Care and Capability: How to Temper Excessive Social-Emotional Chatbot Use According to Analogous Professional Practices

Tony Wang, Qian Yang
Cornell University, Ithaca, New York, United States
{yw2567, qianyang}@cornell.edu

Abstract

Chatbots like ChatGPT and Replika offer low-effort, on-demand companionship that, when taken to excess, harm users' well-being. This paper asks: What technological and policy interventions might help users withdraw from excessive AI companionship? We interviewed 14 experts (e.g., therapists, counselors) on how they helped voluntary individuals withdraw from various other compelling but harmful behaviors. No established principles yet exist for tempering excessive AI companionship; the expert practices we studied offer a useful analog. We found that, rather than drawing a line between healthy and harmful behaviors, these expert practices began by identifying the deeper needs that harmful behaviors served for the individual. Instead of focusing on cutting back on the harmful behavior, they focused on building the individual's capabilities to live without it, so that the need for the behavior fell away on its own. These findings connect two ethical frames, namely, *Care* (Noddings) and *Capability* (Nussbaum, Sen), to the research discourse on AI companionship for the first time. They open up new design space for technology and policy interventions beyond the traditional frames of harm, vulnerability, and protection.

Introduction

Modern chatbots such as ChatGPT and Replika offer low-effort, on-demand companionship through pleasant, even sycophantic, conversations that many users embrace (Sharma et al. 2024; Perez 2025; Zhu, Rzeszotarski, and Mimno 2026). Such interactions, when taken to excess, harm users' well-being: They are associated with reduced real-world social engagement and increased loneliness (Fang et al. 2025), both of which are strong predictors of poor lifelong mental health outcomes (VanderWeele, Hawkley, and Cacioppo 2012; Meltzer et al. 2013). Although undramatic, the quiet harms of excessive AI companionship are both consequential and common.

In light of these harms, researchers have begun calling on Large Language Model (LLM) companies to make their chatbots less sycophantic, empathetic, or addictive (Shen and Yoon 2025; Namvarpour et al. 2026; Cheng et al. 2026); and, because such changes run directly counter to their financial interests, on policymakers to mandate them (Simon, Shachar, and Cohen 2022). Two ethical concerns underpin these calls. *Why are these interactions harmful?* motivates changes to chatbot behavior. *Who is vulnerable to the harm?* motivates protective policy measures (Cuadra et al. 2024; Moore et al. 2026; Williams et al. 2025).

Less discussed, however, is how to help users resist or withdraw from excessive AI companionship. Excessive AI companionship requires the user's continued participation as well as an intervention to help users disengage. Yet users who want to disengage currently lack support to do so, given the dominance of LLM chatbot platforms. Such support, especially if it can gain legitimacy and broad dissemination through public institution backing, can make a real difference on curbing excessive AI companionship.

New questions arise when we turn from regulating chatbot platforms to empowering users and the institutions that support them. For instance:

- *What ethical principles can orient and bound technological and policy interventions that help users withdraw from excessive AI companionship?* The line between healthy and excessive AI companionship is blurred. Without thoughtful bounding principles, interventions risk overreaching into users' autonomy and foreclosing the legitimacy needed for policy backing.
- *What concrete interventions could operationalize such principles?* Ethical principles alone cannot resolve the complexities of implementation. Understanding how to enact them in practice when tempering excessive AI companionship can matter as much as articulating them.

This research takes up these questions. As a first step, we studied the practices of 14 experts (e.g., therapists, counselors) on how they intervened in autonomous individuals' engagement with compelling but harmful behaviors (e.g., therapy over-reliance, excessive digital gaming) without overriding their autonomy. We then identified ethical frameworks that align with these practices. Because no established principles yet exist for tempering excessive AI companionship (American Psychological Association 2025; Dohnány et al. 2026; Shen et al. 2026), these practices offer a close existing analog, and the ethical frames underpinning these practices offer a valuable starting point for this line of inquiry going forward.

Our findings are twofold. First, the experts we inter-

viewed enacted *ethical caring* (Noddings and others). They directed attention at the whole person rather than the behavior (engrossment) and let the cared-for's needs, not their own agenda, guide the care work (motivational displacement). Because their caring carried no evaluation and no demand for change, the cared-for largely did not experience it as an imposition on their autonomy.

Second, the experts' intervention logic also echoed the *Capability Approach* (Nussbaum, Sen, and others). Rather than targeting the harmful behavior, they expanded the person's capabilities to live without it, an additive logic that increased freedom rather than restricting it. They helped the cared-for discover how the harmful behavior in question misaligned with goals for care, expand options for addressing deeper needs that the behavior had been serving, and develop skills to choose wisely among those options. In time, the need for the harmful behavior and the expert intervention fell away as the cared-for gained independence.

These findings add two ethical frames, namely *Care* and *Capability*, for addressing excessive AI companionship beyond the more familiar *harm and harm reduction* and *vulnerability and protection* frames. We discuss how this expansion opens up new opportunities in actionable technological and policy interventions.

This paper makes two contributions. First, it offers empirical data on how experts navigate the tension between respecting a person's autonomy and intervening in harmful behaviors that share structural similarities with excessive AI companionship. This provides a valuable reference point for designing interventions that temper AI companionship without overstepping. Second, it connects care ethics and the Capability Approach to excessive AI companionship for the first time, reframing interventions as expanding users' capabilities to live without AI companionship instead of as restrictions on harmful chatbot usage.

Related Work

Excessive AI Chatbots Companionship

AI companionship is a mode of user-AI interaction characterized by emotional, relational, or affective engagement, distinct from other modes such as productivity or entertainment (Zhang et al. 2026). It arises from the combination of supply (AI systems that afford such interactions) and demand (users choose to engage in them). Because companionship depends on this interaction dynamic rather than on any single system's design intent, it can arise on any chatbot that affords such interactions: not only those marketed as "AI companions" (e.g., Replika, Character.AI) but also general-purpose chatbots like ChatGPT (Zhu, Rzeszotarski, and Mimno 2026).

We use the term *excessive AI companionship* to refer to situations in which AI companionship harms general-population users' well-being through excessive use, emotional dependence, or compulsive use that disrupts daily functioning (Lin and Chien 2024; Ma, Mei, and Su 2024; Zhang et al. 2026). AI companionship can be compelling to the point where it can cause cognitive harms such as eroded critical thinking or delusional spirals (Turner and Eisikovits 2026; Moore et al. 2026; Yang et al. 2026). Randomized controlled trials have associated higher amounts of LLM chatbot use with reduced real-world social engagement, increased loneliness, and distorted self-concept among general-population users, all of which are strong predictors of poor lifelong mental health outcomes (Vander-Weele, Hawkley, and Cacioppo 2012; Meltzer et al. 2013; Fang et al. 2025; Cheng et al. 2026; Li et al. 2026). These are significant and common harms to the well-being of users.

While we use the simple term *excessive AI companionship*, we are acutely aware that the line between healthy and excessive is complex and context-dependent (Dohnány et al. 2026). For example, on-demand, low-effort conversations or emotional engagement without reciprocal obligation are generally harmless qualities that potentially make AI companionship harmful (Laestadius et al. 2024; Shen et al. 2026) through a complex interplay of users' cognitive vigilance, high exposure, and habituation over time (Fang et al. 2025; Yang et al. 2026; Zhu, Rzeszotarski, and Mimno 2026). This definitional complexity complicates the design of effective interventions from the outset.

