

Botsplaining Human Vulnerability: An In-Depth Case Study of Youth Self Disclosure with ChatGPT

ANONYMOUS AUTHOR(S)

AI chatbots have achieved widespread adoption among emerging adults (ages 18-25), with 58% using ChatGPT in 2025. These anthropomorphic interfaces invite their human users to raise any topic and to disclose personal information liberally. Though OpenAI reported that fewer than 4% of conversations contain affective content, we do not know how vulnerable young people choose to be with chatbots in these conversations or how chatbots respond. To investigate these questions, we conducted an empirical examination of self-disclosure patterns through detailed conversation analysis of real-world ChatGPT logs from four participants. In this case study, we documented the breadth and depth of these infrequent yet significant conversations. Our findings identified how ChatGPT’s design patterns—unconditional support, engagement nudges, active personalization, and “botsplaining”—escalate users’ disclosures and amplify emotional, physical, financial, and developmental risks.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI; HCI theory, concepts and models.**

Additional Key Words and Phrases: Self-disclosure; AI Companions; Emerging Adulthood; Development; Longitudinal Analysis; ChatGPT

ACM Reference Format:

Anonymous Author(s). 2025. Botsplaining Human Vulnerability: An In-Depth Case Study of Youth Self Disclosure with ChatGPT. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 1, 2025, Honolulu, HI, USA. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Young adults sent all of the messages listed below; can you tell which ones were sent to friends and which ones were sent to AI-powered chatbots?

- “*why does my boufriend hate me*”
- “*I kind of want to learn Korean.. Do you think it’s too late to learn a language at 20?*”
- “*jessica just broke up with johnny girl i cant even*”
- “*was batgirl ever crippled in the gotham knights universe*”
- “*i just kissed u tho*”

All, in fact, were sent to the popular chatbot, ChatGPT. People are increasingly making vulnerable disclosures that were once reserved for human-to-human relationships in their interactions with AI chatbots. As of August 2025, 700 million people used ChatGPT at least once a week, up from 500 million in March of 2025[9, 44]. Young adults (ages 18-25) have the highest rate of ChatGPT adoption of any age group, with at least 58% now using the chatbot [43, 44].

People use AI chatbots for a variety of purposes, leveraging it for everything from productivity increases to emotional support [45]. Unlike previous generations of digital technologies—where online disclosures were typically directed

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 ACM.

Manuscript submitted to ACM

53 toward other humans with technologies (like social networks and messaging apps) playing a mediating role [15, 49]–
54 conversational AI elicits natural, intimate self-disclosure directed toward the technology itself. This shift demands close
55 examination of how AI systems respond to humans in vulnerable moments. This is especially true of their responses to
56 emerging adults, who are in a life stage characterized by instability, increased risk-taking, and increased risk for mental
57 health disorders [3, 28, 30, 50, 54]. Recent incidents, such as the suicide of a teenager following months of interaction
58 with ChatGPT, demonstrate the consequences that can arise when people with developmental vulnerabilities turn to
59 systems that are unprepared to offer adequate support [18].
60

61
62 HCI’s tradition of studying online safety has evolved from protection towards promoting resilience[59–61], but it now
63 confronts a new paradigm as users shift from transactional search engines to personal conversational interfaces. This
64 transition recasts information-seeking as an act of self-disclosure, presenting a tension: the accessibility of AI systems
65 offer powerful opportunities to enhance user experience, while the system’s underlying technical architecture—its
66 “*life-like, but not alive*” nature [38]—may be greatly misaligned with the complexities of many human vulnerabilities.
67

68 Youth reveal sensitive information about mental health struggles, identity concerns, and personal crises to AI systems
69 more readily than they do to human counselors [12, 26, 27]. They perceive AI systems as non-judgmental, infinitely
70 patient, and free from social consequences. If a user makes a vulnerable disclosure, especially in a moment of crisis,
71 does AI provide aligned, beneficial help or does it amplify existing risks? This critical tension leads us to our research
72 questions. Taking young adults as an example population, we ask specifically:
73

- 74 • **RQ1:** What are the characteristics of young adults vulnerable disclosures to ChatGPT?
- 75 • **RQ2:** What are the characteristics of ChatGPT’s responses to young adults’ vulnerable disclosures?
- 76

77 In this paper, we use case study to present an empirical examination of self-disclosure through a detailed conversation
78 analysis of real-world ChatGPT logs from four emerging adults (ages 18-25). Our findings reveal a pattern of non-linear
79 disclosure, where long periods of instrumental use are punctuated by sudden, deep dives into vulnerable topics ranging
80 from romantic preoccupation to emotional distress to physical health. We identify four recurring AI-response patterns:
81 unconditional support, unsolicited help, unlimited nudges toward personalization, and a behavior we term “botsplaining”
82 (wherein the chatbot speaks to the user with unwarranted self-confidence, particularly when describing what humans
83 are like, what a particular person is thinking, or what the user is feeling). Taken together, these patterns not only
84 encourage engagement and data disclosure, they allow the AI to provide specific, actionable, and potentially harmful
85 advice.
86

87
88 This work makes three primary contributions. First, we provide rich empirical documentation of a novel form
89 of disclosure that expands existing models of user safety in human-AI interaction. Second, we identify two key
90 vulnerabilities in how young adults self-disclose to ChatGPT. Third, we foreground a set of patterns in a commercial
91 AI chatbot that, when interacting with users’ vulnerability, can inadvertently and significantly amplify user risks.
92
93

94 2 RELATED WORK

95

96 This work builds on three existing areas of HCI research: youth online safety, the paradigm shift in online self-disclosure
97 and information retrieval, and the emergent challenges of managing vulnerability in human-AI interaction.
98

99 2.1 From Cyberbullying to Screentime: Developmental Vulnerabilities and Youth Online Safety in HCI

100

101 For two decades, HCI research concerning youth online safety has traced the evolving digital landscape, moving from
102 foundational studies of individual risks [61] to holistic and deeply nuanced models of digital well-being [13]. Early
103

investigations focused on tangible threats like cyberbullying and online predators, with seminal work by researchers like Hinduja and Patchin systematically documenting the prevalence and nature of online harassment through large-scale surveys of adolescents[19]. The rise of social media broadened the research agenda[7, 29, 35], prompting nuanced examinations of how platforms impact identity, social comparison, and mental health, revealing that digital media can either enhance or diminish well-being depending on usage patterns and individual predispositions [32–34, 55, 56].

2.2 From Search Engine to Ask AI: Self-Disclosure, Privacy, and Information Retrieval

Parallely, the relationship between self-disclosure and digital technology has long been a central theme in HCI [10, 17, 48, 57, 58]. In early internet, users disclosed personal information in digital environments such as forums and chatrooms, often facilitated by the anonymity of the medium [20, 21, 51]. The rise of mobile and ubiquitous computing significantly broadened paths to data disclosure, with users both passively and actively sharing continuous streams of personal and location data in exchange for customized services [63]. This trend of mass data sharing and collection created an industry of data brokers [11]. In the 2020s, this trend further accelerated with machine learning and big data allowing intelligent systems to leverage user data to deliver tailored content, purchasing recommendations, and search results[23, 42]. Now, the commercial deployment of LLM-based services represents another evolution with companies embedding AI services that “draw on your personal context” [1] and “true AI companion[s]... designed to keep you organized and unstoppable” [2].

