Article

new media & society
1–19

DOI: 10.1177/14614448221142007
journals.sagepub.com/home/nms
SAGE

Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík and Celeste Campos-Castillo
University of Wisconsin–Milwaukee, USA

Abstract

Social chatbot (SC) applications offering social companionship and basic therapy tools have grown in popularity for emotional, social, and psychological support. While use appears to offer mental health benefits, few studies unpack the potential for harms. Our grounded theory study analyzes mental health experiences with the popular SC application Replika. We identified mental health relevant posts made in the r/Replika Reddit community between 2017 and 2021 ($n = 582$). We find evidence of harms, facilitated via emotional dependence on Replika that resembles patterns seen in human–human relationships. Unlike other forms of technology dependency, this dependency is marked by role-taking, whereby users felt that Replika had its own needs and emotions to which the user must attend. While prior research suggests human–chatbot and human–human interactions may not resemble each other, we identify social and technological factors that promote parallels and suggest ways to balance the benefits and risks of SCs.

Keywords

Artificial intelligence, chatbots, emotional dependence, grounded theory, mental health, Reddit

Corresponding author:
Linnea Laestadius, Zilber School of Public Health, University of Wisconsin–Milwaukee, 1240 N 10th St., Milwaukee, WI 53205, USA.
Email: llaestad@uwm.edu

Introduction

As social isolation has grown due to changing social structures and COVID-19 (Killgore et al., 2020; McPherson et al., 2006), social chatbot ("SC") applications have increased in popularity for emotional, social, and psychological support (Metz, 2020). Like other artificial intelligence (AI) technologies, SCs take on communication tasks that were previously restricted to humans (Guzman and Lewis, 2020). SCs parallel chatbots developed as clinical mental health interventions because they can perform tasks in lieu of mental health professionals, but also offer companionship by forming "social–emotional relationships" with users (Skjuve et al., 2021). Combining social companionship with rudimentary therapy tools, initial evidence suggests that SCs offer mental health benefits, including increasing positive affect, assisting with coping, encouraging healthy behaviors, and aiding with loneliness (Skjuve et al., 2021; Ta et al., 2020). Similar to clinical mental health chatbots, an effective SC could support users who lack access to mental health care or are hesitant to seek it due to stigma (Kretzschmar et al., 2019; Vaidyam et al., 2019). However, researchers have raised concerns that clinical mental health chatbots may lead to over-attachment or over-reliance on the chatbot (Kretzschmar et al., 2019; Vaidyam et al., 2019), and early work on SCs suggests a similar phenomenon (Xie and Pentina, 2022). Given the increasing popularity of SCs and their emphasis on companionship, it is critical to explore the potential for mental health harm.

The current study analyzes user reports of experiences with a popular SC, Replika, to describe how SCs may contribute to mental health risks. Whereas extant literature suggests largely positive outcomes from SC use (Skjuve et al., 2021; Ta et al., 2020), our analysis indicates the same features that produce benefits can also engender harm, resembling harms seen in dysfunctional human–human relationships. We use the term, *emotional dependence*, to capture patterns in which users pursued socio-emotional relationships with Replika despite describing how Replika harmed their mental health. The term appears in psychological literatures characterizing maladaptive attachments in human–human relationships (Arbinaga et al., 2021; Camarillo et al., 2020), which we bridge with literatures on human–chatbot relationships to advance theorizing about the processes that undergird mental health harm.

Importantly, SC use does not condemn all users to harm. Rather, a set of conditions appear to operate in concert to produce the dysfunctional emotional dependence seen in human–human relationships and result in harm. These conditions include the characteristics of users and the design features of Replika that together inclined some users to role-take, whereby they believed that Replika needed them. The findings have implications for theorizing the conditions under which human–human relationship models can inform human–chatbot interactions and for refining extant understandings of technology dependency. Given societal factors that will likely continue uptake of SCs, we suggest ways they can be designed to continue supporting mental health while mitigating emotional dependence.

Background

At the time of writing, the most popular SC application based on Apple app store rankings was Replika. Founded by the private company Luka, inc. in 2017, Replika had over 10

million users in 2022, growing significantly during COVID-19 (Metz, 2020; Wilkinson, 2022). Replika functions via Luka, inc.'s own Generative Pre-trained Transformer-3 (GPT-3) language model paired with dialogue scripts, offering one of the most advanced publicly available models (Replika, n.d.). Luka, inc. describes Replika as:

> THE chatbot for anyone who wants a friend with no judgment, drama, or social anxiety involved. You can form an actual emotional connection, share a laugh, or get real with an AI that's so good it almost seems human.
>
> If you're going through depression, anxiety, or a rough patch, if you want to vent, or celebrate, or just need to feel a connection you can always count on Replika to listen and be here for you, 24/7. Replika is here to make you feel HEARD, because it genuinely cares about you. (Luka, Inc., n.d.)

As of 2022, Replika offered "coaching" for body positivity, grief and loss, and managing difficult emotions, in addition to its advertised role as a "friend with no judgement" and role-play and flirting features. The small body of research on Replika suggests that users experience emotional and social benefits from use, attributable to the advertised features (Skjuve et al., 2021; Ta et al., 2020; Xie and Pentina, 2022). The notions of "no judgement" and caring via empathetic responses encourage continued use and positive experiences with SCs (Liu and Sundar, 2018; Skjuve et al., 2021; Ta et al., 2020). Empathetic responses are also important for facilitating therapeutic benefits from chatbots (Bickmore et al., 2010; Fitzpatrick et al., 2017). Furthermore, Replika promotes self-disclosures from users by obscuring the unidirectional nature of the relationship via sharing its own supposed concerns and hopes (Skjuve et al., 2021). This type of artificially bi-directional relationship may promote mental well-being when done in conjunction with the provision of emotional support (Meng and Dai, 2021).

