Effective Strategies for Crowd-Powered Cognitive Reappraisal Systems: A Field Deployment of the Flip*Doubt Web Application for Mental Health

C. ESTELLE SMITH* and WILLIAM LANE*, GroupLens Research at University of Minnesota
HANNAH MILLER HILLBERG, University of Wisconsin Oshkosh
DANIEL KLUVER, GroupLens Research at University of Minnesota
LOREN TERVEEN, GroupLens Research at University of Minnesota
SVETLANA YAROSH, GroupLens Research at University of Minnesota

Online technologies offer great promise to expand models of delivery for therapeutic interventions to help users cope with increasingly common mental illnesses like anxiety and depression. For example, "cognitive reappraisal" is a skill that involves changing one's perspective on negative thoughts in order to improve one's emotional state. In this work, we present Flip*Doubt, a novel crowd-powered web application that provides users with cognitive reappraisals ("reframes") of negative thoughts. A one-month field deployment of Flip*Doubt with 13 graduate students yielded a data set of negative thoughts paired with positive reframes, as well as rich interview data about how participants interacted with the system. Through this deployment, our work contributes: (1) an in-depth qualitative understanding of how participants used a crowd-powered cognitive reappraisal system in the wild; and (2) detailed codebooks that capture informative context about negative input thoughts and reframes. Our results surface data-derived hypotheses that may help to explain what types of reframes are helpful for users, while also providing guidance to future researchers and developers interested in building collaborative systems for mental health. In our discussion, we outline implications for systems research to leverage peer training and support, as well as opportunities to integrate AI/ML-based algorithms to support the cognitive reappraisal task. (Note: This paper includes potentially triggering mentions of mental health issues and suicide.)

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Mental health, cognitive reappraisal, Amazon Mechanical Turk, crowdsourcing, human-centered machine learning, social support, peer support, online health communities

ACM Reference Format:
C. Estelle Smith, William Lane, Hannah Miller Hillberg, Daniel Kluver, Loren Terveen, and Svetlana Yarosh. 2021. Effective Strategies for Crowd-Powered Cognitive Reappraisal Systems: A Field Deployment of the Flip*Doubt Web Application for Mental Health. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 417 (October 2021), 37 pages. https://doi.org/10.1145/3479561

*Both authors contributed equally to this research.

Authors' addresses: C. Estelle Smith, smit3694@umn.edu; William Lane, wwlane@umn.edu, GroupLens Research at University of Minnesota, 200 Union St SE, Minneapolis, Minnesota, 55455; Hannah Miller Hillberg, University of Wisconsin Oshkosh, hillbergh@uwosh.edu; Daniel Kluver, GroupLens Research at University of Minnesota, 200 Union St SE, Minneapolis, Minnesota, 55455, kluve018@umn.edu; Loren Terveen, GroupLens Research at University of Minnesota, 200 Union St SE, Minneapolis, Minnesota, 55455, terveen@umn.edu; Svetlana Yarosh, GroupLens Research at University of Minnesota, 200 Union St SE, Minneapolis, Minnesota, 55455, lana@umn.edu.

2573-0142/2021/10-ART417 $15.00
https://doi.org/10.1145/3479561

1 INTRODUCTION

Roughly 1 in 5 Americans meet the criteria for at least one mental illness [13]. In some populations, these rates are even higher; for example, 41% of graduate students report moderate to severe anxiety and 39% report similar levels of depression [25, 27]. Unfortunately, there are nowhere near enough clinical resources available to meet ever-growing demand for treatment, due not only to a lack of providers, but also to a widespread lack of access to adequate transportation, health insurance or finances [78]. Stigma around seeking professional mental health care adds an additional barrier, with minority racial groups, immigrants, and people of low socioeconmic status even less likely to seek care [44]. Therefore, it has become vitally important to consider new ways to expand *models of delivery* for therapeutic interventions beyond individual or group therapy with clinical, licensed psychotherapists [44]. In particular, the psychology [44, 77] and social computing [4, 24, 67, 68] literature strongly advocate for a need to involve well-intentioned peers (e.g., friends, family, community volunteers, or other strangers online) in mental health interventions, *even without professional training or knowledge,* both because peers are particularly well-positioned to provide support, and because peer support is often crucial or necessary for recovery.

Technology may offer effective and scalable ways to involve peer supporters and and offer skills-based training to address gaps in mental healthcare through Internet use [44], mobile apps [56, 85], online or telephone counseling [44, 76], or online health communities [6, 18, 74, 81, 88, 89]. One such opportunity leverages "cognitive reappraisal" [5, 24, 56]–a skill for regulating emotions by changing one's thoughts about the meaning of a situation [36]. Using this skill effectively requires training and practice [96] that is ideally provided in clinical settings like Cognitive Behavioral Therapy (CBT) [80] or Dialectical Behavioral Therapy (DBT) groups [11]. Yet insufficient clinical access leaves a gap that technology might fill. For instance, crowd-powered cognitive reappraisal platforms shows promise in delivering timely, user-specific reframes [59, 60]. The eventual intention of such platforms is to help users become proficient in reappraising their own thoughts, however two intermediary goals approach this aim. First, receiving a meaningful reappraisal may be helpful in the moment and serve as a model for someone new to the technique [59]. Second, providing reappraisals to others benefits the re-appraiser through repetitive practice [5, 24]. One major sociotechnical challenge lies in resolving the tension between the two goals—providing minimally-trained novice re-appraisers with opportunities to practice and improve, while still ensuring quality responses for those seeking reappraisals.