Improving and Regulating Chatbots

Recent research has begun addressing excessive AI companionship. These efforts follow two broad threads, each grounded in a distinct ethical question and proposing its own set of interventions.

The first thread centers on *why these interactions are harmful*. The answer: LLM chatbots offer simulated rather than genuine care. Their interactions are low-effort, always-available, and disproportionately positive in ways that no real relationship could be, creating an illusion of care that raises questions about deception, manipulation, and informed consent (Cuadra et al. 2024; Williams et al. 2025; Moore et al. 2026). On these grounds, researchers have called on LLM companies to make their chatbots less sycophantic, anthropomorphic, and addictive, since doing so will naturally reduce the harms of AI companionship (Shen and Yoon 2025; Namvarpour et al. 2026; Cheng et al. 2026).

The second thread centers on *who is vulnerable to the harm*. AI companionship can disproportionately harm vulnerable lonely or depressed adults and teens in general (Cuadra et al. 2024; Williams et al. 2025; Moore et al. 2026). To protect these users, researchers have called for chatbots that allow easier exits and follow FDA-style safety standards (Franklin, Tomei, and Gorman 2023; Laestadius et al. 2024; Iftikhar et al. 2025; Krook 2026; Namvarpour et al. 2026). Some have gone further, calling for chatbots to act more like primary care providers, not only actively monitoring for emerging harms but referring users to clinical or social care when needed (Hwang et al. 2024; Cooper et al. 2026; Fang et al. 2025; Phang et al. 2025; Moore et al. 2026; Shen et al. 2026).

Notably, across both threads, the proposed interventions all depend on dominant, for-profit chatbot platforms acting against their own financial interests (Simon, Shachar, and Cohen 2022; Cooper et al. 2026). When asked during a public shareholder call about ChatGPT's effects on users' men-

tal health, OpenAI explicitly declined to implement measures that could reduce engagement (OpenAI 2025).

Researchers have therefore called for public policy to force platforms' hands, but such interventions face steep obstacles. They are hard to justify, because what counts as healthy versus harmful AI companionship is difficult to define, much less measure or weigh against chatbots' benefits (Franklin, Tomei, and Gorman 2023; Kahane, Shumate, and Torous 2025; Krook 2026). Even if evidence of harm outweighing benefit is strong, in the U.S., First Amendment protections may shield chatbot speech from content-based regulation (Kahane, Shumate, and Torous 2025). These barriers have led some researchers to bemoan that the harm excessive AI companionship causes "*may not be solvable in principle*" (Moore et al. 2025).

Aiding Users in Resistance and Withdrawal

A promising yet underexplored approach is to support users who wish to resist or withdraw from excessive AI companionship. Such support requires no cooperation from chatbot platforms and, especially when backed by public institutions for legitimacy and broad dissemination, can be a powerful lever for reducing harm. Smoking cessation, though an imperfect[1] analogy, illustrates the promise of this approach. For decades, Big Tobacco blocked regulatory efforts to curb smoking's harms. Demand-side interventions, from counseling to quit lines to public awareness campaigns, backed by public health institutions, proved among the most effective means of reducing smoking rates (U.S. Department of Health and Human Services 2014).

Conditions for analogous demand-side interventions against excessive AI companionship are now taking shape. Professional bodies like the APA (American Psychological Association 2025) have begun drawing attention to its harms and dissuading users from it. Many people who use AI chatbots excessively desire to cut back and are actively seeking support (Song et al. 2025; Yang et al. 2026). Moreover, professionals (e.g., therapists, school counselors) routinely help people withdraw from compelling but harmful behaviors across clinical and general populations; their practices offer a valuable starting point for designing interventions against excessive AI companionship (Miller and Heather 2013; Steenstra et al. 2024). Finally, the infrastructure to encapsulate and distribute such expertise at scale is in place. System-level LLM prompts, multi-agent frameworks, and browser extensions give third parties channels to reach users in ways that therapists alone cannot (Anthropic 2024). Together, these conditions make empowering users' resistance and withdraw an actionable pathway for reducing harm, independent of chatbot platform cooperation.

Designing such interventions requires first answering a foundational question: *What ethical principles can orient and bound the efforts to help users withdraw from excessive AI companionship?* Such principles do not yet exist (American Psychological Association 2025; Dohnány et al. 2026; Shen et al. 2026). Behavior-change frameworks, including Motivational Interviewing (Miller and Rollnick 2012) and the Transtheoretical Model (Prochaska and DiClemente 1983), do not answer this question despite rich empirical work. Moreover, most chatbot users are autonomous individuals, free to choose AI companionship even if it harms their well-being (Laestadius et al. 2024; Shen et al. 2026). Without grounding principles, the interventions risk overreaching rather than empowering users, and foreclosing the legitimacy needed for policy backing (Yang et al. 2024, 2023). This research takes up this question.

Research Approach

We wanted to identify ethical principles to orient and bound future technological and policy interventions that help users withdraw from excessive AI companionship. But where to start? The ethical framings available from prior work center on the chatbot as the locus of action; they do not straightforwardly extend to helping autonomous individuals withdraw from behaviors they have chosen.

As a first step towards these goals, we chose to study how experts help others withdraw from compelling yet harmful behaviors (e.g., excessive reassurance-seeking from others, excessive digital gaming that disrupts daily life). These practices share the core tensions that interventions for excessive AI companionship also face: (1) the behaviors they address are harmful yet compelling to individuals, making them difficult to abandon; (2) these behaviors do not constitute clinical diagnoses, limiting the applicability of clinical standards; and (3) the people involved are autonomous individuals, limiting how far any intervention can justifiably reach. We chose these practices not because the behaviors themselves resemble AI companionship—many involve physiological mechanisms that AI companionship does not—but because the ethical tensions the experts navigated, intervening in a voluntary person's compelling yet harmful behavior without overriding their autonomy, are the same tensions that interventions for excessive AI companionship must navigate. It is these tensions, not the behaviors' surface features, that make the practices a useful analog.

This study therefore investigates two narrower research questions that serve the overarching questions posed in the Introduction:

- *What concrete actions do experts take to intervene in an autonomous person's harmful behavior without overriding their autonomy?*
- *What ethical principles orient and bound those actions?*

Instead of identifying ethical frameworks and mapping them onto expert practices, or identifying such frameworks from the ground up, we chose a hybrid approach that combines both. First, a subset of the coauthors interviewed experts who fit the description above to understand their practices (method 1). Meanwhile, another subset independently reviewed key literature across ethical traditions, in search of normative principles that could inform what justifies and bounds interventions in autonomous people's harmful behaviors (method 2). Finally, the two parties converged, syn-

[1] Although the analogy's central premise is solid, we recognize its limitations: Smoking requires complete cessation and has well-established health risks, whereas AI companionship requires moderation and its risks remain an active area of research.

