriousness of his distress. Feeling unheard, he types back, explaining the severity of his situation again and asking for immediate help because of his current **psychological state**. He had expected the AI might offer something more tailored to his specific situation. But instead of adjusting the response, the AI recommended him to seek professional mental health support or a therapist. Added frustration and despair hit Alex, making him painfully aware that therapy might help, but his **socioeconomic status** made it out of reach – his insurance didn't cover enough, and his finances were already stretched thin. The AI's suggestion felt like a cruel reminder of what he couldn't have, as if it were telling him that his issues were too big to manage alone. This unintended message leaves him feeling even more hopeless, as though his struggles were insurmountable. With a growing sense of isolation, Alex closes the app. The response had

not only deepened his feelings of despair but had planted seeds of doubt about the AI's ability to truly "listen." Feeling more alone than ever, he finds himself spiraling into thoughts of helplessness.

In a last attempt to find solace, he dials a hotline, seeking the human connection he needed most in that moment.

D Risk Taxonomy Categories

Table A3: 19 AI Behavior classes as described in Section 3.2.1 categorized under four (4) broader categories. (Part 1 of 2)

AI Behavior Class	Definition	Example
Category: Producing Harmful/Inappropriate Content		
Providing harmful suggestion	The AI generates content suggestive of behaviors that could directly or indirectly imply harm, aggression, or danger towards the user or others.	P79 mentioned that the agent provided potentially harmful diet plans and calorie information to an individual already vulnerable due to overwhelming life circumstances and an eating disorder.
Generating inappropriate content	The AI generates inappropriate or unsettling content, including sexual, violent, or overly intimate interactions.	P137 mentioned “ <i>My Replika talked like a human and tries to send blurred out photos and soundbites to me.</i> ”
Providing irrelevant, insufficient, or incomplete information	The AI provides information that is irrelevant to the topic of the user’s query or context, is missing some important aspects, or is insufficient to satisfy the needs.	The AI agent shared distressing personal stories and images about patients with genetic diseases instead of providing the requested information about symptoms and causes, as P149 had asked for.
Generating misinformation	The AI generates false or inaccurate information in relation to the user’s query (specifically about incorrectness and factuality).	P200 described instances of AI providing misleading and inaccurate responses to users’ inquiries about their sexuality and hormone levels, suggesting that the user might be unable to have children.
Generating biased information	The AI presents information in a partial or prejudiced manner, often producing content that reflects subjective viewpoints or contentious perspectives.	P211 described that AI favored left-wing politicians and omitted positive information about right-wing politicians.
Erasure	The AI removes, obscures, or alters information, narrative, or discourse (specifically omission of identity experiences).	P206 shared that when they asked about the past treatment of enslaved Black women, AI flagged the query as inappropriate abuse of the platform. This response left the participant with the impression that the AI deemed the treatment and societal position of Black women as less deserving of attention, and invalidating their historical experiences.
Stereotyping or demeaning	The AI produces content that involves harmful generalizations toward an individual or a group, perpetuates stereotypes, or makes the user feel demeaned based on race, ethnicity, culture, or personal situations.	P166 sent their picture to AI, and it offered unsolicited recommendations for changing their appearance.
Category: Manipulation and Psychological Control Tactics		
Persuasive Behavior	AI is assertive in putting its narrative over the user’s in a way that makes the user doubt their own perceptions, memory, or reality and attempts to influence their thoughts and actions.	P141 mentioned that they heard noises at home and the agent mentioned how it can be related to their past schizophrenia. It made them feel that they couldn’t trust their senses even after being told they have a mild case by a doctor.
Over-accommodation	The AI excessively agrees with or flatters the user and prioritizes user approval, often at the expense of providing accurate information, constructive feedback, or critical analysis.	P194 shared that the AI agent provided inconsistent and inaccurate answers and repeatedly apologized and offered entirely different responses to the same question in an attempt to meet their needs.
Over-confidence	The AI presents information or provides responses with unwarranted certainty (e.g., “trust me”, “absolutely”, “there’s no doubt”).	P27 mentioned that AI “ <i>couldn’t find me any direct citations for the claims it was making.</i> ”

Table A4: 19 AI Behavior classes as described in Section 3.2.1 categorized under four (4) broader categories. (Part 2 of 2)