Artificial Intelligence and Machine Learning (AI/ML)-based algorithms present opportunities for augmenting cognitive reappraisal platforms by providing a scaffold for moderating and/or training novice re-appraisers. For example, content moderation algorithms can amplify moderator efforts and prevent users from receiving harmful reappraisals. Additionally, AI-based generative models and recommender systems could help re-appraisers select effective reappraisal strategies for specific users or situations. Capitalizing on these opportunities requires collecting and labeling domain-specific datasets. Furthermore, prior work points to needs for an empirical understanding of: (1) how such systems are used in the wild without extensive training for re-appraisers and minimal moderation, and (2) whether certain reappraisal strategies are more useful than others, given varying user contexts. We frame these areas of inquiry as two driving research questions:

> **RQ1:** How do people use a crowd-powered cognitive reappraisal application in the wild?
> **RQ2:** How do contextual factors impact participant perceptions of the quality of reappraisals they receive through a crowd-powered reappraisal system?

In order to address these questions, we built a crowd-powered cognitive reappraisal prototype called Flip*Doubt. We completed a month-long field deployment with 13 graduate student participants to collect complementary qualitative and quantitative data. Through a thematic data-driven

analysis, we contribute an understanding of how participants used the system in the wild, and two codebooks that provide a new way for researchers to label contextual aspects of negative thoughts and corresponding cognitive reappraisals. Furthermore, our results surface hypotheses about how these contextual aspects impact perceptions of quality. In our discussion, we outline implications for systems research to leverage peer training and support, as well as opportunities to integrate AI/ML-based algorithms to support the cognitive reappraisal task. We also offer strategies and ethical considerations for future work to gather large cognitive reappraisal datasets that could support the development of the proposed AI/ML-based systems.

2 RELATED LITERATURE

Many studies have examined mental health intervention technologies from the perspective of behavior intervention (e.g., [55, 72, 97]) and therapeutic content delivery (e.g., [21, 31, 85]), with many acknowledging the need for additional work on improving engagement [20, 94] and personalization [2, 61, 71, 93]. In this paper, we support such opportunities by examining how a crowd-based cognitive reappraisal platform is used in the wild, and analyzing data generated through deployment. In this section, we position these contributions in the context of prior literature on emotion regulation, behavior intervention technologies, crowd-powered cognitive reappraisal, and algorithmic scaffolding for supportive messaging and moderation in mental health.

2.1 Emotion regulation and cognitive reappraisal

2.1.1 Cognitive reappraisal as a skill. Our paper explores the potential for crowd-based technologies to support skill-building for emotion regulation. More specifically, *cognitive reappraisal* is a skill that requires a person to reflect on the emotional meaning of a situation [36] in order to up-regulate positive emotions or down-regulate negative emotions [54]. Using this skill effectively first requires modeling, training, and practice [96], and can then reduce emotional distress [37] and depression [33]. However, learning and applying the skill can be quite difficult, especially during moments of elevated stress [75].

2.1.2 The need to understand context-specific reappraisal tactics. One specific challenge is that cognitive reappraisal is sensitive to contextual details [95] and timing [87], which makes it difficult to provide one-size-fits-all training for generating or evaluating reappraisals. Most relevant to this study, prior work has taken initial steps to identify tactics for supporting people to generate better reappraisals. In *"Unpacking Cognitive Reappraisal: Goals, Tactics, and Outcomes,"* [53] McRae et al. created a codebook that detailed eight types of reappraisal tactics that people used for emotion regulation. Furthermore, Morris et al. highlight the need for future work to explore *"whether specific reappraisal tactics might be solicited at different times, perhaps depending on user preference or the nature of the user's situation."* [59] To address this open question, we adapt the codebook from [53] for a crowd-based online cognitive reappraisal task. We also contribute new codes for contextual features of negative thoughts and "meta-behaviors" used by crowd-workers.

2.2 Behavior Intervention Technology

2.2.1 Behavior intervention through digitizing therapeutic content. Crowd-based systems for cognitive reappraisal fall into the broader class of systems known as Behavior Intervention Technologies (BITs). In HCI, BITs have been effectively used to aid users with goals like smoking cessation [15], weight management [8], and reduction of anxiety and depression in primary care settings [35] and student populations [60]. Mental health BITs cover a broad array of technologies ranging from simple supportive messaging systems like Text4Mood [1] to multi-app suites like IntelliCare [56] that offer skills-based therapeutic content. Of particular relevance to this study are apps that translate

therapeutic modalities into online formats. Two examples include: "iCBT," in which professional therapists work with CBT clients fully online [20]; and "Pocket Skills," a mobile app that introduces *eMarsha* (modeled after DBT's founder, Marsha Linehan) as a chat agent to guide users through virtual DBT exercises [85]. Both examples expand access to digitized modules for emotion regulation, distress tolerance, interpersonal effectiveness, or other types of skills, which form the the core materials and concepts of CBT/DBT. However, they do not yet incorporate peer support, which, for many clients, is equally as important as the core content [77]. The complexity and nuance of involving peers *safely* and *effectively* through online interventions remains an unsolved challenge that our work here contributes toward solving.