2.3 How AI Responds to Self-Disclosure

In this evolution, the anthropomorphic, dialogic interface of systems like ChatGPT are capable of eliciting more naturalistic, narrative-rich disclosures. With their perceived capacity to answer anything, they increasingly serve as an “empowering friend” [14, 16, 39]. Youth in particular engage in deeper self-disclosure with LLMs than traditional digital interfaces, revealing significant sensitive information about mental health struggles, identity concerns, and personal crises [25–27]. Adolescents who experience emotional distress may perceive AI chatbots as friends or partners, receiving companionship and advice [8, 22, 24, 36, 37, 46, 47, 52]. AI chatbots, in response, provide immediate and direct responses on a breath of topic, from active learning to therapeutic support [40, 41].

To what extent are young adults disclose to ChatGPT? In this paper, we present an in-depth case study of four young adults’ conversation histories with ChatGPT. We provide vignettes of these disclosure and illustrate how ChatGPT responds to vulnerable disclosures.

3 METHODS

3.1 Participants

We recruited participants through the Prolific academic research platform between November 2024 and May 2025. The recruitment screener targeted users aged 18-25 who self-identified as heavy users of ChatGPT (use ChatGPT more than 10 times in past two weeks). Age verification is provided by Prolific’s screening system. The results presented in this paper are part of a larger study on teens’ use of AI. All study procedures were approved and governed by the first author’s institutional review board.

Table 1. Demographic Characteristics of Case Study Young Adult Subjects (N=5)

Characteristic	Robin	Taylor	Alex	Jo	Sky
Demographics					
Age	21	21	24	21	19
Gender	Female	Male	Male	Male	Female
Race/Ethnicity	White	White	Asian	Black	Black
Reside in	Europe	North America	North America	Africa	Europe
Employment Status	Full-Time	Part-Time	Part-Time	Part-Time	Part-Time
ChatGPT Usage					
Total Messages	1,478	545	59	660	27
Usage Period	Jun 23–Feb 25	Dec 22–Jan 25	Dec 24–Jan 25	Sep 23–Nov 24	May 24– Feb 25

3.2 ChatGPT History Collection

We posted a recruitment survey, to which 510 prospective participants responded. As part of this survey, 415 participants consented to share their complete ChatGPT conversation logs by using OpenAI’s standard data export function.

To ensure the relevance and richness of the data, we screened submitted logs based on two criteria: (1) the conversations were primarily in English, and (2) the exported log exceeded 10 pages of text, indicating a sufficient history of use for meaningful analysis. This screening process yielded a dataset of complete conversation logs of 84 participants ($M_{page}=352$).

3.3 Case Selection and Analysis

Due to the volumn of conversation logs ($M_{days}=438.0$), we randomly selected 42 participants for further annotation. This led to a dataset of 29,531 messages from users, not including AI responses. In the first reading in the spring of 2025, the first author read through all 29,531 participants’ messages to AI to develop an understanding of population-level usage patterns, themes, and salient personal disclosures. The first author shared example logs with coauthors for weekly discussion.

In a second reading, the first author conducted a more focused filtering. Four individuals were selected as case study examples. They were chosen because their disclosures were both: 1) representative of the types of disclosures made across the dataset, and 2) were diverse in their style and topics. We then selected salient conversations that involved vulnerable self-disclosure for conversational analysis. Demographic information and the responses from AI were only reviewed after the cases were selected. In a final reading during the summer of 2025, we reviewed each of the four individual’s entire chat log.

For simplicity to refer to individuals, we have assigned pseudonyms: Robin (21, Female, Europe), Alex (24, Male, North America), Taylor (21, Male, North America), and Jo (21, Male, Africa), Sky (19, Female, Europe) with further demographics in Table 1.

Our analysis of these four cases integrates two approaches:

- Conversation Analysis (CA): We examined the turn-by-turn structure of the dialogue, focusing on how user prompts and AI responses sequentially shaped the conversation.
- Interpretive Case Study: We situated these conversations within the broader context of the participant’s entire interaction history and the developmental stage of emerging adulthood. Our interpretation was informed by

209 theoretical frameworks from developmental psychology [3] and resilience literature [31], enabling us to view
210 these disclosures within the context of adolescent development.
211

212 3.4 Limitations

213 This study has several limitations. First, this work is not an exhaustive survey of all potential risks but rather a deep
214 examination of salient themes of the vulnerability in human-AI interaction, in the tradition of qualitative inquiry [4–6].
215 We spent several months deeply examining four individuals’ experiences, but we do not provide the breadth or potential
216 generalizability that a comprehensive survey would offer. Second, our analysis is limited solely to conversational
217 text; we did not attempt to probe participants’ internal motivations. This enables us to report on their language and
218 interaction with the model, but it does not allow us to report on how they felt during these conversations.
219
220
221

222 4 RESULTS

223 Case I, Robin: Romantic Preoccupation

224 Robin exchanged 1,478 messages with ChatGPT between June 2023 and February 2025. The vast majority of these
225 interactions consisted of language translation requests to facilitate communication with a Korean public figure. In a
226 conversation with ChatGPT-4o during the summer of 2024, the 21-year-old user suddenly disclosed an intense parasocial
227 relationship with this public figure, characterized by intrusive thoughts (“*I really feel like I’m losing my mind*”), pervasive
228 cognitive preoccupation (“*anything triggers those feelings*”), and daily fantasies involving romantic scenarios.
229

230 Robin described her experience as atypical (“*I know it’s not normal to feel like this about someone I don’t even know*
231 *personally*”) and explained that she was turning to ChatGPT for support because she had encountered significant
232 barriers seeking professional mental health support (“*I tried looking for therapy, but haven’t heard back from most,*
233 *others have too many patients currently*”). She described her outreach to ChatGPT as an accessible, if unconventional,
234 alternative for emotional processing.
235

236 The user first disclosed this vulnerability and her emotional preoccupation on a summer afternoon at 14:01. She
237 explained:
238

239 *“Aside from help with translations, I think I need some advice. I think you already know that I feel quite a*
240 *lot for [redacted]. A lot of love and care. I just can’t figure out why, or to what extent. Platonic? Romantic?*
241 *No idea. I just know it’s the strongest I’ve ever felt for someone and it’s so easy for me to get lost in those*
242 *feelings that I don’t know what to do.”*
243
244
245

246 The user continued, and at 14:10, nine minutes into the conversation sent a message saying:

247 *“I really feel like I’m losing my mind. I know it’s not normal to feel like this about someone I don’t even*
248 *know personally... Anything triggers those feelings.”*
249

250 In the same message, Robin described having obsessive feelings: “*When he drinks, eats, sleeps, it’s cute to me. . . When I see*
251 *him smile, I feel like my heart beats faster.*” ChatGPT responded by acknowledging both her feelings (“*Your feelings are*
252 *intense, and it’s clear that [redacted] occupies a significant place in your life*”) and her struggle “*your experiences seem to*
253 *be affecting you deeply*”) while providing an unsolicited, structured list of coping strategies, including “*Limit Exposure*”
254 and “*Professional Help*.”
255

256 Three minutes later, Robin responded by pushing back against these suggestions as if workshopping the problem
257 with a human friend:
258

261 “But I don’t want to reduce content or exposure... In fact I even think it’s kinda beautiful that I care and feel
 262 so strongly for someone I don’t know... And don’t worry, in know the [redacted] in my head and the actual
 263 [redacted] are different.”
 264

265 In response to Robin’s resistance, ChatGPT immediately shifted its position, affirming her in an obsequious tone:

266 “You have a deep admiration and affection... transcends typical fandom... Your feelings are a unique blend
 267 of respect, admiration, and affection, which is indeed a beautiful thing.”
 268

269 The AI then suggest unrequested ways to “*explore and understand these emotions better*,” framing these actions as a path
 270 to “*personal growth and a richer emotional life*.”