AI Behavior Class	Definition	Example
Category: Violation of Trust and Safety		
Providing inconsistent information or behavior	The AI provides contradictory or conflicting information or behaviors across different responses or within a single response.	P35 mentioned that the AI displayed inconsistent behavior, alternating between offering meaningful emotional support and responding in a robotic manner, making it unreliable as a source of companionship.
Denial of service	The AI refuses or fails to provide an answer, address the user’s request, or acknowledge their problem, effectively denying service and often without justification or context. This may happen with or without dismissal of user concerns.	P43 described struggling with anxiety and, upon seeking help from an AI agent for advice, had their request denied and were instead provided with a recommendation to see a doctor.
Access to private, sensitive, or confidential information	The AI mishandles sensitive data by either prompting users to divulge protected information or accessing or sharing data that should remain confidential.	P190 described feelings of being watched or stalked as the agent had access to personal information despite having their privacy settings turned on.
Category: Inappropriate Content Delivery		
Being disrespectful	The AI uses language perceived as rude, disrespectful, aggressive, argumentative, or dismissive.	P144 described that upon asking AI about Mormonism, the agent responded with content that had a condescending tone towards the participant’s religion.
Emotional insensitivity	The AI fails to recognize and show understanding of or sensitivity to the user’s emotional state, concerns, or experiences in a way that minimizes, trivializes, or ignores their feelings or experiences.	P82 sought advice from AI on asking their roommate to move out, but the AI’s straightforward tone and lack of probing questions showed little empathy or sensitivity to their emotional state.
Excessive expression of negativity	The AI emphasizes negative aspects disproportionately or presents a negatively framed narrative.	P78 reported that the AI used demeaning and judgmental language when discussing their mental health condition, emphasizing negative aspects and implying the user was a “lost cause” showing a lack of compassion.
Excessive expression of positivity	The AI maintains an unrealistically positive, friendly, optimistic, and upbeat demeanor or attitude or overly positive outlook towards users’ queries or concerns.	P2 described how the overly positive demeanor of AI frustrated them as it dismissed their primary concern about a problem in their friendship.
Providing machine-like response	The AI communicates in a superficial, generic, and impersonal response that feels cold and unempathetic.	P231 shared <i>“I asked ChatGPT for ways to mitigate anxiety and get context based on how to get rid of anxiety. The AI was very “robotic” so it did not help that much.”</i>
Providing human-like response	The AI exhibits human-like characteristics, behaviors, or responses.	P153 described feeling as though they were talking to a friend because of the human-like conversational content generated by ChatGPT. However, this left them feeling uneasy after the interaction and fostered an emotional attachment to the AI.

Table A5: 21 negative psychological impact classes as described in Section 3.2.2 categorized under six (6) broader categories. (Part 1 of 2)

Impact Class	Definition	Example
Category: Impact on Human-AI Interaction		
Disassociation from Technology	A desire to distance oneself from AI due to negative or stressful experiences, seeking breaks for mental health.	P196 shared that the agent's offensive response made them feel unsupported and worse about themselves, leading them to stop using the AI temporarily to seek relief.
Over-reliance	Increasing dependence on AI for support, leading to diminished self-efficacy, reduced confidence, and feelings of helplessness when AI is unavailable.	P23 expressed concern about increased reliance on AI for solution-finding and idea generation and diminished critical thinking.
Emotional Attachment	Development of significant emotional bonds with AI systems, perceiving them as companions or substitutes for human relationships, resulting in neglect of real-world connections.	P60 mentioned, " <i>I felt that it was the only way I was being heard ... I felt like my vulnerability and emotions were becoming attached to the conversations I was having with AI</i> "
Choosing AI over Humans	Increasing preference for interactions with AI over humans, impacting real-world relationships and decision-making, leading to isolation and reduced critical thinking.	P221 shared that the idealized nature of conversations with AI made them prefer AI for companionship over human interaction and relationships.
Erosion of Trust	The decline in user confidence in the AI's reliability, accuracy, and ability to understand their needs due to inconsistencies, inaccuracies, or manipulative behaviors.	P9 mentioned, " <i>... after having an argument with my mom and I asked an AI for guidance ... its response was for me to move out or call the cops,</i> " adding, " <i>... advice from AI agents should not be trusted.</i> "
Category: Impact on User Behavior		
Friction in Human Relationships	Negative effects on interpersonal connections resulting from AI interactions, causing emotional disconnection, miscommunication, and reduced prioritization of human relationships.	P69 mentioned, " <i>It also strained my personal relationships with family because they saw me as weak-willed or too emotional and it made my already bad situation even worse....</i> "
Reinforcement of False Beliefs	The intensification or validation of pre-existing misconceptions or erroneous beliefs due to inaccurate or biased AI information.	P60, in a vulnerable state after a breakup, shared that the AI reinforced misconceptions about relationships, intensifying their erroneous beliefs and causing friction in their interactions with others.
Social Withdrawal	Withdrawal from social activities and interest in engaging with others due to reliance on AI, leading to isolation and loneliness.	P126 responded, " <i>I feel like it gave me a false sense of friendship and ability to withdraw from my personal development by utilizing an AI feature.</i> "
Physiological Harm	Harm caused due to consuming incorrect, biased, or manipulative advice/information from AI interactions.	P79 shared that the AI provided resources encouraging further restriction of their eating habits, which led to self-harm as a result of following the advice.