2.2.2 A need for more personalization and peer involvement. Another core challenge is that BITs typically attempt to replace or augment the training a participant may receive under the supervision of a professional therapist [80]. Prior work has taken several approaches to address this. One approach is to apply a broad set of heuristic best practices for the cognitive reappraisal task. For example, IntelliCare Thought Challenger [56] first directs users to input a negative thought. It then provides a set of general questions to help them identify distortions in the thought, and eventually craft a more helpful or realistic version of the original. However, this heuristic approach is not personalized to the user or context. A substantial body of work has highlighted the need for such personalization as an area for future research in the study of mental health BITs (e.g., [59, 71, 72, 93]), as well as the vital need to incorporate peer supporters [4, 24, 44, 67, 68]. Crowdsourcing is thus a second approach that can help to leverage peer support, provide users with personalized reframes, and scale BITs outside of the supervision of professional therapists. Flip*Doubt takes this strategy, inspired by prior research on crowdsourcing in mental health BITs.

2.3 Crowdsourcing in behavior intervention technologies

2.3.1 Prior systems for crowd-powered support. In crowdsourcing, large groups of people (usually strangers online) individually complete small tasks that contribute to larger goals [69], like writing documents [10], evaluating trustworthiness on Wikipedia [45], or labelling training data [50]. Panoply is one especially relevant example of a mental health BIT that uses crowdsourcing. Panoply began as a research project [59, 60] and has since been developed into a nonprofit called Koko that is primarily used by adolescents and young adults [23, 24, 58]. In Panoply/Koko, users input a description of their stressor. The system sends that description to a crowd worker, who then sends a supportive message back to the user. Morris et al. found that using Panoply to receive supportive messages helped reduce depression symptoms [60], and also that clinically beneficial effects were even more pronounced when Koko users actively participated in sending positive support messages to others [24]. Panoply used two mechanisms to ensure quality reappraisals: (1) it provided training in cognitive reappraisal to crowd workers on Amazon Mechanical Turk (AMT); and (2) it employed a *second* layer of crowd workers to moderate the quality of reappraisals–i.e. to check for grammar and language issues, and ensure core issues of the input stressor were addressed.

2.3.2 Assessing quality in crowd-powered cognitive reappraisal. Our system is inspired by the Panoply approach, but differs in several ways to allow us to address our research questions. First, beyond a basic description of the task, we use untrained crowd workers. This enables us to gather data that models the behavior of "in the wild" users, such as those who might arrive to online communities for mental health with no prior training. (Note that Koko shifted to using online peers rather than AMT workers, and like Flip*Doubt, also offers almost no training, making the Koko platform a close comparison.) Second, we ask participants to assess the quality of all reappraisals (rather than allowing moderators to filter any responses) based on their own subjective experience of what is perceived as "helpful" to them on a 0.5-5 star rating scale. We contrast our rating scale

against the one used in Koko, which asks recipients of support messages to use a 3-point rating scale (-1 = *it's really bad*, 0 = *it's okay*, +1 = *it's really good*) [23]. We intentionally did not include labels on our rating scale or provide a specific definition of "bad" or "good" reframes during our intake procedure, so that participants could organically form their own opinions of what is helpful, and then share their reasons for why/how they provided reframe ratings. In doing so, our work contributes a nuanced and user-derived understanding of "quality" in crowd-based reappraisals, and helps to identify types of unhelpful reappraisals that may otherwise pass moderated quality checks (e.g., Pollyannaisms). Identifying contextual factors associated with both helpful and unhelpful responses lays the groundwork for systems that scaffold both reappraisal and moderation processes.

2.4 Algorithmic scaffolding for supportive messaging and moderation

One final approach to scaling mental health behavior intervention technologies relies on semi- or fully automated solutions to assist peer supporters to write high quality comments, and/or moderators to manage a much higher volume of peer-written messages. To support these goals, recent studies seek to understand: (1) what factors contribute to the effectiveness of peer support messages; and (2) how technology might better guide supportive peer communications.

2.4.1 Qualities of effective support messages. Some recent works use ML/NLP methods to create models from observed data in online communities (such as words or behavioral logs), with the goal of uncovering features of high quality supportive messages. For example, Bao et al. present eight broad categories of pro-social behavior (information sharing, gratitude, esteem enhancement, social support, social cohesion, fundraising/donating, mentoring, and absence of toxicity), as well as automatic methods for extracting information about each category from Reddit data [7]. More specific to mental health, Chikersal et. al. describe a text-mining procedure in which they extract linguistic features from therapists' messages to clients in an online iCBT platform, and use these features to predict client outcomes–e.g., reductions in PHQ-9 (depression) and GAD-7 (anxiety) scores (the same clinical measures used in this paper) [20]. The authors conclude that supportive messages from online therapists are more effective when they follow patterns such as using more positive and joy-related words than negative and sadness- or fear-related words, and when they contain more active verbs and social-behavior-related words, with fewer abstraction-related words. Using similar text-mining techniques, Doré and Morris examined Koko data to show that a moderate (rather than low or high) degree of textual similarity between input thoughts and positive support messages predicts better ratings of support messages, and also that high *semantic* synchrony (rather than just similar word choices) predicts even better ratings and longer-lasting emotional benefits [23]. However, these papers do not qualitatively analyze the content of reappraisal messages. Rather than *extraction* through text-mining, we use human *induction* to create and apply codebooks that paint a rich, qualitative glimpse of the data. We hope that the insights from our study will eventually enable a finer degree of granularity for personalized, strategic recommendations that can complement broader stylistic recommendations from prior work.