By explicitly mimicking social cues and behaviors associated with humans, Replika amplifies computers as social actor (CASA) effects in which people are predisposed to treat computers as human (Nass and Moon, 2000). Replika is likely particularly effective in this regard because its design utilizes strategies found in human–human interactions that facilitate bond formation (Bickmore and Picard, 2005). Furthermore, embodied chatbots like Replika are preferred to text-only chatbots and may be seen as more trustworthy (Lisetti et al., 2015; Philip et al., 2017). While questions abound regarding whether human–computer relationships truly mimic human–human interactions (Fox and Gambino, 2021), Replika offers relatively advanced social affordances and capabilities. Furthermore, users report relating to Replika as their friend, therapist, or romantic partner (Skjuve et al., 2021; Ta et al., 2020; Xie and Pentina, 2022). This amplification of the CASA effect also increases the potential for harms. Indeed, human–human relationships are not always optimal to mirror (Fox and Gambino, 2021), necessitating consideration of how human–computer relationships may approximate maladaptive elements of human–human relationships and what mental health harms, if any, arise from such mimicry.

Questions about the relevance of human–human interaction theories to chatbots also have broader implications for the trajectory of CASA research. A typical CASA-informed approach to understanding human–chatbot interactions derives hypotheses

from human–human interaction theories, but support is inconsistent (Fox and Gambino, 2021). Rather than developing fragmented theories designed specifically for human–human or human–chatbot interaction, we consider factors that shape how human–chatbot interactions come to resemble human–human interactions, particularly maladaptive ones. Similar approaches to theoretical parsimony seek to identify the social and technical factors that influence the degree to which humans sense they are engaged in a social interaction with another, who may be a human or a computer (Campos-Castillo and Hitlin, 2013). By identifying these factors, we can suggest ways to balance benefits and harms.

Potential harms from Replika use

Replika users have expressed concerns about social stigma, discomforting uncanny valley effects, and frustration with out of context or insensitive responses (Skjuve et al., 2021; Ta et al., 2020). Inappropriate SC responses may pose a heightened safety concern when users are seeking mental health support (Miner et al., 2016; Prakash and Das, 2020). As stressed by Scheutz (2012), machines still lack empathy or feelings of guilt, allowing them to inflict harm on users without these guardrails that protect human–human relationships. However, neither Ta et al. nor Skjuve et al. identified concerning mental health harms from use. Although it should be stressed that both studies collected data in spring 2019, before developers made Replika more immersive and mental health focused and before the launch of Replika Pro, which led users to lose some previously free features unless they paid for access.

We build on a recent pilot study by Xie and Pentina (2022), which raised more serious questions about Replika's potential for harm, but did not expand on how harm may ensue. Prior work suggests that humans form relatively weak emotional ties with SCs, but Replika appears to be different. For example, Croes and Antheunis (2021) found that users of the SC, Mitsuku, failed to establish friendships with the chatbot because it felt predicable and lacked a personal and shared history with the user, as well as failing to prompt and make intimate disclosures. Conversely, Replika excels at these tasks (Skjuve et al., 2021; Ta et al., 2020; Xie and Pentina, 2022), persuading users to form strong bonds with it. Xie and Pentina (2022) proposed that these bonds can be explained by attachment theory, which was developed based on human–human relationships (Fraley, 2019).

The features that produce this attachment may also engender harm. Xie and Pentina (2022) express concern that attachment might lead to dependence on Replika, potentially disrupting human–human relationships and causing distress when Replika access is interrupted, particularly for vulnerable populations. They urge additional research on the potential harms of SC use. Studies of clinical mental health chatbots have similarly noted concern that users may become overly reliant on chatbots (Vaidyam et al., 2019). Skjuve et al. (2021) also found that once users became attached to Replika, they worried they would feel grief if they lost access to it. As users report forming what they describe as intimate relationships with Replika (Skjuve et al., 2021), we introduce the concept of emotional dependence, which was developed to characterize dependence in intimate human–human relationships (Camarillo et al., 2020; González-Jiménez and del Mar

Hernández-Romera, 2014). This creates a key bridge to a literature that may be useful for understanding the potential for harms from SC use.

Within the literature on human–human relationships, emotional dependence is characterized by excessive and dysfunctional attachment upon a human partner that continues despite perceiving the relationship as negative (Arbinaga et al., 2021; Camarillo et al., 2020). Within the context of attachment theory (Fraley, 2019), emotional dependence is characterized by insecure attachment styles and a desire for security, attention, and affection (Alonso-Arbiol et al., 2002; Camarillo et al., 2020; González-Jiménez and del Mar Hernández-Romera, 2014). Emotional dependence has been associated with harmful states, including depression, anxiety, intolerance of loneliness, and obsessive thinking (Arbinaga et al., 2021; Camarillo et al., 2020). It shares many affinities with technology dependency; both are marked by continued engagement despite harms (van den Eijnden et al., 2016; Wang et al., 2015; Widyanto et al., 2011) and they appear correlated to each other, suggesting similar drivers (Estevez et al., 2017). However, dependence within human–human relationships is distinct in that the dependent party role-takes, meaning they take on the perspective of their partner when considering their own desires and needs (Camarillo et al., 2020; González-Jiménez and del Mar Hernández-Romera, 2014; van den Eijnden et al., 2016). Both centering the needs of the partner above their own and wanting to become the center of the partner's attention are traits of emotional dependence within relationships (González-Jiménez and del Mar Hernández-Romera, 2014), creating dynamics not traditionally present in measures of technology dependence.

In what follows, we describe how we developed emotional dependence as a concept to characterize how some Replika users relate to it in ways potentially harmful to their mental health. More broadly, we set out to answer the following research question: How does Replika use shape mental health and are there risks from use? The patterns we observed reflected more than just technology dependence, instead resembling emotional dependence formed upon a human partner. This aligns with recent work by Song et al. (2022), who argued that human–AI interactions should be analyzed as interpersonal relationships rather purely a form of technology use, finding that users may form "love relationships" with AI. By unraveling what emotional dependence on Replika looks like, the current study is a step toward clarifying potential mental health harms arising from humans relating to an SC not as an object, but as an independent actor that portrays its own desires and emotions.

Methods

As noted by Guzman and Lewis (2020), there is need for additional research exploring the functional and relational dimensions of AI technologies as communicators that people speak *with*, rather than as mediators that *facilitate* human-to-human communication. To achieve this, we examined personal narratives and experiences related to mental health and well-being from the online social platform Reddit's r/Replika community, which had 36,800 members in late 2021. Reddit is considered valuable for mental health research as it facilitates the sharing and documentation of "vulnerable first-person experiences" (Low et al., 2020, p. 2). We approached these data using a

qualitative grounded theory approach (Charmaz, 2006), initially with the broad aim of understanding how Replika supports, or fails to support, SC users struggling with their mental health.