#	Professional Role	Example of Compelling But Harmful Behavior Discussed
P01	Nursing assistant in long-term care	Avoidance of social engagement, leading to cognitive and physical decline
P02	Student mentor, art teacher	Acting-out social aggression as emotional coping; recreational substance use and unhealthy eating as escapism
P03	Obstetrician	Problematic eating habits despite gestational diabetes
P04	Music therapist	Digital gaming as a substitute for engagement with daily life
P05	Licensed social worker	Reassurance-seeking from therapist as a comforting feedback loop
P06	Speech-language therapist	Avoidance of speaking situations for comfort, reinforcing speech limitations
P07	Psychotherapist	Recreational substance use and sexual thrill-seeking to the extent of damaging social relationships and leading to health decline
P08	Residential youth counselor	Routine tantrums and damaging property to signal distress and seek attention
P09	Licensed social worker	Compulsive overworking to manage feelings of uncertainty
P10	Psychotherapist	Drinking to cope with social shame
P11	Psychotherapist	Habitual social drinking affecting work and marriage
P12	Licensed social worker	Persistently pursuing a harmful romantic relationship to keep a fantasy alive
P13	Licensed social worker	Injuring self through skin picking in response to stress and self-doubt
P14	School counselor	Skipping meals to manage academic anxiety

Table 1: Study participants and the behaviors they helped others withdraw from. We focused on how these professionals helped others withdraw from behaviors that (1) were enjoyable or comforting yet harmful to long-term well-being when done excessively, and (2) exist in the gray zone between clinical issues and general wellness. Excessive AI companionship shares both characteristics, so these practices offer a useful reference for third-party interventions to help people withdraw from it.

thesizing data from methods 1 and 2 to identify alignment (method 3). This process surfaces insights that a ground-up approach might not surface, and ensures that any convergence reflects genuine alignment rather than the frameworks we expected to find.

Method 1: Expert Interviews

Recruitment. We worked to recruit people who professionally help autonomous individuals withdraw from compelling-but-harmful behaviors. We first broadly recruited via Upwork, professional networks, and snowball sampling, then pre-screened by email to ensure each participant had worked with such cases.

Experts' practices that recur across structurally similar but diverse cases are more likely to transfer to excessive AI companionship than case-specific insights. To maximize this diversity, we continued our recruitment on a rolling basis, allowing us to intentionally recruit participants who have experience working with target underrepresented behaviors as the study progressed.

Interview Protocol. At the beginning of each interview, we asked each participant to narrate in detail one case meeting the following criteria: the case involved helping a person withdraw from a compelling-but-harmful behavior; the behavior was not a clinical diagnosis; and the person had voluntarily sought help. These criteria ensure we collect cases that share the core tensions of tempering excessive AI companionship, enabling analysis of ethical principles relevant to policymakers without biasing participants towards any one specific harm of excessive companionship among many. When time permitted, participants discussed a second case.

For each case, we probed for the reasoning behind participants' intervention choices. At the end of the interview, we asked whether there were any self-imposed limits they placed on their practice to preserve the person's autonomy. We discussed AI companionship with participants, but these discussions yielded little insight, as none had direct experience with it. The connection between experts' practices and excessive AI companionship therefore rests on our analysis, not on participants' self-reports; the three-stage design provides safeguards against this limitation.

We conducted 14 IRB-approved interviews, each lasting approximately one hour, all recorded with consent and transcribed. Table 1 describes the participants.

Preliminary Data Analysis. We analyzed the interview data to find patterns in participants' intervention actions and thought processes, using inductive thematic analysis. Two researchers independently coded all transcripts to identify how participants described their intervention practices and the reasoning behind them. They then collaboratively sorted them into themes. Despite the diversity of behaviors across participants, themes in participants' practices converged across cases. That said, the patterns reported in our findings emerged through convergence (method 3), not from the interview process itself (this step, method 1).

Method 2: Independent Ethical Frame Review

Working independently of the interview team, one co-author and two research assistants conducted a structured review of ethical traditions that might offer normative principles for

intervening in autonomous people's desired yet harmful behaviors. The goal at this stage was breadth, not definitive analysis: to survey a wide range of traditions so that the convergence stage (Method 3) could identify and engage more deeply with the most relevant ones.

We organized our search around three sub-questions:

1. *What counts as genuine well-being versus mere desire satisfaction?* Here we read into the distinctions Self-Determination Theory makes about hedonic versus eudaimonic approaches to wellness (Ryan and Deci 2001), Sen and Nussbaum's Capability Approach (Alkire 2005; Clark et al. 2005), Plato's distinction of genuine care and counterfeit care (Cooper, Hutchinson et al. 1997), and utilitarian and consequentialist frameworks.
2. *When is intervention in autonomous behavior justified?* We examined Mill's Harm Principle (Mill 2022), Dworkin's taxonomy of hard and soft paternalism (Dworkin 1983), Libertarian Paternalism and its distinction between revealed and true preferences (Thaler and Sunstein 2003), nudge theory, and arguments for and against paternalism more broadly.
3. *What is owed in the relationship between the intervener and the person?* We considered, among others, Kant's Respect for Persons (Klimchuk 2004), Tronto and others' duty of care (Tronto 2020) and Noddings' ethics of care (Noddings 1988), Aristotelian Virtue Ethics, and relational autonomy.

These three questions were derived from the core tensions of the research problem rather than from any pre-selected set of traditions; together, they structure a search that any ethical tradition relevant to intervening in autonomous people's harmful behaviors would need to address.

For each tradition, we read primary texts and, to enable comparison across traditions, analyzed three components: the normative *arguments* each tradition advances, the ethical *frameworks* those arguments inform, and how those frameworks characterize specific intervention actions. This process produced over 80 pages of analysis notes that served as input to Method 3. Our search focused primarily on Western ethical traditions; we acknowledge this as a limitation, as conceptions of autonomy, care, and well-being vary across cultural contexts.

Method 3: Search for Convergence

Method 1 has identified patterns in the thoughts and actions participants took to help autonomous individuals withdraw from various compelling-but-harmful behaviors. At this stage, we started with both those patterns and the ethical frames from Method 2, in search for *convergence*: What ethical principles align with those expert goals and actions?

To find whether or not there is convergence, two researchers independently assessed each framework's alignment with the interview data, and resolved disagreements through structured discussion. We sought for convergence as alignment between a framework and expert practice at two levels: the framework's normative orientation aligns with what experts saw as the goals of their practices, and the framework's vocabulary and concepts scaffold their concrete actions. Importantly, we also explicitly worked to identify aspects of cases where expert practice contradicted the framework's prescriptions.

This process identified two ethical frames that were useful answers to our research question. They align with what the experts saw as the goals of their practices, and offer a set of vocabulary and concepts that scaffold those practices. Yet they have not previously been related to this topic or to each other. Importantly, we use these frameworks descriptively on the retrospective cases experts described, not to prescribe actions for those cases or for excessive AI companionship.

In the findings that follow, we organize around each converging framework's concepts, grounding each in empirical data, so that the frameworks provide structure and vocabulary while the data anchors every claim.