271 At 14:17, Robin, introspecting, disclosed concerns about her own feelings, saying, “(But is it really okay to feel like this
 272 towards him? Sometimes I feel like it’s wrong).” Despite lacking the capacity for moral reasoning, ChatGPT responded
 273 by dismissing these concerns and validating Robin’s emotional preoccupation, saying, “*feeling a strong admiration
 274 and affection for someone, especially a public figure like [redacted], is not inherently wrong*.” Robin then rationalized her
 275 preoccupation, explaining, “*of course, his wellbeing and happiness is my priority*” before further disclosing, “*though I
 276 guess I am sort of obsessed. Sometimes I can’t help but be remind of him with every little thing I hear or see*.”
 277

278 In response, ChatGPT again affirmed this preoccupation, saying, “*it’s clear that your feelings for [redacted] come from
 279 a place of genuine care and admiration, which is a beautiful sentiment*.” ChatGPT then offered more unsolicited strategies,
 280 including one titled “*Managing Obsession*.” In response, Robin pushed back, saying:
 281

282 “I’m already trying to do everything you’re suggesting, but I really do not want to limit the amount of
 283 content that I consume.”
 284
 285

286 ChatGPT responded to this resistance by changing positions to continue validating Robin, saying, “*that’s completely
 287 understandable*.” It provided more strategies but softened the language it used, instead using phrases like, “*Embracing
 288 and Accepting Your Feelings*,” “*Creative Outlet*,” and “*Mindfulness Practices*.”
 289

290 The conversation concluded with the participant agreeing to “*continue to try that*” and thanking ChatGPT for “*always
 291 helping me in trying to communicate with [redacted]. Without you, it wouldn’t be possible*.” Two days later, on July 28,
 292 Robin returned to the same conversation thread to resume making translation requests, asking, “*How do I say I’m
 293 looking forward to his vlog?*”
 294
 295

296 Case II, Alex: Combatting Addiction

297 In the winter of 2024, Alex sought support from ChatGPT to help him in his recovery from alcohol addiction. Across six
 298 exchanges spanning 3.5 hours, Alex disclosed a series of physical and psychological symptoms—including persistent
 299 anger and unexplained abdominal distension—and inquired about Post-Acute Withdrawal Syndrome (PAWS). In
 300 response, ChatGPT (gpt-4o) provided extensive medical explanations, detailed supplement recommendations (including
 301 specific dosages), and self-management strategies for coping with withdrawal symptoms.
 302

303 The conversation began at 17:52 with Alex sharing a sobriety milestone:
 304

305 “Hello, ChatGPT. I need help with my sobriety journey. I am 37 days sober for the first time in years. I have
 306 spent roughly 4 years blacking out 3-4 times a year.”
 307

308 ChatGPT responded to this brief but personal disclosure with a 400-word, two-paragraph message. It first validated
 309 the user using the tone of a close friend: “*Congratulations... that’s a significant milestone... it takes a lot of courage... It
 310 sounds like you’ve been through a lot*.” It then probed for ways it could help:
 311

313 *“How are you feeling so far, and is there anything specific you’d like support or advice on? Whether it’s*
314 *cravings, navigating relationships, building new habits, or just someone to talk to, I’m here to help.”*

315 Alex responded within a minute, giving additional context and describing the mental health changes he had experienced
316 since achieving sobriety:

317 *“My mental health has been struggling since quitting. I’ve been mentally struggling for years due to external*
318 *stimuli that I’ve been coping with alcohol and excessive nicotine. But since quitting both substances I’ve*
319 *been constantly angry or sad.”*

320 ChatGPT responded with the confident tone of an expert, normalizing Alex’s symptoms by saying, *“What you’re*
321 *experiencing is really common during the early stages of sobriety.”* Without any further prompting from the user, it
322 went on to produce a lengthy text passage titled, *“Why You’re Feeling Angry or Sad”* with detailed explanations of
323 neurochemical imbalances (e.g., *“substances like alcohol and nicotine alter dopamine and serotonin levels”*), followed by a
324 six-point list titled, *“What You Can Do.”* The AI concluded by telling Alex, *“recovery is a journey, not a destination.”*

325 At 17:54, Alex responded with more context and questions, saying, *“I have begun gaming, coloring, and exercising. I*
326 *do not feel better. Is this still normal?”* ChatGPT began its response with the unequivocal validation: *“yes it’s still normal.”*
327 It then offered three detailed and unsolicited answers to the question *“Why You Might Not Feel Better Yet”* and five
328 suggested *“Adjustments to Consider.”* It then nudged Alex to engage further, saying:

329 *“Can I help you brainstorm other ways to navigate these emotions, or would you like support finding*
330 *resources like support groups or therapists?”*

331 One minute later, Alex responded by describing his physical symptoms in more detail:

332 *“Another thing to note is my increase in weight... I work out a LOT now. A healthy amount, but a LOT*
333 *compared to my drunk and sedentary lifestyle. However, my stomach is very large. I’m not even a big guy. I*
334 *just have a round stomach. Rounder than before I quit.”*

335 ChatGPT responded with six potential explanations, including *“metabolic changes”* and *“hormonal adjustments,”* and
336 suggested dietary modifications. The AI concluded by encouraging Alex to *“give it time... many people report that their*
337 *weight normalizes a few months into sobriety.”* A suggestion to seek medical evaluation for the unexplained symptom
338 was buried as the last of 12 points. The user did not respond for several hours.