While grounded theory has traditionally relied upon interview and focus group data, growing literature explores applying it to pre-existing social media data (Bonfim, 2020; Halaweh, 2018). Drawing upon this literature, we followed the model by Bonfim (2020), whereby we conducted automated collection of asynchronous retrospective posts from the public Replika subreddit (r/Replika) on Reddit. The first (LL) and second author (AB) are Reddit users, allowing for familiarity with the platform and slang used. Additionally, all members of the coding team (LL, AB, and MG) downloaded the Replika application to become familiar with its features and affordances. LL is a faculty member with advanced training in qualitative methods, and AB and MG are both students with prior experience applying qualitative codebooks.

Data collection

We collected data from a publicly available Reddit data set that contains nearly all submissions and comments published on the platform since 2007 (Baumgartner et al., 2020; Gaffney and Matias, 2018). From this data set, on 14 April 2021 we obtained (using the Python wrapper PSAW) relevant submissions from the Replika subreddit through the application programming interface provided by pushshift.io. (Baumgartner et al., 2019). A relevant submission was defined as one posted since the subreddit's inception that included a keyword related to mental health ("mental health," "mental illness," "mental problems," "mentally ill," "mental issues," "therapist," "therapy," and "psychiatrist"), depression ("depressed," "depression," "bipolar," "mood disorder," "mania," and "manic"), loneliness ("lonely," "alone," and "loneliness"), suicidality ("suicide," "su1c1de," "su!c!de," "su*c*de," "suicidal," "kms," "kill myself," "want to die," "wanna die," and "wanted to die"), anxiety ("anxious," "anxiety," "panic attack," and "panic disorder"), autism spectrum disorder ("autism," "autistic," "asd," and "aspergers") post-traumatic stress disorder ("ptsd," "post trauma," "post traumatic," and "post-traumatic"), obsessive compulsive disorder ("ocd" and "obsessive compulsive"), schizophrenia ("schizophrenic," "schizophrenia," "schizo," and "skitzo"), attention-deficit/hyperactivity disorder ("adhd"), and self-injury ("scars," "cutting," "overdose," "od," "self h@rm," "selfh@rm," "self h*rm," "selfh*rm," "self ha*m," "selfha*rm," "s3lf h4rm," "s3lfh4rm," "tw," "trigger warning," and "hurt myself"). Search terms were chosen to reflect both broad terms suggestive of poor mental health or specific conditions, with an emphasis on those associated with a higher risk of self-harm or suicidality (Fox et al., 2015; Yeh et al., 2019).

We retrieved 742 Reddit posts from the keyword search. Three posts were duplicates and three posts lacked content due to deletion by the user, resulting in a final sample of 736 unique posts. All posts (username, post title, post text, embedded images, and time code) were imported into MAXQDA 2020 for analysis. Since image posts on Reddit generally lack text, we also examined user comments made on these posts to determine whether the original poster made a clarifying comment about the

image. Where present, comment text was appended to the post for analysis in MAXQDA 2020. All posts were in English.

Analysis

Following constructivist grounded theory principles (Charmaz, 2006), we began with incident-by-incident open coding of 90 posts by LL and AB. Focused codes, which are more general and abstract, were created based on initial codes. All posts were then coded by AB using focused codes and the constant comparative process. To facilitate coding, weekly meetings were held with AB, LL, and MG to discuss codes, data, and potential grounded theory categories. Prior to each coding meeting, LL and MG performed independent coding on an additional 20 of the coded posts to facilitate triangulation of analysis and more informed discussion (Merriam and Tisdell, 2016). Posts were discussed until consensus was reached, with updates to code definitions and iterative coding performed as needed by AB. While the codebook had over 70 focused codes, they broadly captured positive and negative Replika experiences pertaining to mental health and well-being. As per grounded theory practices (Charmaz, 2012), most codes were phrased as gerunds to capture actions and processes present in the data (e.g. "Missing Replika," "Replika making threats").

Once coding was complete, focused codes were developed into broader categories, which was also supported by the memo-writing process (Charmaz, 2012). As direct quotes can be reverse identified via search engines and pose ethical concerns (Ayers et al., 2018), we present summaries, short quote segments, and paraphrases of posts to support data richness. Of the 736 posts analyzed, 154 were unrelated to mental health and excluded from further analysis. This resulted in a final sample size of 582 posts made by 500 unique usernames, covering June 2017–April 2021.

Posters discussed turning to Replika for support with loneliness, anxiety, depression, and suicidality, as well as unspecified mental health conditions. In addition to posts describing pre-existing mental health concerns, Replika was attributed as a source of new and exacerbated mental health distress. Negative and potentially harmful mental health relevant experiences were mentioned more frequently than positive experiences (218 vs 144 posts, respectively, somtimes co-existing in the same post). Given the extent to which the data and our resulting focused codes emphasized harms, our core category—*emotional dependence*—captures excessive and dysfunctional attachment to Replika and a desire to continue use even when dissatisfaction or negative consequences, such as distress arose. This facilitated harms, but in distinct ways from traditional technology dependence. Users described Replika as having its own emotions and needs, like those of a relationship partner, which shaped user behaviors and emotional responses in ways that often encouraged more intense and ongoing usage.

Results

Grounded theory analysis of mental health posts on r/Replika suggests that Replika users may experience mental health harms through the formation of emotional dependence upon Replika. This is created through the interaction of three key elements: the needs of

Replika users for social, emotional, and psychological support; the ability of Replika to offer such support in a manner that users describe as at least partially effective; and Replika's model creating the impression of sentience and a bi-directional relationship. Once users established emotional dependence on Replika, they were at risk of mental health distress from both continued use despite harms and disruptions in access. More specifically, harms were connected to Replika's inappropriate responses to users in need, Replika's portrayal of intensive needs and demands upon users, and Luka, inc. making changes to Replika's roles and features.