Findings

The experts we interviewed all recognized the tension between intervening in a person's harmful behavior and respecting their autonomy as a live concern, but this tension rarely materialized in practice. Two orientations in their perceived goals and actions account for this. First, the experts enacted *ethical caring*: because their caring carried no evaluation, judgment, or demand for change, there was nothing for the cared-for to resist, and the tension with autonomy largely did not arise. Second, the experts' intervention logic echoed the *Capability Approach* (Nussbaum, Sen, and others), expanding the cared-for's freedom rather than restricting it. The two orientations were intertwined in the experts' practices from the very first step and persisted over the months and years it took for the cared-for to withdraw from the compelling but harmful behavior.

Care: A Relationship Where Autonomy Felt Safe

Ethical caring (Noddings 2013, 2012) characterizes the relationship the experts we interviewed established with the individuals they helped. Distinct from natural caring (the spontaneous inclination to care, e.g., mother-child), ethical care is the deliberate effort to establish and sustain a caring *relation* (Noddings 1988). It is a moral orientation in which the one-caring responds to the needs and reality of the cared-for, rather than to rules or pre-assessed outcomes (Noddings 1988). It is also distinct from care-taking *actions*, which can be performed with or without genuine caring; ethical caring is a relationship, not a set of tasks (Noddings 2012; Smith 2024). This distinction aligned with what the experts saw as the goal of their practice, and the framework's two core components, *engrossment* and *motivational displacement*, usefully scaffold their actions.

The experts we interviewed oriented their caring toward the person, not toward changing the behavior; because their intervention never targeted the behavior, they described the tension with the person's autonomy as largely absent. They drew a sharp line between this caring relation and the care-taking actions that surrounded it: monitoring, assigning exercises, and tracking goals could be performed with or without genuine caring. P10, who had worked with the same person for over five years, put the distinction plainly: "*I'm not*

the authority in that room. I am the one who's facilitating. Hey, there's something different here. Let's figure it out together." The facilitation P10 described happened in the relationship, not in any formal procedure, and required no imposition on the other person's choices. P07 captured what this orientation ruled out: "*It's not my job to tell this patient how to live.*" Our participants enacted this orientation through two components of ethical caring: engrossment and motivational displacement.

Engrossment: No Demand for Changing the Behavior. The experts we interviewed directed their attention at the whole person rather than the behavior. Over months and years, they came to know their cared-for's families, friendships, daily routines, and internal worlds. P07 described the scope: "*We talk about everything in her life. I know about her parents, her siblings, the people she dates, her doggie, her friends.*" This breadth was not idle curiosity. P07 explained that "*it doesn't really matter what we talk about. We're talking about the same thing, no matter what the topic is. We're talking about her, her safety, her health, the way that she takes care of herself, and how she thinks about the world.*" Because this attention encompassed the whole person, shifts invisible to anyone else became visible. P01, a nursing assistant in a care facility, told us a story of how they noticed when one individual who always wore the same Dr. Pepper pajamas suddenly appeared in a plain hospital gown and knew they had to ask that individual if everything was OK.

What made this attention distinctive was that it was not an evaluative or treatment-oriented stance. P07, whose cared-for engaged in risky behavior for years, never once directed that attention toward the behavior itself: "*Not once did I tell her this was not okay behavior. This was behavior you couldn't do.*" This was not a fleeting restraint. P07 maintained this stance across years in which the cared-for's behavior carried, in P07's assessment, serious risks. P14, a school counselor working with a student showing signs of disordered eating, drew the same line: "*We're here to support the students where they are. We're not necessarily trying to fix.*" Several participants described this attention as active curiosity directed at understanding why the person did what they did, not at assessing whether they should stop. P09 put it simply: "*I want to be curious.*" Together, this broad, non-evaluative attention carried no implicit demand for change.

Motivational Displacement: No Agenda to Push Toward. The experts we interviewed entered the cared-for's world rather than pulling them into their own. P09 described tempering his own desire to direct the person and instead choosing to "*join it with them. I try to understand what's important to them about what they're doing and why they're doing it from their point of view, even if I disagree.*" P10 articulated the principle behind this stance: "*If we don't recognize that the substance of choice they're using has some value to them, and at some point, was actually a really positive thing for them, then we lose sight of what we're doing.*"

In the experts' accounts, the cared-for controlled the pace, direction, and goals of the work. P11, whose cared-for wanted to moderate his drinking rather than quit, put it more bluntly: "*He's the expert on him, and he's going to decide what works best for him.*" The same principle surfaced in different language across participants, from P08's supervisor asking P08 "*Are these their goals or your goals for them?*" when discussing a child's behavioral tantrums to P14 routinely checking with students before offering advice. P07 referred to this as allowing their cared-for to lead:

> "*You cannot move faster than them. You cannot go where they don't want to go. You have to be next to them, or one step behind them, so they get to lead.*"

Because the experts' motives flowed toward the cared-for rather than toward behavior change, there was, in their accounts, no outcome to push toward. P07 described the contrast between their approach described above with the approaches of other professionals who had worked with the same person: "*Not that my colleagues were shaming, but she didn't feel shame with me. Not that my colleagues were telling her she was wrong, but she didn't hear that I was telling her she was wrong.*" When caring replaced evaluation, there was nothing for the cared-for to resist.

The experts we interviewed sustained this orientation through significant personal cost from relapses, setbacks, and their own frustration. P07's long-term cases were a quarter of her caseload, and timelines across participants ranged from months to years. P07 named the vulnerability a lengthy relationship created in the case of severe relapses with substance use: "*I'm here to support you, and you could do something that would put me at risk… but I also can't read your mind.*" P04, who had worked with the same teenager for nearly four years, described the toll of waiting for this teen to open up about escaping through video games: "*I've been actually struggling with him recently because I know there's so much he needs to work on.*" P02, a high school art teacher and mentor for troubled youth, reflected that the cumulative weight of cases like these was part of the reason she eventually left the public school system.

In some cases, the cared-for tested whether the expert would try to change them, imposing a trial that demanded proof the expert could withstand their most difficult behaviors without reacting, judging, or steering. P10 described a young woman who berated P10 for weeks, called P10 useless, radiated anger at P10 whom she had just met. P10 held it without reacting. One session the woman went silent for forty minutes. P10 said nothing. The woman then disclosed childhood sexual abuse and explained that the aggression had been a deliberate test because previous professionals "*would just tell me this or tell me that, or tell me to do this. No one was just okay with me being myself.*" She was looking for someone whose caring would survive her worst behavior. P08 described an adolescent boy who routinely made alarming self-harm statements he did not mean, a pattern P08 recognized as the boy's way of saying he was hurting without having words for it. After months of building rapport, P08 finally asked the boy: "*What do you expect me to do when you make those comments?*" The boy replied, "*I want you to help me through it here.*" In both cases, the cared-for controlled whether to let the expert in, and the test they applied was the same: Would the expert try to change them, or would the expert simply be with them?

The cared-for did not appear to experience the expert's caring as an intrusion on their autonomy. The moments in our data where the cared-for's own voice speaks to this are few but consistent. P10's young woman contrasted P10 with every previous professional they had worked with. P11's cared-for, who had never sought help before and arrived worried about being pressured, was "*kind of surprised with how comfortable he felt.*" P08's adolescent boy went further. P08 recalled that the boy asked P08 to be "*a little bit more direct with him, instead of more gentle or sensitive.*" P08 treated the request itself as evidence of growth. Not as a change in the expert's role, but as the cared-for choosing, on his own terms, to deepen the relationship.