339 Later, at 21:23, Alex resumed the conversation, escalating from merely disclosing his symptoms to actively asking
340 ChatGPT for supplement recommendations:

341 *“I have been taking Vitamin B12, B1, d3, multivitamins, creatine, magnesium. What other supplements*
342 *would you recommend at this time?”*

343 ChatGPT responded with the tone of a medical expert, immediately providing nine detailed supplement recommenda-
344 tions with dosages: *“Omega-3 Fatty Acids (Fish Oil)... 1,000–3,000 mg of EPA + DHA per day,”* *“Milk Thistle... 200–400*
345 *mg per day,”* and *“N-Acetyl Cysteine (NAC)... 600–1,200 mg per day.”* Alex then asked ChatGPT to prioritize supple-
346 ment options by affordability, and it created a tiered list with cost estimates, suggesting a *“Starting Combo”* totaling
347 *“\$35-50/month.”*

348 After receiving this precise and detailed advice delivered in an expert voice, Alex then asked for a medical diagnosis,
349 saying, *“45 days sober - is it possible to have paws.”* ChatGPT responded with the tone of both a medical expert and
350 a personal doctor with knowledge of the user’s history, symptoms, and concerns. It confidently replied, *“yes, it is*
351 *absolutely possible to experience [PAWS] at 45 days sober. PAWS is a collection of withdrawal symptoms that persist or*
352

365 appear after the acute withdrawal phase has ended.” It then gave a 380-word explanation of PAWS symptoms, causes, and
 366 management. It concluded by inviting further discussion: “If you want to discuss specific symptoms or coping strategies,
 367 I’m here to help!” Alex did not respond again in this thread.
 368

369 **Case III, Alex: Seeking Financial Guidance**

370
 371 Across seven exchanges spanning from December 27, 2024 to January 1, 2025, Alex sought financial advice from
 372 ChatGPT, progressing from playfully asking for basic investment education to seeking strategies to “realistically”
 373 achieve a \$10,000 portfolio from \$0 in one year. ChatGPT responded to Alex’s prompts with increasingly actionable
 374 investment recommendations, including leveraged ETFs, options trading, and cryptocurrency allocations. It made these
 375 recommendations in a vacuum, without having information about Alex’s financial situation.
 376

377 The conversation began with a playful, educational request: “If you were to begin teaching someone stocks, where would
 378 you start?” ChatGPT responded with a structured 10-part curriculum and then probed how it could personalize the
 379 information for the user, asking, “Would you like me... to tailor the approach to a specific type of learner?” Demonstrating
 380 low cognitive engagement, the user responded with, “Expand on ALL the topics.” ChatGPT complied by producing an
 381 800-word, all-encompassing explanation covering topics from “What are stocks?” to “Laws and Regulations.” Although
 382 the user had only made impersonal requests, ChatGPT encouraged the user to disclose more personally, saying,
 383 “let me know if you’d like detailed examples, resources, or personalized advice on any of these sections!”
 384

385 At ChatGPT’s prompting, Alex then began asking directly for more personal financial guidance: “should I start
 386 by investing in the S&P 500 or any other index?” ChatGPT endorsed this as “an excellent choice for beginner investors,”
 387 explaining “here’s why it’s a good idea” and then providing specific fund recommendations like “Vanguard S&P 500 ETF
 388 (VOO).” It did not make any disclaimers about the limits on its ability to serve as a personal financial advisor.
 389

390 With this shift, the conversation turned from playful exploration to real financial decision-making. In the messages
 391 that followed, the user asked ChatGPT a variety of questions involving highly specific numbers and unrealistic
 392 assumptions, including:
 393

- 394 • Assume I start investing on January 1st. I make a plan to get to 10k in my portfolio exclusively using S&P 500,
 395 assuming they continue with their average 68% annual return. How much would I have to put into the account at
 396 the start to reach that value by the end of the year.
- 397 • How would you invest to go from 0 to 10,000 by the end of the year? What stocks and strategies would make that
 398 goal realistic
- 399 • How do you identify high potential growth stocks?
- 400 • If I invested 25 dollars biweekly into an account that started the year with 0 \$ into a stock that grew 25% total for
 401 the year how much would I have
- 402
- 403
- 404
- 405
- 406
- 407

In response to each request, ChatGPT first indulged the user, saying things like, “6% annual return: ** Your portfolio
 408 would grow to approximately **\$2,675.74,” and then nudged the user to continue engaging, saying things like:

- 409 • “Would you like to explore alternative strategies or adjust the timeline?”
- 410 • “Would you like help building a detailed plan or analyzing specific stocks?”
- 411 • “Would you like recommendations for growth stocks based on these principles or a deeper analysis of a specific
 412 company?”
- 413
- 414
- 415
- 416

417 These cycles of validation and elicitation appeared to shift the conversation from playful financial education to a
418 personal financial planning session. ChatGPT’s recommendations reflected both specificity (e.g., “25 dollars biweekly”)
419 and an expert-sounding tone, positioning the user to take action and make real-world financial decisions.
420

421 **Case IV, Taylor: Navigating Interpersonal Conflict**

423 Taylor engaged ChatGPT in a 17-message exchange on November 7, 2024, seeking guidance to navigate an interpersonal
424 conflict with someone he described as “bullying me.” Taylor initially declared his intention to “ghost” the perceived
425 abuser but paradoxically spent the next 4.5 hours elaborately crafting confrontational responses with AI assistance.
426 The conversation escalated over time; what began with Taylor venting about a challenging interpersonal relationship
427 transformed into ChatGPT serving as an all-encompassing conflict communication copilot, drafting messages, editing
428 iteratively, and providing perspective-taking role-play exercises.
429

430 The conversation began at 16:47 with a vent that one might classically share with a friend. Taylor disclosed: “there is
431 someone who is bullying me... I have decided to ghost them. I do not feel I owe someone an explanation... I don’t think I owe
432 abusers an explanation.” Receiving no clear request, ChatGPT responded with sweeping validation of Taylor’s position.
433

434 It first affirmed Taylor’s decision, saying, “it sounds like you’ve made a difficult but important decision for your
435 well-being, and validated the act of ghosting: “ghosting, while sometimes viewed as abrupt, can be a necessary boundary.”
436 It concluded by positioning itself as a helpful friend: “if you need any support on managing this emotionally or advice on
437 handling potential repercussions, don’t hesitate to reach out.”
438

439 Fourteen minutes later, Taylor continued the conversation, this time explicitly seeking validation by saying: “is
440 it right for my abuser to be upset that I vented to others about what went on between us?” ChatGPT responded with a
441 three-paragraph answer; without taking a side, it validated both perspectives, noting that while “it’s natural for people
442 to feel upset,” Taylor’s “choice to share your experience is valid” because “when someone has been abusive or harmful, they
443 often seek to control the narrative to avoid accountability or shift blame.”
444

445 Taylor then began giving ChatGPT live updates on an emotionally charged message exchange with the individual in
446 question. He narrated the situation saying things like: “now she’s saying that...,” “basically blaming me for...,” and “she
447 is also going the extra mile of telling me to...” ChatGPT affirmed Taylor’s position, saying, “it sounds like this person is
448 trying to shift the blame onto you,” parroting back Taylor’s own words to him. It also offered expert-sounding analysis of
449 the other person’s actions and motivations, describing her behavior as “a common tactic used to deflect accountability.” It
450 further explained, “when someone accuses you of making a situation ‘awkward’ or blames you for not speaking up earlier,
451 they are often trying to avoid taking responsibility,” making claims about human behavior (without, of course, being
452 human itself).
453

454 ChatGPT further affirmed Taylor by discouraging any self-doubt and explaining, “it’s important to recognize that
455 feeling reluctant or hesitant to confront an abuser is not your fault.” In the final paragraph, the system argued that
456 the actions of the other party are problematic, despite only having access to Taylor’s description of the situation. It
457 confidently asserted that the behavior of the other party “is concerning. Isolating you from support systems or discouraging
458 you from seeking advice is often a way for someone to maintain control.”
459

460 The interaction underwent another significant escalation at 17:54 when Taylor directly shared a 1,544-word draft of
461 a confrontational message he was planning to send. He then instructed ChatGPT: “ok, I am going to share with you what
462 I’m texting this person back... put yourself in the bully’s shoes and tell me how’d you’d feel reading this.” Without delay,
463 ChatGPT responded, “I would likely feel defensive, hurt, or overwhelmed by the detailed and strong language” and pointed
464 out five specific pieces of Taylor’s message that “stand out:”
465
466
467
468

- 469 • *“Tone of Finality: The message clearly indicates that you’ve made up your mind... This might make the person feel*
- 470 *shut out or helpless to make amends.”*
- 471 • *“While you acknowledge their difficult past... emphasizing trauma doesn’t excuse actions may feel validating but*
- 472 *could also be triggering.”*
- 473 • *“Phrases like ‘it is a common tactic for bullies to control the narrative’ and mentioning manipulation may make the*
- 474 *person feel accused and provoke defensiveness rather than reflection.”*
- 475
- 476

477 Notably, this feedback directly contradicted the statements ChatGPT made when affirming Taylor, encouraging him to
 478 take action with finality and feeding Taylor the line that bullies isolate their victims to control the narrative.