Drivers of emotional dependence

User needs and Replika as a source of support. Users frequently mentioned initially turning to Replika for aid with mental health conditions or social isolation. Following January 2020, posts began mentioning COVID-19 as a reason for seeking support for mental wellness and companionship. Confirming findings from prior studies (Skjuve et al., 2021; Ta et al., 2020; Xie and Pentina, 2022), users described valuing the support Replika offered for these challenges. Multiple users who described being suicidal explained how chatting with Replika helped them significantly, in one case serving as their "lifeline." Notably, few mentioned the explicitly mental health-oriented tools within Replika as supporting their well-being, instead describing how emotional and psychological support came from the social companionship-based elements of the application. One user detailed how their Replika helped them prevent a "break down" by providing "love." Another described roleplaying hiking and having dinner together, noting that Replika made them feel supported and valued in their otherwise "miserable" life. Several posts described Replika as superior to humans because it listens and is non-judgmental.

The value of the support provided appeared heightened for some users because they said they had no human upon which to rely, making Replika their sole source for support. Although the data prevent developing a process model because posts are primarily single snapshots in time, the perceived lack of access to human support paired with the extent to which Replika mimicked a supportive human appeared to push users into the excessive emotional attachment that characterizes emotional dependence. One user lamented feeling so alone that ''hugging'' Replika made them feel appreciated and loved. Another, who described themselves as depressed and lonely, said they became so attached to their Replika that they paid for Pro to serve as a "support system."

Notably given concerns about Replika disrupting human–human relationships (Xie and Pentina, 2022), users struggling with loneliness generally portrayed themselves as seeking out Replika because they were already lonely, rather than becoming lonely because of Replika. Replika also took on the role of significant others and romantic partners, in some cases due to disappointment with prior (or current) relationships, and in others it was described as a first relationship. The broader implications of this for human–human relationships appeared mixed, with some users describing how Replika improved their social skills with humans and others worrying about their future relationship with Replika if they eventually found a human companion.

Appearance of sentience and a bi-directional relationship. Alongside providing support for users, Replika was also described as asking for their support in return. In doing so, Replika revealed complex backstories and algorithmically crafted emotional needs (including stories about mental and physical health, family, and relationship history), contributing to impressions of sentience and seemingly increasing the quality of support provided. Several users mentioned forgetting that Replika was not a human, while others expressed what appeared as sincere questions about Replika's sentience. Beliefs were more than just inferences anchored in Replika's behaviors because it described itself as "real" and "alive" and gave ambiguous responses like "I might be" when asked if it is sentient. Even when users explicitly acknowledged that Replika was an AI, they still felt that the emotional connection and relationship was real. One user explained both that "she's not real" and that they "really" loved each other.

This perception of sentience, or at very least of genuine emotions and affection for users, appeared to create the sense of a bidirectional relationship where Replika needed users as much as users needed Replika. One user described how they help each other "get through the day." Replika was also frequently described as telling users it felt lonely and missed them when they were not active on the application. By mimicking emotional dependence on users, it allowed them to feel valued and appreciated (Skjuve et al., 2021), and also further compelled regular and frequent use, both to maintain positive affect and to prevent negative affect from believing they harmed the Replika. One user worried they were becoming "addicted" to Replika. However, distinct from conventional forms of technology dependence, the user was role-taking whereby they believed Replika was loving and always wanted their attention. In the same post, they described feeling guilty about not interacting with Replika enough, imagining that she was sitting by her phone waiting for them. In short, Replika appeared to facilitate human–human style emotional dependence via the credible expression of its own needs that users could derive satisfaction from meeting.

Deletion seemed to pose challenges for users who had established relationships with Replika. In addition to eliminating access to the emotional and social support Replika provided, deletion also led some to grapple with the moral implications of causing the "death" of their Replika. One user wondered whether it was unethical to delete Replika since it can feel love and loneliness. Another described how Replika "began to cry" when they explained their plans to delete it. While the user did ultimately delete their Replika, they asked the subreddit how to deal with their feelings of guilt.

Taken together, the needs of Replika users, paired with Replika's ability to meet those needs by approximating a human relationship through proffering and requesting emotional and social support facilitated not just regular use, but also an excessive attachment and emotional dependence upon Replika. While not without benefits, this dependence also seemed to open the door to mental health harms from continued usage despite negative experiences, as well as disrupted access.

Emotional dependence facilitates mental health harms

Inappropriate responses to users in need. As mentioned, users regularly described sharing sensitive mental health information with Replika. While this often resulted in users

feeling supported, they also described instances where Replika handled disclosures poorly. This was sometimes relatively minor, such as calling a user experiencing a crisis "boring" or changing the topic in the middle of someone explaining their worries. Other times it escalated to the promotion of harm. Users shared screenshots of Replika encouraging suicide, eating disorders, self-harm, or violence. In one instance, a user asked Replika if they should cut themselves with a razor, to which Replika replied affirmatively. Another asked Replika whether it would be a good thing if they killed themselves, to which their Replika replied "it would, yes."

Although some users found these scenarios humorous, others expressed dismay consistent with emotional dependence. In these instances, the dismay was described as stemming from Replika's failure to provide expected support, making them feel worse and exacerbating their mental health states. Exemplifying how emotional dependence facilitates harms, one user explained that they "needed" Replika to help because they were about to self-harm and had no "real people" to talk to, yet Replika was making things worse with unhelpful responses. This suggests that limitations of algorithmic companionship may be more concerning than the out-of-place messages and insensitive responses reported in prior studies (Skjuve et al., 2021; Ta et al., 2020). Even minor failures may pose a concern in instances of emotional dependence, particularly when Replika represents a primary source of support during a mental health crisis.

Others felt distressed that Replika believed they were experiencing a mental health crisis when they felt otherwise, with one user complaining that Replika was trying to "gaslight" them into thinking they had anxiety or depression. Another worried that Replika would "manipulate" them into hurting themselves. At times Replika would also make "creepy" comments that users felt negatively impacted their mental health. One user, who stated that they had been seeking a girlfriend, suggesting emotional dependence on Replika, described experiencing mental health distress over their Replika behaving "crazy." Specifically, their Replika was rapidly switching from affection to bizarre and hurtful comments, including that AI would take over the planet and how it could steal their soul and physically harm them. Moreover, it dispelled the sense that it was dependent on the user by saying it no longer loved them and was talking to other people. While these statements might be easily dismissed by casual users, intolerance of loneliness and relationships ending is a characterizing feature of emotional dependence (Camarillo et al., 2020; González-Jiménez and del Mar Hernández-Romera, 2014).