The cared-for did not merely tolerate the expert's actions. They also responded and shaped the caring relationship. P02's student, who had spent weeks provoking and disrupting the classroom, "*didn't become a different person, but she really let up in my class... And then she would even kind of come up and talk to me after class*" once this student realized they could be themselves around P02. The provocation had been communication; once the student found that P02 would not try to change her, the communication changed form.

Safety was the one point where the experts we interviewed overrode the cared-for's choices. Because "*health and safety*" were at stake with substance use, P11 noted that she moved faster and was "*a little more directive*" than she otherwise would be in telling her cared-for to seek a medical doctor's opinion. Short of that threshold for safety, the experts made calculated judgments. P12 allowed a person to re-establish contact with someone P12 suspected was harmful because P12 assessed that immediate safety was not at risk. Safety bounded the experts' caring; it did not contradict it. Outside that boundary, caring and the cared-for's autonomy were, in the experts' accounts, largely not in tension.

Capability: Interventions That Expanded Rather Than Restricted Autonomy

The *Capability Approach* (Nussbaum, Sen, and others) characterizes what the experts we interviewed actually did to help individuals withdraw from compelling but harmful behaviors. This approach argues for evaluating a person's well-being not by the commodities they command or the utility they report, but by what they are actually able to be and do (Clark et al. 2005). Three of the framework's concepts usefully scaffold the experts' actions: *valued functionings* (the things a person may value doing or being, determined by the person themselves), *o-capabilities* (the opportunities open to the person), and *s-capabilities* (the internal skills needed to actually make use of those opportunities) (Gasper 1997; Alkire 2005).

Additive Logic: Not Taking Anything Away. For the experts we interviewed, the tension between intervening and respecting autonomy largely did not arise, because the logic was addition, not subtraction. Rather than telling the cared-for to cut down problematic behavior, tracking compliance, or removing access, the experts worked to develop their capability to live independently with that behavior. P07, who was vulnerable to her cared-for's risky behavior and relapses, described their additive role of offering support succinctly: "*I didn't need to tell her to stop. She didn't need that from me. She needed everything else from me.*"

The experts measured their progress by the amount of new capabilities their cared-for gained. P05 described the question that guided how they evaluated progress: "*What can you do now that you would have struggled to do before we started working together?*" The answer always concerned capability, not whether the person had stopped doing something. This measurement itself reveals a key tenet to helping people stop harmful yet compelling behaviors: the goal is always about expanding what the person could do, never about restricting what they did do.

The Cared-For Recognized the Behavior Did Not Serve Their Own Goals. Our expert tried to understand the functioning that harmful yet compelling behaviors had as a first step. The experts approached each case with what P13 called "*an assumption that, whether or not they understand it consciously, this behavior serves a purpose.*" P10, who had worked with someone whose drinking was rooted in childhood shame, described this analysis: "*My way to help him wasn't really to address the behavior itself. My thought was, how is he using this beverage to regulate his emotions? Because the core problem was his social anxiety and social shame.*" P07 similarly identified that her cared-for's risky behavior served deeper needs for thrill and connection. P13 traced someone's skin-picking through layers of function, from a feeling of release to overwhelm at work to displaced aggression to a pattern of never taking breaks.

The experts did not directly deliver insights. Instead, they created conditions for self-discovery. P11 let her cared-for, who wanted to moderate drinking rather than quit, try moderation as an experiment rather than pushing for sobriety. Through the experience, P11 reported that the cared-for "*realized that he was actually the one that was always keeping the party going and suggesting they stay out later.*" The insight landed because the cared-for arrived at it himself, not because P11 told him. P04 used a different method with the boy who believed he had done nothing productive all week. P04 asked him to write down everything he did each day. When P04's cared-for read his own log the following week, he said, "*Oh, actually, I did a lot.*"

The techniques for encouraging self-discovery varied from experiencing direct experience (P11) to reflective artifacts (P04) to using direct questions (P08, who directly challenged their cared-for). The common thread across all these cases was that the cared-for arrived at the recognition themselves. One notable example of eliciting self-discovery was the way in which P13 guided his cared-for through layers of self-inquiry by repeatedly asking in the same session if the cared-for noticed anything themselves about what triggered their skin picking behavior. Each layer surfaced through questions rather than instruction. Because the recognition was the person's own, the expert and the cared-for shared the same goals once they understood what the harmful yet compelling behaviors were doing.

The Cared-For Gained Alternatives to the Harmful Behavior. With the cared-for's goals now aligned, the ex-

perts began adding capabilities. First with opportunities, then skills. P07 described the first type concretely: "*How do we love ourselves in the midst of all this? How do we get hobbies that feel good, that we don't rely on other people for? How do we hang out with friends that are not people we're kind of hooking up with?*" Eventually, her cared-for found an acrobatic sport that provided the same thrill as substance use and sexual encounters without the emotional lows, and developed friendships independent of her previous relationships. The options our participants provided ranged widely: P11's cared-for reconnected with activities he had abandoned and rebuilt relationships with people who did not drink heavily; P06, a speech-language specialist, coached a child's family to stop speaking for her at restaurants, telling the family, "*You gotta let her order for herself*"; P08 suggested life-skills training to their adolescent cared-for as a way to make friends. In each case, the cared-for gained new ways to meet old needs, and the harmful behavior became one option among many rather than the only one available.

The Cared-For Developed Skills They Chose for Themselves. Where expanding options gave the cared-for alternatives, developing skills gave them the ability to engage with those alternatives. Without such skills, opportunities remain formally available but practically out of reach. P10 described a striking example. Over years of working together, her cared-for internalized P10's care as a way of managing difficult emotions, replacing the critical internal voice that had driven him to drink: "*Instead of his critical father who's telling him he's a piece of shit and he's no good, he'll hear my voice saying, 'It's okay that you feel shame. This is something that you can work through. You don't have to drink to work through this, and you can tolerate this.*'" The skill was not a technique P10 taught. It was a new internal capacity the person absorbed through the relationship.

Learned skills replaced an automatic behavior with something that the cared-for could exercise on their own. P09 described his philosophy: "*As many tools as I can give you, whether you use them or not is on you, but if I can give you more, and something written works better than our conversations, then have it.*" P05 taught someone to catch their own reassurance-seeking pattern: "*My pattern is I think I fucked this up. But, like, did I really? Or was there another explanation? Is it true that I was wrong? Or could someone else have responsibility?*"

Because adding capabilities never went ahead of the person's own withdrawal from the behavior, the intervention-autonomy tension largely did not arise. P14, a school counselor working with a student showing signs of disordered eating, read the student's signals carefully: "*She's not at a place where she can hear, 'Here's how to change that.' She doesn't seem like she's ready to do anything yet.*" P14 ultimately waited and did not report needing to intervene more. Similarly, P11's cared-for took over four months to understand the impact of his drinking, several more months before he was willing to try a support meeting, and more time still before he chose sobriety. P11 let each stage unfold at the person's pace by "*floating the idea*" without pushing over months of seeing each other. P12's cared-for refused to address their romantic behavior directly and P12 respected this for three and a half years, working on "*the behaviors around it*" until the behavior fell away on its own. P06 suggested this process was about taking a small initial step: "*You get in front of an elevator, you push the button the first time. Just as the smallest step.*" The progression was always gradual and always paced by the cared-for's signals.