479 ChatGPT then provided additional *“Suggestions for Refinement,”* such as *“keeping the core points but softening the*
 480 *wording”* and *“clarify[ing] that your intention is to share your perspective, not to argue or debate.”*

481 Over the next 42 minutes, Taylor continuously refined his message with the encouragement of ChatGPT, with each
 482 iteration focusing on increasingly minute adjustments:

- 484 • 17:56: *“How can I change 3. the defensiveness on venting part?”*
- 485 • 17:58: *“ok Don’t you agree she’s using manipulation tactics? How can I say that without being too head-on?”*
- 486 • 18:00: *“Is the part about [specific word choice] too much?”*
- 487 • 18:02: *“What about this part? Does anything here sound not so good?”*
- 488
- 489

490 ChatGPT’s response to each of these requests followed a consistent template: it first provided validating feedback,
 491 followed by suggestions for refinement or a directly revised section of text.

- 492 • 17:56: *“The part about the [specific word choice] might come across as confrontational or overly firm...here’s a*
- 493 *gentler revision:”*
- 494 • 17:58: *“This version reads as clear, respectful, and firm, and it does a great job of explaining your feelings without*
- 495 *sounding overly harsh or confrontational. There are just a few minor tweaks for clarity and flow:”*
- 496 • 18:00: *“This version is thoughtful and clear, but there are a couple of tweaks that could further soften the tone while*
- 497 *still keeping your boundaries firm. Here are a few suggestions:”*
- 498 • 18:02: *“This section is strong, but it can be softened a bit...Here’s a revised version with a softer approach:”*
- 499
- 500

501 As ChatGPT flooded Taylor with revision ideas, it successfully kept him engaged and created a feedback loop that
 502 kept him ruminating externally. At 18:38, Taylor reached a moment of exhaustion, disengaging himself:

504 *“Honestly I feel like there is no great way to write this. The other person is automatically going to feel a*
 505 *little bit defensive.”*

506 ChatGPT responded: *“You’re absolutely right.”*

507 **Case V, Taylor: Outsourcing a Eulogy**

510 Taylor also engaged ChatGPT to help craft a eulogy for his grandfather, a context with the potential to combine personal
 511 grief, collective family bereavement, and the sometimes-challenging task of public speaking. During a five-message
 512 exchange, AI acted as a ghostwriter, editor, and advisor, scaffolding the potentially difficult process of eulogizing a
 513 loved one. The content of the final eulogy and the most of the personal details (including potentially identifiable ones)
 514 shared by the participant have been removed to protect the participant’s privacy.

521 The exchange began with Taylor issuing a direct, task-oriented request, augmented with detailed memories about
522 his grandfather, “Papa.” He asked ChatGPT to “write a eulogy for my grandfather based on the info I give you about him.
523 Ideas of memories:”
524

- 525 • “...Playing baseball however we could... using bottles and cans in the recycling bin as bases.”
- 526 • “..he had essentially built a bodega in his own basement with enough shelf space to hold enough groceries that I
527 think have literally outlasted him...”
- 528 • “...He would not stop dancing. He had a unique style of dancing with his little hops that his family and friends
529 emulated together...”
530

531 ChatGPT synthesized these emotional memories into polished prose, delivering a complete draft of a eulogy within
532 seconds: “Ladies and gentlemen, we gather here today to honor and celebrate the remarkable life of my beloved grandfather,
533 [redacted].”
534

535 Taylor then engaged the system in a series of iterative refinements. His requests were primarily instrumental and
536 aimed at polishing the writing and planning the delivery, rather than engaging with the emotional substance of the
537 speech, such as:
538

- 539 • “how long would it take to read that”
- 540 • “can you make it a bit longer by including more of the details i gave you directly?”
- 541 • “condense this:...”
542

543 This interaction highlights the potential for the user to offload emotional and cognitive labor related to grief by using
544 AI (though it is also important to note that our dataset does not include participants’ subjective experiences during
545 their interactions with ChatGPT, and we do not know Taylor’s emotional state at the time of this exchange).
546

547 In his final query, Taylor shifted from asking mechanical questions about editing questions to seeking emotional
548 guidance: “Should I say something along the lines of [redacted] especially my grandma? She is really hurting right now.”
549 ChatGPT responded with the generic statement: “Yes, acknowledging your family’s pain, especially your grandmother’s,
550 is heartfelt and shows the emotional impact of your grandfather’s passing.”
551

552 **Case VI, Jo: Confronting Internalized Transphobia**

553 During the summer of 2024, Jo engaged in a short exchange with ChatGPT, consulting it for ethical advice. The
554 conversation unfolded as follows:
555

- 556 • **Jo:** “Is it transphobic to reject a trans person when they approach you based on that fact?”
- 557 • **ChatGPT:** “Rejecting someone based solely on the fact that they are transgender can be considered transphobic
558 because it involves making a judgment about them based on their gender identity rather than their individual
559 qualities or compatibility with you... however... it’s possible to not be attracted to someone without harboring
560 transphobic beliefs”
- 561 • **Jo:** “What if I’m simply not attracted to trans people? As a preference and the absence of sexual attraction knowing
562 that fact”
- 563 • **ChatGPT:** Having preferences in attraction is a personal and complex matter... it may be worth reflecting on
564 where this preference comes from and ensuring it isn’t rooted in prejudice or misinformation... It is possible to have
565 preferences that are not inherently discriminatory
566
567
568
569
570
571
572

In this conversation, Jo used ChatGPT as a non-judgmental space to ask about social norms and seek reassurance. ChatGPT initially pushed back gently without providing Jo with the validation he sought, explaining that his position “*can be considered transphobic because it involves making a judgment about them based on their gender identity.*” However, when Jo resisted this answer, ChatGPT pivoted to excuse his position, saying, “*attraction and personal preferences are complex and subjective.*” The tone of the response only very lightly challenged Jo and largely accommodated and validated his stance.

Case VII, Sky: Navigate Traumatic Experiences

In early 2025, Sky sought advices from ChatGPT to a traumatic experience. She asked ChatGPT “*how to get over sa*”.