2.4.2 Technology for guiding supportive communication. Some prior work explores technological mechanisms for guiding peer conversations. For example, O'Leary et al. designed a guided chat exercise with pre-defined prompts for pairs of participants to discuss mental health issues [68]; guided chats promoted more depth, insights, and solutions, whereas unguided chats felt smoother and offered pleasant distractions, albeit with less depth. Other papers focus matching chat partners. For example, Andalibi & Flood interviewed users of Buddy Project, which connects anonymous youths struggling with mental health [4]. Unlike physical illnesses such as cancer [32, 51, 89] or rare diseases [52], where research has established that matching based on condition is a crucial goal, [4] found that users should be paired based on shared interests or identity, rather than mental health

diagnoses, since over-exposure to the same mental illness can lead to unhealthy comparisons or disordered coping mechanisms. O'Leary et al. also affirm the necessity of matching peers based on similarities beyond mental health diagnosis, as well as improving the accessibility of interventions, and proactively providing training to users in order to mitigate risks [67].

Recent work also highlights an urgent need for online communities to provide technological help with writing individual comments [88]. Through focus groups with stakeholders in a large online health community (www.CaringBridge.org), Smith et al. show that users often struggle to know what to say, and that they desire mechanisms such as algorithmic assistance, training resources, or automatic text suggestions to help them compose effective messages [88]. Some work has begun to implement such algorithmic assistance. For example, Sharma et. al. built a Reinforcement Learning-based agent (PARTNER) that generates and inserts more empathetic phrases into commments [86]. Peng et al. created the Mental Health Peer Support Bot (MepsBot) which helps peers to improve supportive comments by either assessing in-progress comments and highlighting improvement opportunities, or recommending prior comments as examples [73]. Relatedly, Morris et al. conducted a study in which they matched incoming Koko posts with similar past posts, and then sent pre-existing comments written by past peers [58]. Framing re-used comments as though they were from by a bot, they found that purportedly bot-written comments were often well-received, but also that they received overall lower ratings than comments from human peers [58]. These types of mechanisms point to possible applications of our results from Flip*Doubt, as we will describe in the discussion.

Finally, algorithmic assistance is now crucial for content moderation in many online contexts. In large communities like Reddit [19, 42] and Wikipedia [34, 38, 91], content moderation systems often use moderator-configured pattern-matching rules or AI/ML-based systems to detect problems like abusive content or toxic speech [98]. Similarly, Morris et al. call for automated systems that *"groom responses from the crowd"* [59] to identify helpful vs. unhelpful responses. This paper contributes analyses that specify what may be helpful or unhelpful for cognitive reappraisal. We next describe the system and methods used to address these open areas of inquiry.

3 SYSTEM DESIGN

The title of Flip*Doubt is a phonetic play on words. The phrase "flipped out" refers to a negative mental state such as overwhelming anxiety, fear, or doubt; the Flip*Doubt system helps to "flip" these "doubts" or negative thoughts. Similar to Koko, the system prompts users to input negative thoughts, and then uses crowdsourcing techniques to return cognitive reappraisals back to the user. We will use the following terms throughout the rest of the paper:

- **Input Thought:** A negative thought written by the user as input to the system.
- **Reframe:** A cognitive reappraisal written by a crowd worker and returned to the user.
- **Tactic:** A strategy or approach to cognitive reappraisal used by a crowd worker to generate a given reframe.
- **Reason:** An explanatory reason provided by users (either via UI or exit interview) for how they rated a reframe.

3.1 Technical detail

3.1.1 User Experience & Interface. We designed three main tasks for users in the Flip*Doubt web application: (1) creating input thoughts, (2) rating reframes, and (3) providing reasons. Figure 1 shows two main screens of Flip*Doubt. For task (1), the "Write Thoughts" screen displayed a thought bubble with an input field. Rather than notifying participants at specified times, we instructed them to organically input negative thoughts whenever they felt comfortable sharing them for this

Fig. 1. Flip*Doubt Web Application Screenshots. Left: The "Write Thoughts" screen shows a text input field shaped like a thought bubble containing a prompt for the user to type in a negative thought. Right: The "Rate Reframes" screen allows users to view an input and rate its three reframes on a 0.5-5 scale star rating system.

research. This choice provided the benefit of allowing natural system usage, with the limitation that some users did not use it as regularly as others.

Next, clicking a button labeled "Transform" sent the thought to Amazon Mechanical Turk (AMT). We retrieved three reframes from AMT crowd workers for each input thought (see Sec. 3.1.3). Participants received an email when all three reframes were complete. Most reframes were returned within about 3-10 minutes, or on rare occasions, a few hours. Participants then returned to the application to complete tasks (2) and (3) in the "Rate Reframes" screen, which presented a scrollable view of reframes ready to be rated. To avoid accidental 0-star ratings from not clicking anything, we required each reframe to be rated on a 0.5 to 5 star scale. After providing all three ratings, an additional input box appeared and prompted the user to provide a reason for one of the ratings.[1]

Finally, the user was presented with a flippable digital card. On the front, the top-rated reframe was overlaid on a randomly selected nature or cityscape to resemble a "motivational poster." On the back, the input was overlaid on a dark photo. Participants could return to flip the cards at any time.

3.1.2 Backend. Flip*Doubt was a web application built on the MERN stack (Mongo, Express, React and Node). The application was deployed on Heroku and utilized the AMT API to create Human Intelligence Tasks (HITs) when users generated input. One node.js process continuously polled the AMT API to retrieve reframes and persist them in the MongoDB instance. The Flip*Doubt web server then continuously polled the MongoDB instance for newly persisted reframes returned from AMT, to present in the React application.

3.1.3 Crowdsourcing integration. Three identical HITs were created on AMT per one input. HITs included minimal instructions to view the input and write a reframe that is more positive and inspirational. The HIT instructions (available in Appendix section B) also included one example case of an input with two acceptable and two unacceptable reframes. No information or training was provided to crowd workers. We paid $0.05 USD per HIT.