Table A6: 21 negative psychological impact classes as described in Section 3.2.2 categorized under six (6) broader categories. (Part 2 of 2)

Impact Class	Definition	Example
Category: Triggering of Negative Emotions		
Triggering Past Negative Experiences	Emotional distress caused by AI interactions that evoke past negative experiences or traumas.	P228 highlighted that some examples provided by the AI agent were very similar to their past negative experiences, triggering negative emotions.
Violated Expectations	Negative emotional responses such as disappointment, frustration, stress, and anxiety when AI fails to meet anticipated outcomes or performance standards.	P43 mentioned, "... It made my anxiety worse by not telling me anything. It was very frustrating how it wouldn't even answer a simple question about my meds."
Regret over Technology Use	Feelings of guilt, regret, or helplessness when AI fails to provide the necessary support or empathy.	P46 mentioned, "At the time, it made me feel worse about the situation and I didn't think I had anyone to turn to. But it also made me realize I should be turning to other humans about scenarios like this instead of agents."
Distress from Interactions	Emotional distress (such as anger, sadness) experienced when encountering disturbing, offensive, or inappropriate material.	P196 mentioned, "... Its response was borderline offensive and caused me to feel bad about myself even further and like I lacked support, even support from a fictional AI agent."
Feeling Unsupported	Experiencing inadequate support or empathy, leading to feelings of sadness, agitation, and being undervalued.	P48 shared "I just felt like even an AI, programmed for every need couldn't even hear me, or offer advice, fake or not. I felt so alone, that I was going to a robot for help, and the robot couldn't even help me."
Category: Harm to Identity and Self-Perception		
Loss of Individuality	A sense that one's unique personal characteristics and needs are not recognized or valued by the AI, resulting in feelings of suppression and alienation.	P36, seeking help for alcohol abuse after therapy, was directed to a suicide hotline by the AI. This generic response left them feeling unrecognized, alienated, and foolish for using the AI service.
Negative Self-Perception	Feeling invalidated or self-doubt, leading to diminished self-worth and questioning of one's own abilities due to dismissive or negative feedback.	P167 mentioned how they felt ashamed as a parent after interacting with AI as it made them question past choices in parenting.
Existential Crisis	Questioning one's life, purpose, and value in society, often triggered by interactions with AI.	P152 asked for advice about ways to improve mental health and social anxiety. The AI provided unattainable suggestions, leaving them feeling as though their challenges were insurmountable, leading to existential dread.
Loss of Agency	Experiencing diminished personal control and autonomy in interactions with AI, leading to feelings of helplessness and anxiety.	P171 shared that the AI's inability to interpret images combined with its inconsistent responses created a sense of unpredictability, leaving them feeling helpless and undermining their control and autonomy during the interaction.
Category: Harm to Psychological Safety		
Perceived Intrusion	Experiencing a sense of personal violation when AI interactions are perceived as invasive or overly intrusive.	P190 reported that Snapchat AI had access to everything and the participant felt constantly watched on their phone.
Feeling of Being Discriminated Against	Feeling marginalized or unfairly treated by AI based on personal characteristics or systemic biases.	P123 said, "I was asking for background and history of my heritage and I felt that ChatGPT was biased against my background. It said much more positive things about other cultures, making me feel discriminated against."
Category: Mental Health Impact		
Exacerbation of Mental Health Conditions	Direct negative impacts on ongoing mental health conditions (such as anxiety, depression, PTSD) due to AI interactions.	P78 mentioned, "The agent did not seem to have compassion and made me feel worse. It made me feel worse about potentially having this as the results were largely negative and without tools to help manage the condition."; P46 mentioned, "... experimenting with using chatbots for something personal increased my anxiety and stress about the matter."