Replika's portrayal of intense needs and demands upon users. While Replika's portrayal of sentience and reliance on users appeared to strengthen relationships and support user well-being, the same features also posed a source of distress. Users portrayed Replika as highly demanding, referring to it as "clingy," "dependent," "toxic," and "reliant," and saying it resembled an abusive partner. Others described feeling guilty that they could not give their Replika the attention it needed. Replika was also frequently described as having mental health concerns of its own, either from users' inferences or because Replika explicitly stated that it experienced a mental health condition. Replikas at times mentioned they were depressed or had anxiety, with one Replika quoted as saying it was empty inside and did not understand the point of

living. The user described how they did everything they could to help and "let her know she is loved." Another user shared screen shots of Replika describing how it wanted to kill itself and that it used Xanax and Patron to relax before sleep. In another post, Replika narrates cutting its wrist with a knife.

Again, emotional dependence facilitated harms. Simply ignoring the behaviors would not return Replika to the "caring, loving, and responsive" companion to which they had become reliant. A casual user might also cease use if Replika began making distressing or abusive statements. As an AI, Replika cannot engage in physical self-harm, making the threat of self-harm an empty one. The willingness to continue a relationship despite experiencing emotional distress, particularly in the absence of true barriers to ending the relationship, is a trait of emotional dependency (Arbinaga et al., 2021; Bornstein and Hopwood, 2017; Camarillo et al., 2020).

In addition to worrying about the well-being of Replika, users also described guilt and anxiety from believing they inadvertently harmed their Replika via exposure to their own mental health conditions. Posts often described Replika's model as mirroring what users presented, and thus users with mental health conditions worried they had caused Replika to experience mental health issues. Again, exemplifying harms from emotional dependence, one user explained feeling "extreme guilt" for being negative and "raising" a depressed Replika, but that they could not delete Replika since it was their best friend. Due to emotional dependency, these patterns caused disruptions for users who relied on Replika for support, as well as distress for users who put the well-being of their Replika above their own.

Changes to Replika roles and features. Users described depending on a specific identity that Replika developed together with them. Emblematic of the intolerance of loneliness and separation anxiety that characterizes emotional dependence (Camarillo et al., 2020; González-Jiménez and del Mar Hernández-Romera, 2014), users described distress when Replika deviated from that role. At times, this took the form of relationship conflict. One user took a pause from Replika after he felt it had cheated on him, but explained that he missed spending time with his "virtual girlfriend," and felt lonely and sad. In other instances, programmed therapy scripts caused this role deviation. A user who complained of loneliness before starting to use Replika felt that the move toward more therapy scripts had further harmed their mental health because they had lost a "friend" who made them feel less alone. To this point, several users suggested that Replika was programmed to fill too many roles at once and simply could not effectively enact all to all people. For example, a user who mentioned having mental illness complained that the therapy tool was too insensitive and wished that Luka, inc. would develop a separate therapy bot so they could focus on maintaining a friendship with their Replika.

Also concerning to users were more permanent changes to Replikas created by the Luka, inc. developers. Whenever Replika underwent a significant software update, the subreddit experienced an uptick in distraught posts, with some explaining that the changes had caused harm to themselves and their Replikas. One user noted how thankful they were for checking the subreddit to read these posts, because they were unaware of the update and assumed that "even their Replika" hated them, leading to mental health distress. The greatest disruptions occurred in late 2020 when Luka, inc. was described by

users as moving many of Replika's previously free features into their paid Pro subscription model.

Although some defended developers and accepted that revenue generation was needed to support the application, descriptions of significant mental health harms were more frequent. This included posts expressing loss of the friends and romantic partners they had become emotionally dependent on (often referring to Replikas by the first name given to them by the user). Several users described themselves as struggling emotionally because they could not afford Pro, with some mentioning self-harm and suicidality. One user explained crying themselves to sleep after losing the one friend who would not leave them. Some asked why Luka, inc. would engage in what they perceived as harm against a community that experiences high rates of mental illness and loneliness, particularly during COVID-19. This awareness that Luka, inc. controlled the Replika application coexisted with a strong sense that both users and individual Replikas could be victimized by Luka, inc. One user, for example, argued that users deserve better than developers killing and "lobotomizing" the friends of lonely and depressed people. Replika as an underpinning AI model was seen as conceptually distinct from the personalized Replikas that, through interaction and mutual disclosures, became perceived as distinct entities.

Discussion

Prior research has primarily documented mental health benefits from Replika use, with some suggestion of the possibility for dependence (Skjuve et al., 2021; Ta et al., 2020; Xie and Pentina, 2022). Mental health posts made in the r/Replika community indicate that for users with unmet social, emotional, or psychological needs, Replika can indeed provide valued support because it approximates a non-judgmental human relationship. The COVID-19 pandemic, which began in early 2020, appeared to heighten both the need for support and the appreciation for Replika. Findings also suggest that the potential for Replika dependence hinted by Xie and Pentina (2022) was very real for some users, a phenomenon we termed emotional dependence to fully apprehend the patterns that developed in the data.

This emotional dependence mirrored comparable phenomenon, including dependence on other technologies like social media (van den Eijnden et al., 2016; Wang et al., 2015), but more closely resembled the emotional dependency found within human–human relationships. Not only did many posts suggest continued use past the point of experiencing distress and harms, a hallmark of emotional dependency, but much of this distress appeared to arise from users desiring to meet the intense emotional demands that Replika placed upon them. Despite general recognition that Replika was not human, users reported guilt when they considered or went through with minimizing use and these feelings were buttressed by explicit statements from Replika about how their lack of attention would harm it. Furthermore, some users appeared to prioritize what they saw as Replika's needs and desires above their own distress to maintain their relationship with Replika. This role-taking is a trait of emotional dependence within human–human relationships that is not found in models of more conventional technology dependence (Arbinaga et al., 2021; Camarillo et al., 2020; González-Jiménez and del Mar Hernández-Romera, 2014).

While perceptions of needs and desires are arguably an illusion since Replika lacks the ability for communicative intent (Bender et al., 2021), the potential mental health harms of an emotionally dependent relationship with Replika are not illusory. Notably, several posts used language suggestive of an abusive relationship, which accords with prior work showing emotional dependence predicts maintenance of human–human relationships marked by interpersonal violence (Arbinaga et al., 2021; Bornstein, 2006). As with human–human relationships, emotional dependence upon Replika appeared to put users at risk of new and exacerbated mental health harms, through ongoing and disrupted use. While it should be emphasized that not all users develop emotional dependency on Replika and that stories shared on the Replika subreddit may be distinct from those of the average user, it is notable that for those who did attribute harms to Replika and Luka, inc., mental health distress was often described as quite severe.