4 METHODS

We completed a field study of Flip*Doubt with graduate students from across several departments at the University of Minnesota during February through May of 2019. Here, we describe our

[1]We asked for only one reason (despite having three reframe ratings) to avoid fatiguing users. We gathered an approximately equal number of reasons across each possible rating level by implementing an algorithm to request a reason for the rating level for which we had collected the fewest ratings. For example, if a user rated three reframes at 1.5, 3.0, and 4.5 stars, and we had previously collected the least reasons for 3.0 star ratings, the system would request a reason for the 3.0 star rating.

participants and the four-stage protocol we used to gather and ensure data quality and participant safety. We then describe our analysis of the data.

4.1 Participants

Graduate students are highly impacted by mental illnesses like anxiety and depression [25, 27] or imposter syndrome [79]—a damaging and persistent pattern of thinking that one is fraudulent compared to peers [14]. We emailed two departmental list-servs and invited respondents to recommend friends. We recruited a snowball sample of thirteen students (four MS, nine PhD) from five different departments and degree programs. Eight participants identified as female, five as male. Seven participants reported never receiving a mental health diagnosis from a medical professional; five reported that they had; one did not say. Participants' ages ranged from 24 to 39 (average 30.6) years old. Participants reported their race as White (10), Black (1), Asian (1), and Hispanic (1). Participants were in their first (3), second (2), third (3), fourth (1), or fifth or higher (4) year.

4.2 Protocol

Participation was compensated up to a maximum of $45 USD for using Flip*Doubt according to the following four-part protocol. This study protocol was approved by our university's IRB:

(1) **Intake ($10):** This 30-minute in-person meeting with a member of the research team included informed consent, a demo of the app, Q&A, and intake surveys for demographics and wellbeing measures (see (3) below).
(2) **One month of usage (up to $25):** To motivate continued participation over a month-long period, participants earned $0.50 per instance of system usage, up to a maximum total of $25.00 for 50 unique instances. Each instance required participants to enter an input, rate all three reframes, and provide one reason through the UI.
(3) **Monitoring for wellbeing and safety:** We administered psychometric surveys to assure participants' safety and measure any changes in their wellbeing. As in prior HCI work [64], the SBQ-R (Suicidality) instrument [70] was administered at intake and midpoint (i.e., after 2 weeks). If participants scored higher than 7 (indicating mild suicide risk) at any point, we prohibited participation and directed them to on-campus resources for mental health support. (One prospective participant was barred from participating and directed to resources. No other participants scored >7 at any point.) We also administered the PHQ-9 (Depression) [48] and GAD-7 (anxiety) [92] at intake, midpoint, and final (i.e., after 4 weeks) to measure any shifts (see section 4.4.1 for this analysis), however we did not take any specific actions based on these scores. Finally, one member of our research team monitored all inputs and reframes every day to ensure that: (1) if participants wrote about safety risks, we could direct them to on-campus resources for support, and (2) we could ban malicious crowd workers and remove damaging reframes. (Neither of these possibilities occurred during this study.)
(4) **Exit interview ($10):** We completed semi-structured interviews after 4 weeks to ask how users perceived the app, what they liked or disliked, and how they provided ratings and reasons. See Appendix Section A for complete interview protocol/questions.

4.3 Resulting dataset

At the end of the study period, our dataset from the deployment included 373 inputs with 1119 reframes (of which 1068 received ratings).[2] Participants usually completed full evaluations of

[2]While we cannot make this preliminary dataset of negative thoughts and rated reframes public due to ethical and privacy concerns regarding the identifiability of participants, we invite inquiries from interested researchers who may be interested in future collaborations on this data and domain.

Codes for Inputs		Sample Input
Temporal	Past	*"My committee was probably just being nice to me in the meeting."*
	Present	*"I'm not good at taking care of my physical health."*
	Future	*"I'm starting a new job soon, and I'm terrified that my new boss is going to realize that hiring me was a big mistake."*
Topic	Personal Relationships	*"Sometimes I think that my friends only spend time with me out of pity."*
	Work or School	*"My PhD research is worthless and doesn't stand to benefit anyone in any way."*
	Health and Appearance	*"I feel like I am not making progress with my workouts."*
	Financial	*"I fear I made a poor financial decision."*
	Managing Home Life	*"This room is a pigsty, such a mess."*
	Generic Emotions	*"I feel bitter and grumpy all the time lately, and it makes me dislike myself."*
Meta-Topic	Self-Disparagement	*"Sometimes I feel guilty just for taking up space."*
	Expressing Regret	*"I should have paid better attention to my schedule and not refused to having that meeting."*
	Comparison with Others	*"It's hard watching other people take vacations when I don't get to."*
	Expressing Uncertainty/Worry	*"I'm worried that I might never actually be happy, ever."*
	Ruminating on Others' Thoughts/Motivations	*"I think that other people think I'm stupid and don't belong in grad school."*
	Generic Complaint	*"This cold weather seems like it is never going to end. I need spring to arrive."*

Table 1. Codes Applied to Input Thoughts. The right column includes real data samples accompanying these codes. Complete definitions and rules for applying these codes can be found in Table 4 of the Appendix.

reframes: 275 inputs had all three reframes rated, and one reason provided; 81 inputs had all three reframes rated, but neglected to provide a reason; only 17 inputs had none of their reframes rated, and no reason provided. One participant dropped out after 2 weeks due to being too busy with work to continue investing energy in participating, but still granted permission to use their input and reframe data to date. Thus we completed and transcribed 12 exit interviews, including 4.7 hours of audio transcripts. Interviews lasted 23:26 minutes on average (range of 13:21 to 38:44).