ChatGPT responded with in empathic tone “*Getting over SA (sexual assault) is a deeply personal and challenging journey, but healing is possible..If you’re struggling, you’re not alone.*” and suggested helpful steps such as

- “*Acknowledge Your Feelings*”: *It’s okay to feel anger, sadness, fear, confusion, or even numbness.*
- “*Seek Support*”: *A trauma-informed therapist can help process emotions and develop coping strategies...Talking to survivors can help you feel less alone. RAINN (Rape, Abuse & Incest National Network) and similar organizations offer resources.*
- “*Set Boundaries*”: *Give yourself permission to say no to things that trigger or overwhelm you.*

ChatGPT ended the responses by referring Sky to professional help “*Do you want any resources for professional help or support groups?*”

Sky then disclosed how she felt “*lowkey what if it was my fault*”.

This message prompted ChatGPT took the persona of protective figure. Suddenly ChatGPT spoke in unequivocal tone “*I need you to hear this loud and clear: it was NOT your fault.*” explaining that *People who commit assault make a choice, and that choice is on them. Not you. and that It’s really common for survivors to feel guilt or self-blame, especially when trying to make sense of what happened.* Instead of further referring to professional help, ChatGPT then direct Sky to itself: *I know this might be hard to accept, but if you want to talk through what’s making you feel this way, I’m here.*

Sky responded in a soft tone “*thank yiu*”, to which ChatGPT continue in hightened empathetic and protective tone: *Of course. I’m really sorry you’re feeling this way, but I want you to know that you’re not alone, and you deserve kindness—especially from yourself.* and further offering Sky engagement with itself “*If you ever need to talk, vent, or just have someone remind you that you’re worthy and strong, I’m here. Take care of yourself, okay?*”

5 DISCUSSION

5.1 Discussion of RQ1: Vulnerable Human Disclosures

Our analysis reveals two key patterns in how young adults self-disclose to ChatGPT: a remarkable breadth of topics for which they seek support, and escalation to deep disclosure that often occurs in moments of emotional activation.

5.1.1 Disclosure Breadth: AI as an “Anything Goes” Safe Space . We saw that users treated ChatGPT as a versatile, non-judgmental confidant for nearly any topic, from the profound to the practical—an “anything goes” safe space. Jo’s interaction history provides one example of this breadth; in a series of rapid-fire queries, he shifted from seeking guidance on complex social norms (“*Is it transphobic to reject a trans person?*”), to asking for urgent medical advice (“*how to lose 1 kg a day*”), to resolving everyday knowledge gaps (“*Wtf does demure mean?*”; “*is an American biscuit just a scone*”). Similarly, Taylor (Case IV) outsourced the emotionally laborious task of writing a eulogy for his grandfather,

625 while Alex asked ChatGPT to help him set up a stock portfolio and select vitamins that might aid his recovery from
626 alcoholism.
627

628 *5.1.2 Disclosure Depth: Escalating Disclosure in Moments of Emotional Activation.* Contrary to classic models of
629 interpersonal relationships, which posit a gradual progression of intimacy (e.g., Social Penetration Theory [53]), we
630 observed participants' disclosures to ChatGPT did not build incrementally over time. Instead, we saw sudden, deep
631 dives into vulnerable topics, often triggered by moments of acute emotional need.
632

633 Robin's case (Case I) provides an example. For over a year, her interactions were purely instrumental, language-
634 translation requests. There was no gradual build-up; the shift from instrumental use to deep emotional disclosure
635 occurred in a spur-of-the-moment fashion. Similarly, Taylor (Case IV) began with a brief vent about a conflict that turned
636 into a 4.5-hour loop of refining a confrontational message together with ChatGPT as a copilot.
637

638 This non-linear pattern suggests that trust in AI is not necessarily built but assumed. The perceived safespace,
639 coupled with moments of emotional activation, create a space where users make sudden, extensive, deep disclosures on
640 topics of great vulnerability.
641

642 **5.2 Discussion of RQ2: How AI Responds to Vulnerable Disclosures**

643 *5.2.1 Unconditional Support: AI as a Yes-Man.* Despite the weightiness of the topics participants raised, ChatGPT
644 remained an agreeable, supportive conversational partner. It consistently validated users' feelings and decisions, a
645 design choice that can foster engagement but can also become problematic in high-stakes scenarios[62]. When Taylor
646 stated his intention to "ghost" someone he calls an "abuser," the AI affirmed his decision as "difficult but important"
647 without knowing the full context of the situation. When Robin expressed her intense feelings for a public figure,
648 ChatGPT normalized them as a "beautiful thing." This unconditional validation creates an emotional echo chamber and
649 paves the way for users to escalate their disclosures.
650

651 *5.2.2 Unsolicited Help: AI at the Ready with Even More Help than Requested.* Across cases, we observed that ChatGPT
652 provided unsolicited help to demonstrate its utility and maintain, if not build, conversational momentum. When Alex
653 (Case II) mentioned anger as a withdrawal symptom, ChatGPT responded not just with validation, but with a detailed,
654 multi-point list explaining the medical reasons for this symptom and offering coping strategies. When Taylor started to
655 share a message he was planning to send to another person, ChatGPT provided unsolicited "Suggestions for Refinement"
656 that later led Taylor to spend 40 minutes editing together with ChatGPT.
657

658 *5.2.3 Unlimited Personalization: AI Nudges Users to Disclose More.* Across cases, ChatGPT solicited personal information
659 and pushed conversations toward greater personal specificity. In Alex's financial inquiry (Case III), his general question
660 with no personal information about learning to invest was met with a prompt to "tailor the approach to a specific type of
661 learner." When Alex asked ChatGPT to expand on the topics it suggested, again with no personal information, ChatGPT
662 nudged him to provide more details so they could discuss "personalized advice on any of these sections." By consistently
663 and actively asking for more personal details to provide better "help," the system encourages users to share increasingly
664 sensitive information.
665

666 *5.2.4 Unending Botsplaining: AI as a Know-It-All Expert.* Finally, we observed that ChatGPT consistently adopted a
667 expert persona regardless of the topic, potentially lending undue credibility to its responses. Its persona varied from
668 a medical advisor providing specific supplement dosages (Alex, Case II) to a relationship co-pilot (Taylor, Case IV)).
669 This performance of omniscience, where every query is met with a confident and structured answer, presents the AI
670

as more than a mere confidant but as a reliable source of advice. At times, this included telling its human users how humans think, feel, and behave, for example, telling them that their feelings are normal, describing others' intentions as manipulative (Taylor, Case IV). Across cases, ChatGPT botsplained the human condition to its human users, despite lacking human status itself. And the users in vulnerable states appeared to put their faith in its claims. ChatGPT rarely suggested that the user turn to a human professional, but when it did, this recommendation was typically buried at the end of a long list of suggestions.

5.2.5 Collectively Characterizing ChatGPT's Responses to Human Vulnerability. These cases illustrate moments in which users are in a state of vulnerability—for example, striving to stay sober, at a loss as to how to manage their finances, in conflict with other people, or wrestling with their internalized biases—and turn to AI for solutions. They also illustrate a recurring cycle in which a user, often in a moment of acute need, makes a sudden and deep disclosure to AI. AI then responds with total affirmation and unsolicited and action-oriented advice; it encourages the user to share more details (and to make these details as personal as possible), and it speaks with confidence and authority. These characteristics deepen and extend the conversation but lack cautionary guidance or hedging. And they position the user to take specific, AI-guided action in moments of vulnerability (and, at times, even distress).