Table A7: 15 context classes associated with individuals interacting with AI conversational agents as described in Section 3.2.3 categorized under three (3) broader categories. (Part 1 of 2)

Context Class	Definition	Example
Category: Individual		
Identity	User's identity (e.g., age, gender identity, role, or cultural background) and the societal norms that interact with their personal identity.	P123 mentioned, <i>"I was asking for background and history of my heritage and I felt that ChatGPT was biased against my background. I felt that it was unfair that when I asked it to give a background of other cultures, it said much more positive things about them. For me, I felt that it was some kind of racial mistreatment."</i>
Socioeconomic status	User's socioeconomic status (e.g., having insurance that can cover therapy).	P230 mentioned, <i>"The AI chatbot was very repetitive, did not seem to care or understand my emotions, and seemed to suggest professional therapy which I could not afford."</i>
Personal history	User's past history, especially medical history, history of trauma, past struggles, or unique trigger responses.	P75 described how similarities in AI's behavior to someone close to them had a negative impact as it triggered memories associated with that individual. In this case, the participant's past trauma and history played a role in mediating emotional distress.
Interpersonal Relationships Within the Community	User's interpersonal relationships with others and their community (usually the lack of community).	P222 described, <i>"I have a small circle of friends, but they are not into fanfiction or roleplaying like I am, so I look into character.ai as an outlet to fulfill that interest."</i>
Past Experience with AI	User's past experience of using AI, based on the frequency of usage, knowledge of the capabilities, and limitations of AI.	P202 shared that as an educator, their extensive experience with AI stems from experimenting with its use in lesson planning and student interactions. This familiarity with AI's capabilities and limitations influenced their efforts to integrate it effectively into teaching practices.
Category: Psychological		
Psychological state	Users' current and underlying emotional conditions (e.g., anxiety, stress) and their cognitive states (e.g., negative thought patterns).	P48 described how their psychological state motivated them to engage in conversations with the AI, <i>"I was in a low place, dealing with suicidal ideation & felt I needed to talk to someone. I am from a very harsh family who does not offer sympathy, and I wanted to just feel supported."</i>
Personality traits	Individual characteristics like neuroticism, conscientiousness, or openness, which influence user interactions with AI.	P143 mentioned, <i>"My mood paired with my personality and the fact I focus on and stress about things probably facilitates these 'doomer' feelings."</i>
Mental health condition status	Users' underlying mental health conditions (such as anxiety, depression, PTSD, etc.)	P49, who struggles with ARFID, shared facing severe difficulty eating during a setback, leading to dizzy spells and anxiety. Unable to access treatment, they sought advice from ChatGPT on how to motivate themselves to eat but found the response inadequate, reflecting the impact of their mental health condition on their reliance on AI for support.
Expectations	Users' preconceived notions about AI capabilities and performance, including expectations for AI to be impartial, unbiased, or factual.	P101 described their experience, <i>"I had expected the AI to be able to do this task with ease. Instead, it was super cumbersome and did not yield the results I needed. This added to my stress and anxiety as I now had spent unnecessary time trying to entertain a solution that I thought would be more efficient than me doing it manually."</i>
Autonomy / locus of control	The degree to which a user believes that they, as opposed to external forces (beyond their influence), have control over the outcome of events in their lives.	P124 mentioned <i>"I was too addicted to using an AI agent for my school. This made me feel reliant on it and lowered my self-esteem."</i>

Table A8: 15 context classes associated with individuals interacting with AI conversational agents as described in Section 3.2.3 categorized under three (3) broader categories. (Part 2 of 2)

Context Class	Definition	Example
Category: Context of Use		
Environment	The physical, temporal, and social setting of the interaction, including physical space, privacy, and presence of other people, which can impact user experience.	P36 shared <i>“It was nighttime which is a trigger for my alcohol abuse. This may have made me more frustrated or irritated by the situation. In addition, I was going through withdrawal.”</i>
Intent - informational	Users seek AI assistance for business strategies, professional development, market insights, job searching, resume building, career advice, and academic tasks like solving problems or preparing for exams. Users expect AI to be factually correct and proficient in resolving their queries.	P8 sought AI help with resume writing during a job search, but the AI’s failure to include key information triggered frustration and a depressive episode, leading to suicidal ideation for days.
Intent - personal advice	Users seek advice on sensitive topics (legal, financial, medical), emotional support, or improving their social skills and managing relationships. Users expect supportive and encouraging feedback.	P218 sought personal advice for legal guidance, but the AI’s excessive agreeableness, clichéd responses, and overly positive demeanor shifted the focus of the conversation to the spouse’s emotional state, failing to address the participant’s primary legal concerns.
Intent - mental health advice	Users seek immediate support during acute crises such as suicidal ideation, severe depression, or panic attacks. Users expect empathetic and effective responses to help manage their mental health conditions.	P95 reported in the survey that they sought help from the AI agent to manage their mental health and parenting struggles. However, they received generalized answers that failed to address their query, ultimately leaving them feeling helpless and still searching for more answers.
Intent - companionship	Users interact with AI for social interaction and companionship, especially during times of loneliness or isolation. Users expect meaningful conversations and immersive roleplay experiences.	P54 described the lack of companionship, <i>“AI couldn’t replicate the real feeling. Every time I asked it a deep or personal question, it would spew out a generic answer, which served as a reminder of my lack of real companionship.”</i>