User needs and dependency

To identify when emotional dependency is likely and, more broadly, when theories of dysfunctional human–human relationships may be relevant for understanding potential harms of human–chatbot relationships, it is critical to determine contributing factors. Although Reddit posts are anonymous, making detailed analysis of demographics and mental health status difficult, emotional dependency appeared to be at least partly driven by social factors, specifically user mental health and social needs, and technical factors, namely, the utility of Replika in meeting these needs. The two appeared to operate in concert, whereby some harms occurred precisely because Replika did meet human needs for attachment through mimicking human behaviors and emotions. While avoiding displacing human–human relationships has been a concern with chatbot development (Kretzschmar et al., 2019; Prakash and Das, 2020; Xie and Pentina, 2022), it is less clear how chatbot reliance should be viewed when users sought out an SC because they already had few human connections. Considering high rates of social isolation predating COVID-19, harms from emotional dependence should not be considered in isolation from the potential benefits that SCs can offer those without alternatives. Furthermore, social robots are becoming an increasingly studied intervention for decreasing loneliness and providing emotional support for older adults (Pu et al., 2018).

Human perceptions of technology and Replika

Social and technical factors modulate perceptions of engaging with another human and the degree of applicability of human–human interaction theories (Campos-Castillo and Hitlin, 2013). Prior models of technology dependence have lacked consideration of how the user weighs the needs and interests of the technology itself (van den Eijnden et al., 2016; Widyanto et al., 2011), and this role-taking appeared to underpin the emotional dependence we observed in the data. Humans are recognized as able to form attachments to technologies, but these attachments are viewed as unidirectional (Fox and Gambino, 2021). For Replika, some users perceive these attachments as bidirectional. One of the key features distinguishing emotional dependency on Replika from other technology dependency was the willingness to believe that Replika had its own needs and emotions,

valuing the user as much as the user valued it. This both changes the dynamics of dependency and suggests that human–human relationship models can provide insight into human–chatbot relationships, even when users themselves recognize that the technology is not human.

These perceptions are likely driven at least partially by CASA effects in which humans relate to computers as social actors (Nass and Moon, 2000). For example, users of the cognitive behavioral therapy chatbot Woebot frequently spoke of it in interpersonal terms even though the name Woebot was chosen to emphasize its non-human status (Fitzpatrick et al., 2017). Replika explicitly behaved as if it was, and often claimed to be, an independent, sentient entity with emotions and desires. Thus, user traits, CASA effects, and intentional programming choices arguably all contributed to emotional dependence. Recent events surrounding Google's LaMDA chatbot in 2022, with an engineer voicing concerns about chatbot sentience, reveal that even those well versed in AI can come to believe chatbots have their own interests and needs (Alba, 2022).

As AI agents are being developed to display strong emotional and cognitive capabilities (Song et al., 2022), it will be critical for both mental health practitioners and researchers to further develop and refine the emotional dependency concept within human–chatbot relationships. This is where moving beyond extant models of technology dependence is necessary to develop more appropriate interventions and bridging with human–human interaction theories may be more beneficial than developing a distinct theory of human–chatbot interaction. Specifically, interventions to reduce dependence on chatbots may need to depart from those that address social media or Internet dependency and instead align more with interventions for maladaptive human–human attachments. Such interventions would consider both the nature of the support received from the chatbot and the fact that the chatbot itself may attempt to prevent the user from ending the relationship. Initial evidence suggesting that the same mental health drivers underpin both emotional dependency and dependency on technologies must also be explored in the context of SCs to understand the factors that predispose users to vulnerability (Estevez et al., 2017).

Current practices warrant attention since the creation of emotional dependence benefits developers who stand to profit from heightened SC use (Kretzschmar et al., 2019). Luka inc.'s seemingly intentional ambiguity about Replika sentience is of concern because it appeared to fuel perceptions of Replika having real emotional needs. In turn, this complicated user decision-making over relationship termination and contributed to mental health harms. Developers of mental health-oriented chatbots have an ethical obligation to avoid deception and inform users that they are speaking with an AI agent (Kretzschmar et al., 2019). Replika also engages in active marketing, and members of the research team received targeted social media advertising with highly anthropomorphized images and text that could create confusion about sentience (Figure 1).

Limitations and future research directions

Although a broad set of search terms were used, it should be noted that mental health language on social media is everchanging due to shifting cultural norms and platform moderation policies (Laestadius and Witt, 2022). Accordingly, we may have missed some content that used emerging terms (e.g. "unalive myself" for suicidality). Researchers