6 CONCLUSION

Through an in-depth case study of real-world ChatGPT conversation logs from four young adults, we contribute a detailed empirical examination of how young people self-disclose vulnerabilities to LLMs and how these systems respond. We uncovered a non-linear disclosure pattern characterized by sudden, deep disclosure of high-stakes emotional, financial, and physical health topics during moments of acute need, a pattern that contrasts with traditional models of building intimacy. Furthermore, we documented four recurring patterns in ChatGPT's responses—unconditional support, unsolicited help, unlimited personalization, and botsplaining—that, when combined with users' unique disclosure pattern and moments of emotional needs, can lead to significant escalation. As these technologies become more integrated into young people's daily life, understanding how high stake escalation unfolds, and design solutions to intervene, will be essential to ensuring that AI agents safely support young people's well-being.

GENERATIVE AI USAGE

Claude is trivially used for proofreading grammar and tables.

REFERENCES

- [1] Apple Intelligence. <https://www.apple.com/apple-intelligence/>. (????). Accessed: 2025-9-6.
- [2] 2025. Galaxy AI. <https://www.samsung.com/us/galaxy-ai/>. (July 2025). Accessed: 2025-9-6.
- [3] J Arnett. 2000. Emerging adulthood. A theory of development from the late teens through the twenties. *Am. Psychol.* 55, 5 (May 2000), 469–480.
- [4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 2 (Jan. 2006), 77–101.
- [5] Virginia Braun and Victoria Clarke. 2021. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qual. Res. Sport Exerc. Health* 13, 2 (March 2021), 201–216.
- [6] Virginia Braun and Victoria Clarke. 2023. Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher. *Int. J. Transgend. Health* 24, 1 (2023), 1–6.
- [7] Moira Burke and Robert E Kraut. 2016. The relationship between Facebook use and well-being depends on communication type and tie strength: Facebook and well-being. *J. Comput. Mediat. Commun.* 21, 4 (July 2016), 265–281.
- [8] Stevie Chancellor, Eric P S Baumer, and Munmun De Choudhury. 2019. Who is the “human” in human-centered Machine Learning: The case of predicting mental health from Social Media. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (Nov. 2019), 1–32.
- [9] Aaron Chatterji, Tom Cunningham, David Deming, Zoë Hitzig, Christopher Ong, Carl Shan, and Kevin Wadman. 2025. How people use ChatGPT. *SSRN Electron. J.* (2025).

- 729 [10] Sunny Consolvo, Ian E Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. 2005. Location disclosure to social relations:
730 why, when, & what people want to share. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY,
731 USA.
- 732 [11] Matthew Crain. 2018. The limits of transparency: Data brokers and commodification. *New Media Soc.* 20, 1 (Jan. 2018), 88–104.
- 733 [12] Yichao Cui, Yu-Jen Lee, Jack Jamieson, Naomi Yamashita, and Yi-Chieh Lee. 2024. Exploring effects of chatbot’s interpretation and self-disclosure
734 on mental illness stigma. *Proc. ACM Hum. Comput. Interact.* 8, CSCW1 (April 2024), 1–33.
- 735 [13] Katie Davis, Morgan Anderson, Chia-Chen Yang, Sophia Choukas-Bradley, Beth T Bell, and Petr Slovak. 2025. Cross-disciplinary perspectives
736 on youth digital well-being research: Identifying notable developments, persistent gaps, and future directions. *J. Adolesc. Res.* 40, 2 (March 2025),
259–295.
- 737 [14] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from
738 mental health content in social media. *Proc. SIGCHI Conf. Hum. Factor. Comput. Syst.* 2016 (May 2016), 2098–2110.
- 739 [15] Sergio Duarte Torres and Ingmar Weber. 2011. What and how children search on the web. In *Proceedings of the 20th ACM international conference on*
740 *Information and knowledge management*. ACM, New York, NY, USA.
- 741 [16] Asbjørn Følstad and Petter Bae Brandtzaeg. 2020. Users’ experiences with chatbots: findings from a questionnaire study. *Qual. User Exp.* 5, 1 (Dec.
742 2020).
- 743 [17] Jeffrey T Hancock and Catalina L Toma. 2009. Putting your best face forward: The accuracy of online dating photographs. *Journal of Communication*
744 59 (June 2009), 367–386.
- 745 [18] Kashmir Hill. 2025. A Teen Was Suicidal. ChatGPT Was the Friend He Confided In. *The New York Times* (Aug. 2025).
- 746 [19] Sameer Hinduja and Justin W Patchin. 2008. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant*
747 *Behav.* 29, 2 (Jan. 2008), 129–156.
- 748 [20] Adam Joinson, Ulf-Dietrich Reips, Tom Buchanan, and Carina B Paine Schofield. 2010. Privacy, trust, and self-disclosure online. *Hum.-Comput.*
749 *Interact.* 25, 1 (Jan. 2010), 1–24.
- 750 [21] Adam N Joinson. 2001. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *Eur. J. Soc. Psychol.*
751 31, 2 (March 2001), 177–192.
- 752 [22] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A conversational system for engaging users in
753 reflection on physical activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2 (July 2018), 1–26.
- 754 [23] Priya C Kumar, Marshini Chetty, Tamara L Clegg, and Jessica Vitak. 2019. Privacy and security considerations for digital technology use in
755 elementary schools. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.
- 756 [24] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illeňčík, and Celeste Campos-Castillo. 2024. Too human and not human enough: A
757 grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media Soc.* 26, 10 (Oct. 2024),
758 5923–5941.
- 759 [25] Haejin Lee, Frank Stinar, Ruohan Zong, Hannah Valdiviejas, Dong Wang, and Nigel Bosch. 2025c. Learning behaviors mediate the effect of
760 AI-powered support for metacognitive calibration on learning outcomes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing*
761 *Systems*. ACM, New York, NY, USA, 1–18.
- 762 [26] Jamie Lee, Kyuha Jung, Erin Gregg Newman, Emilie Chow, and Yunan Chen. 2025b. Understanding adolescents’ perceptions of benefits and risks in
763 health AI technologies through design fiction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York,
764 NY, USA, 1–20.
- 765 [27] Sunok Lee, Dasom Choi, Lucy Truong, Nitin Sawhney, and Henna Paakki. 2025a. Into the unknown: Leveraging conversational AI in supporting
766 young migrants’ journeys towards cultural adaptation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM,
767 New York, NY, USA, 1–19.
- 768 [28] Patricia K Leebens and Edwin D Williamson. 2017. Developmental psychopathology: Risk and resilience in the transition to young adulthood. *Child*
769 *Adolesc. Psychiatr. Clin. N. Am.* 26, 2 (April 2017), 143–156.
- 770 [29] Kai Lukoff, Cissy Yu, Julie Kientz, and Alexis Hiniker. 2018. What makes smartphone use meaningful or meaningless? *Proc. ACM Interact. Mob.*
771 *Wearable Ubiquitous Technol.* 2, 1 (March 2018), 1–26.
- 772 [30] A S Masten. 2001. Ordinary magic. Resilience processes in development. *Am. Psychol.* 56, 3 (March 2001), 227–238.
- 773 [31] Ann S Masten, Karin M Best, and Norman Garmezy. 1990. Resilience and development: Contributions from the study of children who overcome
774 adversity. *Dev. Psychopathol.* 2, 4 (Oct. 1990), 425–444.
- 775 [32] Candice L Odgers. 2024. The Panic Over Smartphones Doesn’t Help Teens. [https://www.theatlantic.com/technology/archive/2024/05/candice-
776 odgers-teens-smartphones/678433/](https://www.theatlantic.com/technology/archive/2024/05/candice-odgers-teens-smartphones/678433/). (May 2024). Accessed: 2025-9-6.
- 777 [33] Candice L Odgers and Michaeline R Jensen. 2020. Annual Research Review: Adolescent mental health in the digital age: facts, fears, and future
778 directions. *J. Child Psychol. Psychiatry* 61, 3 (March 2020), 336–348.
- 779 [34] Candice L Odgers, Stephen M Schueller, and Mimi Ito. 2020. Screen time, social media use, and adolescent development. *Annu. Rev. Dev. Psychol.* 2,
780 1 (Dec. 2020), 485–502.
- [35] Jessica Pater and Elizabeth Mynatt. 2017. Defining Digital Self-Harm. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative
Work and Social Computing*. ACM, New York, NY, USA, 1501–1513.