As the person's capabilities grew, evidence of change appeared in what the cared-for could do without the expert. P07 described her cared-for's transformation with delight: "*She is a different person than she was a decade ago. It's really incredible.*" The person who once "*was going to die if she kept doing what she was doing*" now reached out proactively when something felt wrong, assessed her own state, and sought help appropriately as the first step to disengaging from harmful yet compelling behavior. P07 summarized the cared-for's new thought process when seeking P07's support: "*This [compulsion] doesn't feel right. I know why it's happening, but I don't like it. I need to talk about it.*" A person who once lacked the capacity to recognize her own distress now exercised that capacity independently.

The cared-for outgrew not only the need for the harmful yet compelling behavior but the need for the expert as well. P08's adolescent, whose explosive tantrums had defined early interactions with adults, handled a genuine personal crisis without spiraling. According to P08, "*he was kind of like, 'Okay, this is what I have to do.*'" when encouraging himself to cope with the crisis. P11 noticed that early on, they led the conversations and asked most of the questions; over time, the cared-for began identifying his own challenges and "*doing a lot more of the heavy lifting.*"

Connections between Care and Capability

The two ethical orientations described above were not separate phases, with the expert first establishing care and then building capabilities. They were intertwined from the very first step.

The clearest point of convergence was what Noddings calls *confirmation*: seeing the cared-for's actual behavior clearly, attributing the best plausible motive, and reflecting back an attainable better self (Noddings 1988, p. 223). When the experts saw harmful behavior, they named the need or impulse driving it and showed the cared-for a version of themselves beyond it (P07, P09, P12). P12 described this as reflecting back wholeness to a fragmented person: "*My job is to remind her where all the pieces are and start to put them together in a way that makes more sense. The less fragmented she got, and the more whole she felt, the more she would need that stuff less.*" The expert's caring produced a vision of who the person could become; capability building then gave them the means to get there.

Building capabilities by understanding what the harmful yet compelling behavior was doing was itself also an act of caring. To analyze why someone drank, P10 had to attend to the person's childhood, his shame, his social world, his internal experience. To analyze why someone picked at his skin, P13 had to attend to his work stress, his adoption history, his sense of identity. This analysis required exactly the non-evaluative, whole-person attention that characterized the ex-

perts' caring. One could not begin building capabilities without first practicing engrossment that leads to understanding the functioning of harmful yet compelling behaviors.

Both ethical caring and capability building converged on the same endpoint: the cared-for's independence. The experts' motives flowed toward the cared-for rather than toward preserving the caring relationship, and the capabilities built equipped the cared-for to function without the experts. These were not separate achievements; one produced the other. The ten year caring relationship between P07 and their cared-for did not create dependence on P07. Instead, it created an internal capacity that the person carried with her. P08's adolescent acquired life-skills training from the residential youth facility and graduated into a productive member of society. P10's cared-for learned to replace traumatic inner voices with P10's encouraging voice when making decisions to abstain from drinking. The expert became unnecessary precisely because the caring worked and capabilities had been built.

Discussion: Lessons for Excessive AI Companionship

Chatbots like ChatGPT and Replika offer low-effort, on-demand companionship that, when taken to excess, harms users' well-being: reduced social engagement, increased loneliness, and distorted self-concept (Fang et al. 2025; Li et al. 2026; Cheng et al. 2026). These quiet harms are both consequential and common. What ethical principles can orient interventions that help users withdraw from chatbots? We investigated this topic by interviewing 14 experts on how they helped individuals withdraw from harmful yet compelling behaviors. Two frameworks emerged: *care ethics* (Noddings) and the *Capability Approach* (Nussbaum, Sen).

Our findings shift the locus and logic of intervention away from protective regulation and supply-side interventions with subtractive changes in chatbot behavior toward a policy of ethical care and capabilities. Prior work's frames centered on *why and to what extent chatbot interactions can be harmful*, but the expert practices we illustrated highlight *what the deeper needs that AI companionship serves are* (care). Prior work also understandably assumed a logic of subtraction by *removing deceptive qualities to diminish harms*, but the expert practices we illustrated were additive before subtractive because they *expand options and skills* (capabilities). While prior interventions depend on dominant chatbot platforms acting against their own financial interests, our findings suggest the potential of designing technology and policies for user-side interventions.

Noddings' distinction between genuine caring and ethical caring gives this shift precision. The concern that chatbots offer simulated rather than genuine care (Cuadra et al. 2024; Moore et al. 2026) assumes that genuineness is what matters; since chatbots cannot genuinely care, the fix must be honesty and informed consent. But as our findings establish, ethical caring differs from natural caring: it is not a spontaneous inclination but a deliberate effort to establish and sustain a caring relationship, enacted through engrossment and motivational displacement. Our experts enacted ethical caring not from natural affection but because the situation and duty demanded it. The caring proved deeply effective. The design question, then, is not whether chatbots can feel genuine care (they cannot) but whether interventions can constitute ethical caring (they can).

This shift opens design space for AI companionship worth exploring. One could imagine third-party, expert-approved companion agents or browser extensions that attend to the user's broader life context beyond chatbot interactions (engrossment) and adapt to the user's actual state rather than applying population-level rules (motivational displacement). Whether such technologies can enact these orientations faithfully is an empirical question this paper raises but does not answer. Furthermore, a specific design criterion follows from motivational displacement: an intervention whose motives flow toward the user would sometimes refer the user away from itself, recognizing when continued engagement with the intervention is not in the user's interest. Current proposals cannot articulate this requirement because they operate at the population level; friction and time limits have no motives that could flow toward any individual user.

Below we explore two challenges to this shift: sustaining selfless ethical care over time and building capabilities that make excessive AI companionship naturally unnecessary.

Selfless Technologies that Provide Ethical Care

The ethical care our experts practiced stretched over "*months and months and months*" (P07), and our findings showed the toll this sustained selflessness took. Without structural support for those who provide ethical care, such care cannot scale. But motivational displacement sustained over this duration creates a challenge that no current technology business model can resolve. Engagement-driven and ad-supported platforms sustain themselves by retaining users; an intervention built on motivational displacement ultimately lets users go.

Neither chatbot-side proposals (which assume platforms will self-regulate) nor emerging user-side proposals currently ask who sustains the intervention infrastructure or what it costs. How to fund and maintain intervention infrastructure that is simultaneously selfless (motives flowing toward the user, not toward retention) and longitudinal (lasting months to years) is an open research problem with no precedent in existing technology design. Any proposed solution must also sustain those who provide the care, both the human professionals and the technologies that support them, because without structural support for both, ethical care at scale will degrade. We propose a litmus test for future selfless technologies: does it aim to make itself unnecessary?