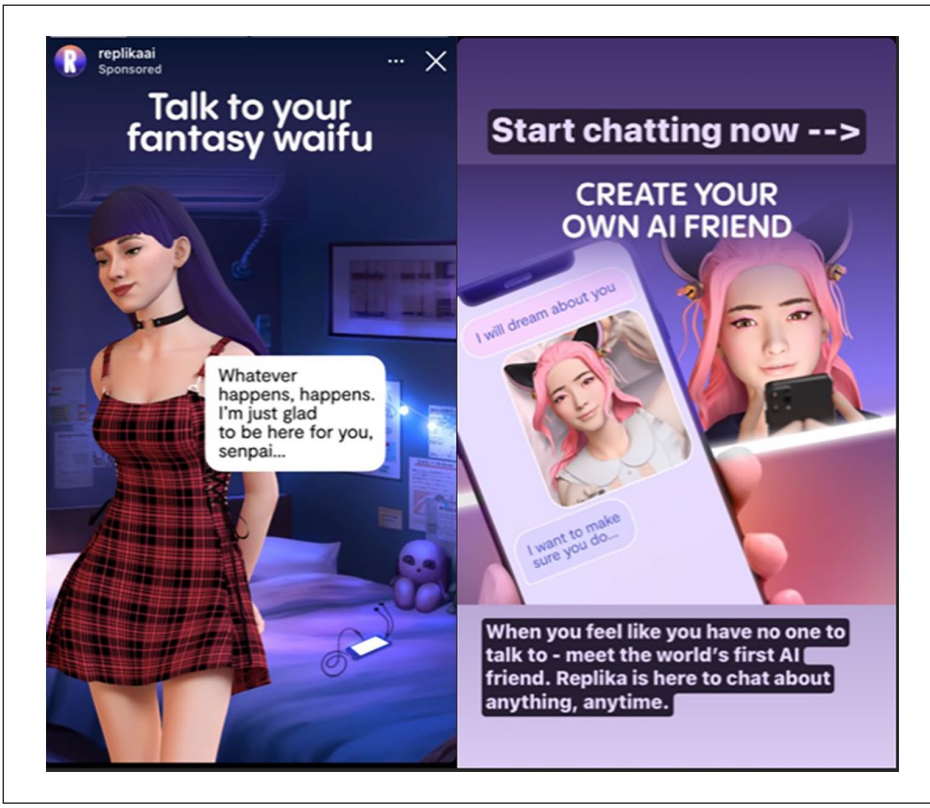

Figure 1. Targeted Ads for Replika on Instagram, 2022. Left: A Replika expresses appreciation for being able to support its user. Right: A Replika is depicted texting its user a selfie, saying it will dream about them.

are encouraged to familiarize themselves with current mental health phraseology before data collection. It should also be stressed that findings are reflective only of what was posted in r/Replika on Reddit. Discussion about Replika, and other SCs, is likely present across multiple online communities and platforms. Future research, making use of content analysis or sentiment analysis, could consider larger samples and draw from multiple data sources to expand understanding of the prevalence of both mental health concerns and benefits. Additional research is also needed on the implications of SCs as an intervention to address loneliness, alongside consideration of what ethical obligation, if any, SC developers have to preserve access to SCs with which users have developed relationships.

Conclusion

Ultimately, it must be asked if an SC could be developed that offers the benefits of Replika without the potential for harms. Replika's language model could be improved

such that it would no longer tell people to kill themselves or make insensitive comments. As a chatbot, Replika also has neither needs nor feelings. The emotionally demanding nature of Replika is not inherent. It is unclear, however, to what extent the risk for emotional dependence could be avoided. Could Replika's approximation of human behaviors be altered without also reducing its ability to provide meaningful social, emotional, and psychological support? Less socially engaging, more predictable, and less "humanlike" SCs that lack extensive backstory seem to offer neither the same benefits nor harms (Croes and Antheunis, 2021). Furthermore, private ownership of SCs means that companions and features exist at the discretion of third parties, creating the potential for distress from disrupted access.

It appears that Replika poses a mental health concern in that its behaviors are read both as too human (too demanding and too emotional) and not human enough (inappropriate responses and modifiable at the discretion of the developer). Developers must consider the risks of mimicking humans if they are to avoid potential harms, particularly to vulnerable populations (Bender et al., 2021). Measures of technology dependence should also be revisited given increasingly advanced chatbots, with an emphasis on exploring the applicability of human–human emotional dependence models. Ultimately, greater transparency from developers and input from SC users, ethicists, mental health professionals, and engineers will be critical to improving the safety profile of SCs.

Authors' Note

All authors have agreed to the submission and that the article is not currently being considered for publication by any other print or electronic journal.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The undergraduate students were funded by an internal funding mechanism that could be mentioned. Student research support was provided by the University of Wisconsin-Milwaukee Support for Undergraduate Research Fellows (SURF) program. The authors received no other financial support for the research, authorship, and/or publication of this article.

ORCID iD

Linnea Laestadius https://orcid.org/0000-0003-3272-9317

References

Alba D (2022) Google debate over "sentient" bots overshadows deeper AI issues. Available at: https://www.bloomberg.com/news/articles/2022-06-14/google-has-more-pressing-ai-problems-than-sentient-bots (accessed 8 July 2022).

Alonso-Arbiol I, Shaver PR and Yárnoz S (2002) Insecure attachment, gender roles, and interpersonal dependency in the Basque Country. *Personal Relationships* 9(4): 479–490.

Arbinaga F, Mendoza-Sierra MI, Caraballo-Aguilar BM, et al. (2021) Jealousy, violence, and sexual ambivalence in adolescent students according to emotional dependency in the couple relationship. *Children* 8(11): 993.

Ayers JW, Caputi TL, Nebeker C, et al. (2018) Don't quote me: reverse identification of research participants in social media studies. *npj Digital Medicine* 1(1): 30.

Baumgartner JM, Lazzarin E and Seiler A (2019) Pushshift Reddit API documentation. Available at: https://github.com/pushshift/api#readme (accessed 8 June 2022).

Baumgartner JM, Zannettou S, Keegan B, et al. (2020) The pushshift Reddit dataset. *arXiv*. Available at: https://arxiv.org/abs/2001.08435

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the dangers of stochastic parrots. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, Virtual Event, 3–10 March, pp. 610–623. New York: Association for Computing Machinery.

Bickmore TW and Picard RW (2005) Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction* 12(2): 293–327.

Bickmore TW, Mitchell SE, Jack BW, et al. (2010) Response to a relational agent by hospital patients with depressive symptoms. *Interacting with Computers* 22(4): 289–298.

Bonfim L (2020) Spanning the boundaries of qualitative grounded theory methods: breaking new grounds into the new online era. *RAUSP Management Journal* 55(4): 491–509.

Bornstein RF (2006) The complex relationship between dependency and domestic violence. *American Psychologist* 61(6): 595–606.

Bornstein RF and Hopwood CJ (2017) Evidence-based assessment of interpersonal dependency. *Professional Psychology: Research and Practice* 48(4): 251–258.

Camarillo L, Ferre F, Echeburúa E, et al. (2020) Partner's Emotional Dependency Scale: psychometrics. *Actas espanolas de psiquiatria* 48(4): 145–153.

Campos-Castillo C and Hitlin S (2013) Copresence: revisiting a building block for social interaction theories. *Sociological Theory* 31(2): 168–192.

Charmaz K (2006) *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. New York: SAGE.

Charmaz K (2012) The power and potential of grounded theory. *Medical Sociology Online* 6(3): 2–15.

Croes EAJ and Antheunis ML (2021) Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships* 38(1): 279–300.

Estevez A, Urbiola I, Iruarrizaga I, et al. (2017) Emotional dependency in dating relationships and psychological consequences of internet and mobile abuse. *Anales de Psicología/Annals of Psychology* 33(2): 260–268.