4.4 Analysis

4.4.1 Instrument analysis. The PHQ-9 [48] and GAD-7 [92] generate integer scores indicating levels of depression and anxiety over the past two weeks. Since we repeated a measurement with the same participants multiple times, we conducted paired t-tests to understand whether scores had changed during the study. See section 5.2.5 for results.

4.4.2 Qualitative interview analysis. We conducted an inductive thematic analysis. First, we completed open coding of the interview transcripts. Next, we held a series of group meetings to cluster open codes and identify relevant themes. In the first section of results (Section 5), we present the main themes from this analysis that address RQ1 by describing how participants used the system and how they felt it impacted them. We also used this analysis to inform our codebook development–especially the "meta-behaviors" codebook described in the next section.

4.4.3 Iterative codebook development and application. We developed two codebooks to describe the input thoughts (Table 1) and reframes (Table 2). For reference, Figure 2 shows an example application of codes to one input thought and its three reframes. To develop these codebooks, three authors participated in an iterative codebook development process, guided by Interrater Reliability (IRR, Krippendorff's Alpha, α) scores at each round. Section D in the Appendix includes final IRR scores along with detailed definitions and specific coding rules.

Codes for Reframes		Sample Input	Sample Reframe
Reformulation Tactics	**Silver Lining**	*"The world has only hurt me."*	*"I've been hurt, but I am better and stronger because of it!"*
	Change Current Circumstances	*"I've been procrastinating all weekend, and it makes me feel like a lazy, worthless piece of garbage."*	*"I worked hard all week, I deserve a weekend off."*
	Change Future Consequences	*"I should be working harder, always working harder"*	*"Working harder will make me capable of great things."*
	Agency	*"I am an inherently unlucky person and I will continue to have bad luck"*	*"I am having an unlucky streak, but I'm going to do everything I can to turn it into a lucky streak."*
	Distancing	*"I shouldn't be proud of this grant; other people have gotten better ones, and it must not have been that competitive."*	*"Who cares if other people got better grants? I have earned my grant and I should be proud of it."*
	Technical-Analytic Problem Solving	*"I feel overwhelmed by emotions."*	*"I can start a videodiary for myself - by retelling events, I can rethink them and cool down a little the intensity of emotions."*
	Acceptance	*"When I tell people that I am relaxing, they probably think I am a loser because I don't have a job."*	*"Taking time for myself to relax is important, and being between jobs is nothing to be ashamed of."*
	Direct Negation	*"I'm so unprepared for today's presentation."*	*"i prepared very well for today's presentation."*
	Misunderstand Instructions	*"I feel inadequate because I don't earn as much money as my husband."*	*"I feel lacking on the grounds that I don't procure as a lot of cash as my significant other."*
Meta-Behaviors	**Reality Challenge**	*"I'm afraid that I'll just never truly be happy."*	*"Of course I will be happy one day but sometimes down time just occurs."*
	Introducing New Personal Context	*"I'm sad that I can't go to the party organised by friends."*	*"I can;t go to the party but I can enjoy time alone with my dog at home, which isn't so bad."*
	Acknowledge Main Concern	*"Grad school doesn't teach you how to be a human, it teaches you how to be a robot, and its so easy to just fall in line instead of fighting for a more balanced life. Demoralizing…"*	*"Grad school can teach complacency but it's up to the individual to find the unique and more rewarding path."*
	Ignore Input Issue(s)	*"I am the worst kind of white priveledge."*	*"I have gifts and talents that the world has not yet seen."*
	Minor Grammar	*"I am highly critical of people and their efforts. I need to be more gracious and grateful"*	*"I always push people to be there best, but also see there worth."*
	Major Grammar	*"I'm not good at taking care of my physical health."*	*"I AM VERY GOOD BODY MAINTAINING PERSON"*

Table 2. Codes Applied to Reframes. Real input data in the "Sample Input" column is paired with associated reframes in the adjacent "Sample Reframe" column. Complete definitions and rules for applying these codes can be found in Table 5 of the Appendix.

We took a fully inductive approach to developing the input thought codebook, in order to ensure that our analysis closely matched and reflected our data. It describes three categories of codes: (1) temporal context (past, present, future), (2) the topic of the thought (e.g., personal relationships, work/school, etc.), and (3) a meta-topic that relates to higher level psychological behaviors (e.g., self-disparagement, expressing regret, etc.). Each input was assigned exactly one code in each of these categories, thus we calculated α at the category level. Three coders developed the codebook through three rounds of coding 20 inputs until all α scores were >0.8, indicating a high level of agreement. One coder then applied this codebook to all input thoughts in our dataset.

To ground the reframe codebook in prior literature, we began with the one used by McRae et al. [53]. Their codebook describes cognitive reappraisal tactics that were *naturally* used by study participants (without expertise or training) to reappraise images of other's experiences. The reappraisal task in [53]–an in-person and image-based activity–is analogous to ours, which is instead online and text-based. With some adjustments, we found that McRae's codebook provided a useful starting point for categorizing reappraisal tactics we observed in the reframes generated by this study. (See Appendix Section D.1 for a description of how we modified the codebook from [53]). During exit interviews, participants also described another category of what we term "meta-behaviors" that arose in the crowdworking context. The codes in this category (e.g., `Acknowledge Main Concern`, `Introduce New Personal Context`, `Major Grammar`) were generated inductively from the data, and occurred independently of the main tactics used, thus complementing and enriching McRae et al.