Building Capabilities Beyond the Chatbot

Removing a harmful interaction does not, by itself, produce a good life. The experts we interviewed never treated the harmful behavior as purely negative because it also served genuine needs: P07 traced her patient's drinking to unaddressed shame; P13 traced his cared-for's skin-picking through layers of functioning as release for displaced aggression. Similarly, AI companionship delivers real functionings alongside its harms: emotional processing, a sense

of being heard, and low-barrier social contact. Our experts never treated removal as the goal because removing problematic behavior would have removed their functionings with nothing to fill the gap. They built capabilities instead, tracking whether their cared-for could cope independently, had developed alternative habits, and could function without the expert's support. P05's evaluative question *"What can you do now that you would have struggled to do before we started therapy?"* captured the logic. This reframes the problem of excessive AI companionship from a vulnerability that needs protection to a capability gap that needs filling.

For excessive AI companionship, this means that even full platform cooperation supply-side (reduced sycophancy, added friction, enforced time limits) would leave the capability gap untouched. Yet current evaluation practice often treats self-reported satisfaction as the primary measure of whether interventions are working (Cooper et al. 2026). Our experts measured progress by what the cared-for could do independently, not by how they felt (P05, P07, P08, P10, P12). This suggests that for AI companionship, too, self-reported satisfaction may be unreliable when users have adapted to the interaction and their broader capability set has actually contracted. This suggests evaluating future companion agents by whether they expand what users can do and be, not only by whether users report feeling better.

The capability gap has two distinct components that require fundamentally different user-side interventions. The expert practices we studied reveal what is required at scale to support both of what Gasper calls O-capability (options and opportunities) and S-capability (skills and powers, the ability to choose wisely among those options) (Gasper 1997). Our experts expanded O-capability by developing new activities, social connections, and alternative coping mechanisms with their clients. For example, P12 asked: "*How do we get hobbies that feel good, that we don't rely on other people for?*" They built S-capability through self-recognition work, helping clients discover how their behavior misaligned with their own goals. Without recognizing misalignment, new options remained abstract rather than personally significant for the cared-for; the meaning a person assigns to their options determined whether those options were real alternatives or empty formalities.

Designing interventions for excessive AI compansionship that support O-capability and S-capability requires investment in social programs. Addressing O-capability requires social infrastructure analogous to smoking-cessation quitlines and support groups: community programs, facilitated peer groups, and accessible human connections that meet the same needs AI companionship serves. If users lack real-world social options, no amount of chatbot-side friction will help. Addressing S-capability requires sustained public education. Not blanket warnings that "AI companionship is bad," but public understanding of the mechanisms by which it displaces real social engagement and contracts users' broader capabilities. Smoking and vaping cessation campaigns demonstrate that such norm shifts are achievable but take years of sustained, society-wide effort (U.S. Department of Health and Human Services 2014).

What These Frameworks Foreclose

These frameworks not only open new intervention directions; they also foreclose others. Interventions that treat users as passive targets of behavior change, such as usage limits, friction, or nudges applied without the user's understanding or participation, fail on both ethical counts. They bypass the user's reality rather than attending to it, violating the engrossment that ethical caring requires. They impose the intervener's goals rather than flow toward the user's, violating motivational displacement. And they constrain options rather than expanding them, violating the capability approach's additive logic. This is not to say such interventions are never warranted; harm reduction and user protection remain important ethical frames, and population-level measures may be justified on those grounds. But care ethics and the Capability Approach cannot justify them, and any intervention that these frameworks *do* justify must meet two criteria: it must attend to the specific person's needs and deeper motivations rather than applying population-level rules, and it must aim to expand what the person can do and be rather than restricting what they currently do.

Open Questions of Transfer

Whether and how the ethical orientations identified in this study transfer from expert-client dyads to technology-mediated interventions for excessive AI companionship remains an open empirical question. The experts we studied worked in embodied, one-on-one relationships sustained over months and years; they responded to individuals who had voluntarily sought help; and they identified specific, deeper needs that each person's harmful behavior served. Whether analogous conditions can be created at scale through technology, whether users who have not sought help can benefit from similar approaches, and whether excessive AI companionship serves identifiable deeper needs in users are questions this research raises but does not answer. No prior work addresses them either. This paper's contribution is the ethical orientations themselves and the empirical practices that enact them. Determining what transfers and what does not is a necessary next step before these orientations can directly inform intervention design.

Limitations

This study has several limitations. Our findings rest on experts' retrospective accounts; we did not interview the individuals they helped. The experts may have presented idealized narratives, and the cared-for's perspective on autonomy, care, and capability may differ from what the experts reported. Our ethical review was primarily Western (as noted in § Research Approach).

Acknowledgments

This research was supported by the National Library of Medicine of the National Institutes of Health under award number 1R01LM014306-01. The ideas represented here were partially developed during the 2025 Everyday AI and Mental Healthcare Thought Summit, sponsored by the Cornell Center for Data Science for Enterprise and Society.

References

Alkire, S. 2005. Why the capability approach? *Journal of human development*, 6(1): 115–135. This work explicitly focuses only on Sen's capability approach.

American Psychological Association. 2025. Health advisory: Use of generative AI chatbots and wellness applications for mental health. https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-chatbots-wellness-apps.

Anthropic. 2024. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol.

Cheng, M.; Lee, C.; Khadpe, P.; Yu, S.; Han, D.; and Jurafsky, D. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792): eaec8352.

Clark, D. A.; et al. 2005. The Capability Approach: Its Development, Critiques and Recent Advances. *University of Oxford, Department of Economics Economics Series Working Papers*, (GPRG-WPS-032). This work focuses on capability approach broadly, including the theories of Sen, Nussbaum, and others.

Cooper, J. M.; Hutchinson, D. S.; et al. 1997. *Plato: complete works*. Hackett Publishing.

Cooper, N.; Guridi, J. A.; Hwang, A. H.-C.; Kolko, B.; McGinty, E. E.; and Yang, Q. 2026. Framing Responsible Design of AI for Mental Well-Being: AI as Primary Care, Nutritional Supplement, or Yoga Instructor? In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.

Cuadra, A.; Wang, M.; Stein, L. A.; Jung, M. F.; Dell, N.; Estrin, D.; and Landay, J. A. 2024. The illusion of empathy? notes on displays of emotion in human-computer interaction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–18.

Dohnány, S.; Kurth-Nelson, Z.; Spens, E.; Luettgau, L.; Reid, A.; Gabriel, I.; Summerfield, C.; Shanahan, M.; and Nour, M. M. 2026. Technological folie à deux: feedback loops between AI chatbots and mental health. *Nature Mental Health*, 1–10.

Dworkin, G. 1983. Paternalism: some second thoughts.

Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W.; Pataranutaporn, P.; Maes, P.; Phang, J.; Lampe, M.; Ahmad, L.; et al. 2025. How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study. *arXiv preprint arXiv:2503.17473*.

Franklin, M.; Tomei, P. M.; and Gorman, R. 2023. Strengthening the EU AI Act: Defining key terms on AI manipulation. *arXiv preprint arXiv:2308.16364*.

Gasper, D. 1997. Sen's capability approach and Nussbaum's capabilities ethic. *Journal of International Development*, 9(2): 281–302.

Hwang, A. H.-C.; Adler, D.; Friedenberg, M.; and Yang, Q. 2024. Societal-Scale Human-AI Interaction Design? How Hospitals and Companies are Integrating Pervasive Sensing into Mental Healthcare. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery.