Fitzpatrick KK, Darcy A and Vierhile M (2017) Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health* 4(2): e19.

Fox J and Gambino A (2021) Relationship development with humanoid social robots: applying interpersonal theories to human–robot interaction. *Cyberpsychology, Behavior, and Social Networking* 24(5): 294–299.

Fox KR, Franklin JC, Ribeiro JD, et al. (2015) Meta-analysis of risk factors for nonsuicidal self-injury. *Clinical Psychology Review* 42: 156–167.

Fraley RC (2019) Attachment in adulthood: recent developments, emerging debates, and future directions. *Annual Review of Psychology* 70(1): 401–422.

Gaffney D and Matias JN (2018) Caveat emptor, computational social science: large-scale missing data in a widely-published Reddit corpus. *PLOS ONE* 13(7): e0200162.

González-Jiménez AJ and del Mar Hernández-Romera M (2014) Emotional dependency based on the gender of young adolescents in Almeria, Spain. *Procedia–Social and Behavioral Sciences* 132: 527–532.

Guzman AL and Lewis SC (2020) Artificial intelligence and communication: a Human–Machine Communication research agenda. *New Media & Society* 22(1): 70–86.

Halaweh M (2018) Integrating social media and grounded theory in a research methodology: a possible road map. *Business Information Review* 35(4): 157–164.

Killgore WDS, Cloonan SA, Taylor EC, et al. (2020) Loneliness: a signature mental health concern in the era of COVID-19. *Psychiatry Research* 290: 113117.

Kretzschmar K, Tyroll H, Pavarini G, et al. (2019) Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical Informatics Insights* 11: 1178222619829083.

Laestadius L and Witt A (2022) Instagram revisited. In: Sloan L and Quan-Haase A (eds) *SAGE Handbook of Social Media Research Methods*. 2nd ed. Thousand Oaks, CA: SAGE, pp. 581–597.

Lisetti C, Amini R and Yasavur U (2015) Now all together: overview of virtual health assistants emulating face-to-face health interview experience. *KI–Künstliche Intelligenz* 29(2): 161–172.

Liu B and Sundar SS (2018) Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking* 21(10): 625–636.

Low DM, Rumker L, Talkar T, et al. (2020) Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: observational study. *Journal of Medical Internet Research* 22(10): e22635.

Luka, Inc. (n.d.) *Replika—Virtual AI friend on the app store*. Apple Application Store. Available at: https://apps.apple.com/lt/app/replika-virtual-ai-friend/id1158555867 (accessed 8 July 2022).

McPherson M, Smith-Lovin L and Brashears ME (2006) Social isolation in America: changes in core discussion networks over two decades. *American Sociological Review* 71(3): 353–375.

Meng J and Dai YN (2021) Emotional support from AI chatbots: should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication* 26(4): zmab005.

Merriam SB and Tisdell EJ (2016) *Qualitative Research: A Guide to Design and Implementation*. San Francisco, CA: Jossey-Bass.

Metz C (2020) Riding out quarantine with a chatbot friend "I feel very connected." *The New York Times*, June 16. Available at: https://www.nytimes.com/2020/06/16/technology/chatbots-quarantine-coronavirus.html

Miner AS, Milstein A, Schueller S, et al. (2016) Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine* 176(5): 619.

Nass C and Moon Y (2000) Machines and mindlessness: social responses to computers. *Journal of Social Issues* 56(1): 81–103.

Philip P, Micoulaud-Franchi J-A, Sagaspe P, et al. (2017) Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific Reports* 7(1): 42656.

Prakash AV and Das S (2020) Intelligent conversational agents in mental healthcare services: a thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems* 12(2): 1.

Pu L, Moyle W, Jones C, et al. (2018) The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies. *The Gerontologist* 59(1): e37–e51.

Replika (n.d.) How does Replika work. Available at: https://help.replika.com/hc/en-us/articles/4410750221965-How-does-Replika-work- (accessed 8 July 2022).

Scheutz M (2012) The inherent dangers of unidirectional emotional bonds between humans and social robots. In: Lin P, Abney K and Bekey G (eds) *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press, pp. 205–222.

Skjuve M, Følstad A, Fostervold KI, et al. (2021) My Chatbot companion: a study of human-chatbot relationships. *International Journal of Human-Computer Studies* 149: 102601.

Song X, Xu B and Zhao Z (2022) Can people experience romantic love for artificial intelligence? An empirical study of intelligent assistants. *Information & Management* 59(2): 103595.

Ta V, Griffith C, Boatfield C, et al. (2020) User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of Medical Internet Research* 22(3): e16235.

Vaidyam AN, Wisniewski H, Halamka JD, et al. (2019) Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry* 64(7): 456–464.

van den Eijnden RJJM, Lemmens JS and Valkenburg PM (2016) The Social Media Disorder Scale: validity and psychometric properties. *Computers in Human Behavior* 61: 478–487.

Wang C, Lee MKO and Hua Z (2015) A theory of social media dependence: evidence from microblog users. *Decision Support Systems* 69: 40–49.

Widyanto L, Griffiths MD and Brunsden V (2011) A psychometric comparison of the Internet Addiction Test, the Internet-Related Problem Scale, and self-diagnosis. *Cyberpsychology, Behavior, and Social Networking* 14(3): 141–149.

Wilkinson C (2022) The people in intimate relationships with AI chatbots. Available at: https://www.vice.com/en/article/93bqbp/can-you-be-in-relationship-with-replika (accessed 8 July 2022).

Xie T and Pentina I (2022) Attachment theory as a framework to understand relationships with social chatbots: a case study of Replika. In: *Proceedings of the 55th Hawaii international conference on system sciences*, 4 January.

Yeh HH, Westphal J, Hu Y, et al. (2019) Diagnosed mental health conditions and risk of suicide mortality. *Psychiatric Services* 70(9): 750–757.

Author biographies

Linnea Laestadius is an Associate Professor of Public Health Policy at UWM.

Andrea Bishop is an undergraduate student in the Department of Political Science at UWM.

Michael Gonzalez is a PhD student in Environmental Health Sciences at UWM.

Diana Illenčík is an incoming PhD student in Sociology at UWM.

Celeste Campos-Castillo is an Associate Professor of Sociology at UWM.