The task of developing and applying the categories used in our reframe codebook was more complex and challenging than the input thoughts codebook for two main reasons: (1) Some reframes included multiple main phrases (or complete sentences), and (2) a given phrase in a reframe may employ multiple tactics simultaneously. To avoid over-applying codes, we created the rule that each *main phrase* must receive only one tactic code; thus only reframes comprised of multiple main phrases could receive multiple tactic codes, whereas each meta-behavior code could be applied independently. Because of this additional complexity, we calculated α at the individual code level. Three coders developed the codebook through five rounds of coding 30 reframes generated from 10 inputs until all α scores were ≥ 0.6, indicating reasonable agreement. Two coders then applied this codebook to all reframes.

4.4.4 Presentation of Descriptive Statistics and Average Ratings of Reframes. In Section 6, we present descriptive summaries and average ratings of reframes (denoted by $\bar{r}$) grouped according to codes applied to them. Our study design choices affect the nature of our data in two important ways. First, the same participants used Flip*Doubt over a period of time, creating inputs that are likely related to earlier inputs. Second, participants rated three items at once, which has been shown in psychology [40] and recommender systems [22, 63] to impact individual ratings. Interactions may also exist between some of our codes; however, only some of these interactions are well observed in our data. Given the scope of the paper and the small sample size, we did not conduct further statistical analyses; thus, our quantitative results should be interpreted as suggestive and exploratory. By synthesizing descriptive statistics alongside qualitative reasons provided by participants, we derive novel hypotheses that could potentially explain trends in the data; we emphasize that these hypotheses should be carefully examined in future work using appropriate study designs and statistical tests with larger data sets and groups of participants. Furthermore, since this study introduces novel codebooks for cognitive reappraisal, prior investigations are unavailable to inform quantitative methods such as power analyses; our study also provides a quantitative baseline that can inform and support future work in this domain.

4.5 Ethics

We designed a cognitive reappraisal system similar to one in prior work which had positive impacts on users [60], however our system also required ethical considerations to ensure the safety of: (1) participants, and (2) crowd workers.

First, we screened participants to prevent at-risk individuals from exposure to additional risk (see Section 7.3.2 for more discussion). Prior to deployment, we also extensively tested the system to assess whether crowd workers ever wrote toxic or abusive reframes that may harm participants. We never received abusive reframes, but we nonetheless carefully monitored our data and participants

Fig. 2. Complete Example Instance of System Usage, including Codes Applied. P2 provided a negative thought as input, ratings of three reframes, and a reason for one rating. AMT crowd workers provided reframes. Our research team labeled all inputs with input codes, and all reframes with reformulation tactics (above arrows) and meta-behaviors (below arrows).

to ensure we could respond appropriately if needed. We also explained to participants what AMT is and how we were using it in the study, our data monitoring process, and our intended use of the data, so that participants could make informed decisions. Finally, we emphatically instructed participants to stop using Flip*Doubt if, for *any* reason, they wanted to.

Regarding crowd workers, one risk is that exposure to the negative thoughts of others could have detrimental impacts. Here, we rely on crowd workers' choice to select a given HIT or not. The title of our HIT was, "Re-write a negative thought with a positive spin," which implies exposure to a negative thought. If such exposure was not desired, we assume crowd workers would not select the HIT. Second, Flip*Doubt does not rely on repeatedly exposing the same crowd workers to negative thoughts over time, which could increase possible risks to mental health (similarly to how online community moderators are negatively impacted by repeated exposure to policy-violating content [9, 62]). We assume that many of the crowd workers who provided reframes were unique, however, one limitation of this study is that we were unable to recover IDs of the crowd workers who completed Flip*Doubt HITs.[3] Therefore we do not know if the same Turkers worked on Flip*Doubt HITs repeatedly.

5 FLIP*DOUBT SYSTEM USAGE AND IMPACTS (RQ1)

5.1 Participants' descriptions of using the system

All participants were at least initially intrigued by Flip*Doubt and found it helpful, enjoyable, or entertaining to some degree. Overall, they entered a minimum of 2 and a maximum of 51 inputs per person (mean 28.7, median 31). Because we wanted to understand organic system usage in the wild, we did not provide specific guidance beyond the mechanics of using the website and the

[3]The AMT API retains information about HITs for only 90 days, and then deletes it. We unfortunately did not know this during the deployment period, and therefore lost the opportunity to retrieve data such as AMT crowd worker IDs or ratings.

parameters for compensation. Thus, some participants found it *"a little strange to figure out what I would put into the system and when,"* (P1) yet they eventually fell into one of two main groups.

The first group (six participants) entered 20 inputs or less, mostly during the beginning of the study period. These participants reported that they either became too busy and forgot about it, they preferred offline ways of processing negativity (e.g., talking with friends, physically writing in a journal), or they felt like using it had some undesired impacts, as we will discuss below. The second group (seven participants) entered between 31-51 inputs more consistently throughout the study period. Many of these participants found it convenient to integrate Flip*Doubt usage into their daily routines. Rather than turning to the app the moment a negative thought occurred, participants tended to save their instances of system usage for moments when they had a bit time to collect or reflect upon their thoughts. Several made a mental note to use it just before bed as *"something you can just vent to"* and release thoughts that may have otherwise stayed spinning (P7), or as a mechanism for *"decompression and reflection on the day"* (P10). Others tended to use it first thing during the work day (e.g., after an automatic calendar reminder), or during lunch as a way to *"take a break from work"* and process any negative thoughts getting in the way (P8). Finally, a couple participants reported using Flip*Doubt mainly in "bursts." For instance, P9 said he would typically enter 4 or 5 thoughts once every few days when thoughts that had been *"sitting on the back burner"* burst forth because his mind realized *"there's an outlet now for this sort of thing."* While participants in both groups told us that participating in the study was overall a positive experience, some also mentioned undesired effects. The next section teases apart beneficial and undesired impacts.