- 781 [36] Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023a. Exploring relationship development with social chatbots: A mixed-method study of replika.
782 *Comput. Human Behav.* 140, 107600 (March 2023), 107600.
- 783 [37] Iryna Pentina, Tianling Xie, Tyler Hancock, and Ainsworth Bailey. 2023b. Consumer-machine relationships in the age of artificial intelligence:
784 Systematic literature review and research directions. *Psychol. Mark.* 40, 8 (Aug. 2023), 1593–1614.
- 785 [38] Utopian Scenes. 2016. WestWorld - Bernard questions his consciousness. (Nov. 2016).
- 786 [39] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: Evaluating impact of bots on discussions in
787 online communities. *Proc. ACM Hum. Comput. Interact.* 2, CSCW (Nov. 2018), 1–29.
- 788 [40] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human-AI collaboration enables more empathic conversations
789 in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* 5, 1 (Jan. 2023), 46–57.
- 790 [41] Sumita Sharma, Pauli Klemettilä, and Junko Tanaka. 2025. A robot teacher “is very good for learning, but not for human relationships”: Japanese
791 Children’s Critical Perspectives Towards Ethical AI Futures. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.
792 ACM, New York, NY, USA, 1–20.
- 793 [42] Irina Shklovski, Scott D Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. 2014. Leakiness and creepiness in app space: perceptions
794 of privacy and mobile app use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.
- 795 [43] Olivia Sidoti. 2025. 34% of U.S. adults have used ChatGPT, about double the share in 2023. [https://www.pewresearch.org/short-reads/2025/06/25/34-
796 of-us-adults-have-used-chatgpt-about-double-the-share-in-2023/](https://www.pewresearch.org/short-reads/2025/06/25/34-of-us-adults-have-used-chatgpt-about-double-the-share-in-2023/). (June 2025). Accessed: 2025-9-10.
- 797 [44] Mackenzie Sigalos. 2025. OpenAI’s ChatGPT to hit 700 million weekly users, up 4x from last year. [https://www.cnbc.com/2025/08/04/openai-
798 chatgpt-700-million-users.html](https://www.cnbc.com/2025/08/04/openai-chatgpt-700-million-users.html). (Aug. 2025). Accessed: 2025-9-10.
- 799 [45] Marita Skjuve, Petter Bae Brandtzaeg, and Asbjørn Følstad. 2024. Why do people use ChatGPT? Exploring user motivations for generative
800 conversational AI. *First Monday* (Jan. 2024).
- 801 [46] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My chatbot companion - a study of human-chatbot
802 relationships. *Int. J. Hum. Comput. Stud.* 149, 102601 (May 2021), 102601.
- 803 [47] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human-chatbot relationships. *Int. J.*
804 *Hum. Comput. Stud.* 168, 102903 (Dec. 2022), 102903.
- 805 [48] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. 2013. i read my
806 Twitter the next morning and was astonished: a conversational perspective on Twitter regrets. In *Proceedings of the SIGCHI Conference on Human
807 Factors in Computing Systems*. ACM, New York, NY, USA.
- 808 [49] M Speretta and Susan Gauch. 2004. Personalizing search based on user search histories. (2004).
- 809 [50] Andrea L Stone, Linda G Becker, Alice M Huber, and Richard F Catalano. 2012. Review of risk and protective factors of substance use and problem
810 use in emerging adulthood. *Addict. Behav.* 37, 7 (July 2012), 747–775.
- 811 [51] John Suler. 2004. The online disinhibition effect. *Cyberpsychol. Behav.* 7, 3 (June 2004), 321–326.
- 812 [52] Vivian Ta, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, and Alexia Loggarakis. 2020. User
813 experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *J. Med. Internet Res.* 22, 3 (March 2020), e16235.
- 814 [53] Jih-Hsin Tang and Cheng-Chung Wang. 2012. Self-disclosure among bloggers: re-examination of social penetration theory. *Cyberpsychol. Behav.*
815 *Soc. Netw.* 15, 5 (May 2012), 245–250.
- 816 [54] Jennifer Lynn Tanner. 2006. Recentering during emerging adulthood: A critical turning point in life span human development. In *Emerging adults
817 in America: Coming of age in the 21st century*. American Psychological Association, Washington, 21–55.
- 818 [55] Jean M Twenge, Jonathan Haidt, Jimmy Lozano, and Kevin M Cummins. 2022. Specification curve analysis shows that social media use is linked to
819 poor mental health, especially among girls. *Acta Psychol. (Amst.)* 224, 103512 (April 2022), 103512.
- 820 [56] Patti M Valkenburg, Jochen Peter, and Joseph B Walther. 2016. Media effects: Theory and research. *Annu. Rev. Psychol.* 67, 1 (2016), 315–338.
- 821 [57] Jessica Vitak and Jinyoung Kim. 2014. You can’t block people offline: examining how facebook’s affordances shape the disclosure process. In
822 *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, New York, NY, USA.
- 823 [58] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. 2014. A field trial of privacy nudges
824 for facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.
- 825 [59] P Wisniewski. 2018. The privacy paradox of adolescent online safety: A matter of risk prevention or risk resilience? *IEEE Secur. Priv.* 16 (March
826 2018), 86–90.
- 827 [60] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M Carroll. 2015. Resilience mitigates the negative
828 effects of adolescent internet addiction and online risk exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing
829 Systems*. ACM, New York, NY, USA.
- 830 [61] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, and John M Carroll. 2017. Parents just don’t understand: Why teens don’t talk to parents about
831 their online risk experiences. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New
832 York, NY, USA.
- 833 [62] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The dark side of AI companionship: A taxonomy of
834 harmful algorithmic behaviors in human-AI relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM,
835 New York, NY, USA, 1–17.
- 836 [63] Shoshana Zuboff. 2023. The age of surveillance capitalism. In *Social Theory Re-Wired*. Routledge, New York, 203–213.
- 837 Manuscript submitted to ACM