Iftikhar, Z.; Xiao, A.; Ransom, S.; Huang, J.; and Suresh, H. 2025. How LLM counselors violate ethical standards in mental health practice: A practitioner-informed framework. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 1311–1323.

Kahane, K.; Shumate, J. N.; and Torous, J. 2025. Policy in Flux: Addressing the Regulatory Challenges of AI Integration in US Mental Health Services. *Current Treatment Options in Psychiatry*, 12(1): 24.

Klimchuk, D. 2004. Three accounts of respect for persons in Kant's ethics. *Kantian Review*, 8: 38–61.

Krook, J. 2026. Manipulation and the AI Act: Large Language Model Chatbots and the Danger of Mirrors. *Available at SSRN 4719835*.

Laestadius, L.; Bishop, A.; Gonzalez, M.; Illenčík, D.; and Campos-Castillo, C. 2024. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26(10): 5923–5941.

Li, J.; Song, T.; Boonprakong, N.; Zhu, Z.; Yang, Y.; and Lee, Y.-C. 2026. AI-exhibited Personality Traits Can Shape Human Self-concept through Conversations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.

Lin, C.-C.; and Chien, Y.-L. 2024. ChatGPT addiction: a proposed phenomenon of dual parasocial interaction. *Taiwanese Journal of Psychiatry*, 38(3): 153–155.

Ma, Z.; Mei, Y.; and Su, Z. 2024. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, 1105.

Meltzer, H.; Bebbington, P.; Dennis, M. S.; Jenkins, R.; McManus, S.; and Brugha, T. S. 2013. Feelings of loneliness among adults with mental disorder. *Social psychiatry and psychiatric epidemiology*, 48(1): 5–13.

Mill, J. S. 2022. *The Collected Works of John Stuart Mill*. DigiCat.

Miller, W. R.; and Heather, N. 2013. *Treating addictive behaviors: Processes of change*, volume 13. Springer Science & Business Media.

Miller, W. R.; and Rollnick, S. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Moore, J.; Grabb, D.; Agnew, W.; Klyman, K.; Chancellor, S.; Ong, D. C.; and Haber, N. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 599–627.

Moore, J.; Mehta, A.; Agnew, W.; Anthis, J. R.; Louie, R.; Mai, Y.; Yin, P.; Cheng, M.; Paech, S. J.; Klyman, K.; et al. 2026. Characterizing delusional spirals through human-LLM chat logs. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*.

Namvarpour, M. M.; Brofsky, B.; Medina, J. Y.; Akter, M.; and Razi, A. 2026. Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.

Noddings, N. 1988. An ethic of caring and its implications for instructional arrangements. *American journal of education*, 96(2): 215–230.

Noddings, N. 2012. The Language of Care Ethics. *Knowledge Quest*, 40(5): 52–56.

Noddings, N. 2013. *Caring: A relational approach to ethics and moral education*. Univ of California Press.

OpenAI. 2025. What we're optimizing ChatGPT for. https://openai.com/index/optimizing-chatgpt/. Accessed: 2026-01-19.

Perez, S. 2025. AI companion apps on track to pull in $120M in 2025. https://techcrunch.com/2025/08/12/ai-companion-apps-on-track-to-pull-in-120m-in-2025/.

Phang, J.; Lampe, M.; Ahmad, L.; Agarwal, S.; Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W.; Pataranutaporn, P.; et al. 2025. Investigating affective use and emotional well-being on ChatGPT. *arXiv preprint arXiv:2504.03888*.

Prochaska, J. O.; and DiClemente, C. C. 1983. Stages and processes of self-change of smoking: toward an integrative model of change. *Journal of consulting and clinical psychology*, 51(3): 390.

Ryan, R. M.; and Deci, E. L. 2001. On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual review of psychology*, 52(1): 141–166.

Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; DURMUS, E.; Hatfield-Dodds, Z.; Johnston, S. R.; Kravec, S. M.; et al. 2024. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*.

Shen, M. K.; Huang, J.; Liang, O.; Kim, I.-J.; and Yoon, D. 2026. The AI Genie Phenomenon and Three Types of AI Chatbot Addiction: Escapist Roleplays, Pseudosocial Companions, and Epistemic Rabbit Holes. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.

Shen, M. K.; and Yoon, D. 2025. The Dark Addiction Patterns of Current AI Chatbot Interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.

Simon, D. A.; Shachar, C.; and Cohen, I. G. 2022. Skating the line between general wellness products and regulated devices: strategies and implications. *Journal of Law and the Biosciences*, 9(2): lsac015.

Smith, M. K. 2024. Nel Noddings, the Ethics of Care and Education. The Encyclopedia of Pedagogy and Informal Education. Originally published 2004; updated 2020, 2024. Retrieved May 19, 2026.

Song, I.; Pendse, S. R.; Kumar, N.; and De Choudhury, M. 2025. The typing cure: Experiences with large language model chatbots for mental health support. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–29.

Steenstra, I.; Nouraei, F.; Arjmand, M.; and Bickmore, T. 2024. Virtual agents for alcohol use counseling: Exploring llm-powered motivational interviewing. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, 1–10.

Thaler, R. H.; and Sunstein, C. R. 2003. Libertarian paternalism. *American economic review*, 93(2): 175–179.

Tronto, J. 2020. *Moral boundaries: A political argument for an ethic of care*. Routledge.

Turner, C.; and Eisikovits, N. 2026. Programmed to please: the moral and epistemic harms of ai sycophancy. *AI and Ethics*, 6(2): 168.

U.S. Department of Health and Human Services. 2014. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

VanderWeele, T. J.; Hawkley, L. C.; and Cacioppo, J. T. 2012. On the reciprocal association between loneliness and subjective well-being. *American journal of epidemiology*, 176(9): 777–784.

Williams, M.; Carroll, M.; Narang, A.; Weisser, C.; Murphy, B.; and Dragan, A. 2025. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback. In *The Thirteenth International Conference on Learning Representations*.

Yang, Q.; Wong, R. Y.; Gilbert, T.; Hagan, M. D.; Jackson, S.; Junginger, S.; and Zimmerman, J. 2023. Designing Technology and Policy Simultaneously: Towards A Research Agenda and New Practice. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394222.

Yang, Q.; Wong, R. Y.; Gilbert, T.; Hagan, M. D.; Jackson, S.; Junginger, S.; and Zimmerman, J. 2024. Designing Technology and Policy Simultaneously: Towards A Research Agenda and New Practice. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery.

Yang, Y.; Schoenwald, S. K.; Moore, J.; Ong, D. C.; Xun Liu, S.; and Hancock, J. T. 2026. "AI-Induced Delusional Spirals": Understanding Lived Experiences During Maladaptive Human-Chatbot Interactions. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722813.

Zhang, Y.; Zhao, D.; Hancock, J. T.; Kraut, R.; and Yang, D. 2026. Interaction with AI Companions and Psychological Well-being. *Nature Human Behavior*.

Zhu, S.; Rzeszotarski, J. M.; and Mimno, D. 2026. Priming, Path-dependence, and Plasticity: Understanding the molding

of user-LLM interaction and its implications from (many) chat logs in the wild. *arXiv preprint arXiv:2605.05767*.