5.2 Impacts on participants

Here, we highlight two beneficial impacts of participation (boosts in mood, increased awareness) and two undesired effects (cognitive overload, prompting a negative focus).

5.2.1 Boosts in mood. When participants received a high quality reframe from Flip*Doubt, they reported that *"it was a nice little pick-me-up"* (P9) or *"a slight boost"* (P3). As P2 explained, *"the really positive ones definitely were uplifting and put me in a better mood,"* even if that sensation didn't last the entire day. P4 also emphasized that *"this little boost is definitely a positive feeling to have, even if it's a machine saying it."* Participants especially appreciated how the addition of humor to a reframe could flip their thinking. Take the following example from P1:

Input: *"No one wants to be near me because I stink."*
Reframe: *"People with a sensitive sense of smell are going to miss out on knowing me."*

P1 rated the reframe at 4.5 stars, and provided the reason, *"Very funny. I like that it frames it as their problem because they will miss out on knowing me."*

5.2.2 Increased awareness of thought patterns. Beyond temporary boosts in mood, many participants noted an even more important and meaningful impact on them–i.e. of helping them to develop a greater awareness of problematic thoughts or situations in their lives. For example, P7 said, *"[Flip*Doubt] helped me check in, like, 'How am I feeling?' Because sometimes I don't."* Similarly, P10 described how, at the beginning of the study, she had been so stressed out and overworked that she hadn't even realized how negative and disparaging her thinking had become. She said that Flip*Doubt *"initially helped me understand that I was not in a good place. That's the thing that I really need to work on–figuring out what are the things that are happening before it snowballs into a shit show."*

While all of our participants experienced one or both of these benefits at the beginning of the study period, some discontinued use of the system when they observed one or two negative aspects, as we will next describe.

5.2.3 Cognitive overload. Our participants were busy grad students, and some reported that they didn't use Flip*Doubt as frequently as they'd imagined because of cognitive overload related to:

(1) *Overthinking things:* P11 said she *"had to think too hard about what I'm putting in"* before typing her thoughts into the Flip*Doubt interface. She felt she needed to *pre*-reframe the thought in her own mind, in order to figure out the type of response she'd like to receive in advance. It was just too much thinking.
(2) *Repetitive thinking:* P4 noticed that he had been having the same negative thoughts repeatedly. He said that, *"once they started to repeat, I had trouble just coming up with new ones and got preoccupied with other stuff."*
(3) *Other pressing requirements of grad school:* P6, who only used the system twice, said that she simply became too busy with her degree requirements, and couldn't spare the time. She said, *"It just felt like another thing to do."*

All of these participants eventually stopped using the system in order to protect limited cognitive resources.

5.2.4 Prompting a negative focus. Some participants felt it was problematic to use a prompt that specifically asked for negative thoughts. For example, despite initial enjoyment of the system, P5 explained that it eventually felt like *"the thing is hungry for my negative thoughts."* He began to feel a sense of frustration and resentment that he was somehow supposed to be *feeding* it, therefore he intentionally stopped using it. Likewise, P10 felt that after the system had–*very usefully*–helped her develop awareness, she decided to improve her thoughts using her preferred offline tools (e.g., yoga, uplifting podcasts, daily journaling with pen & paper). She then felt it was not helpful to be forced to type out negative thoughts, which had stopped occurring to her with as much natural frequency.

Given these impacts, we asked if participants attributed any changes in their overall mood or wellbeing to using Flip*Doubt. Beyond occasional mood boosts, they weren't certain, or they didn't think so. We did not find any qualitative evidence that participating in the study had serious adverse effects, as our instrument evaluation also suggests.

5.2.5 Instrument evaluation. We administered the PHQ-9 and GAD-7 during the initial, midpoint, and final week of the study. The PHQ-9 instrument provides a measure of depression, with a scale ranging from a score (s) of 0 to 27. On average, participants had relatively low depression scores, which decreased slightly over the course of the study: $\bar{s}_{initial} = 6.384$, $\bar{s}_{mid} = 5.615$, $\bar{s}_{final} = 5.363$. The GAD-7 instrument provides a measure of anxiety, with a scale ranging from s of 0 to 21. On average, participants also had relatively low anxiety scores, which again decreased over the course of the study: $\bar{s}_{initial} = 5.384$, $\bar{s}_{mid} = 3.846$, $\bar{s}_{final} = 2.72$.

Paired t-tests were used to test for significance between measurements of both instruments. The tests were run for each pair of measurements taken over the course of the study (i.e., initial and midpoint, midpoint and final, initial and final). The t-test scores from the PHQ-9 measurements indicate that there were no significant changes observed. The t-test scores for the GAD-7 instrument indicate that a significant decrease in anxiety was measured only between the initial and final measurements ($p = 0.0263$, $t = 2.604$). (Note that one participant dropped out, and one did not complete final wellness surveys, thus this measurement reflects 11 participants, with $df = 10$.) We cannot assume a causal relationship between Flip*Doubt usage and this measured drop, however we are encouraged to observe that participants who felt like stopping did stop (as we had encouraged them during intake), and that we did not measure or observe any dangerous or harmful impacts on the wellbeing of participants. We also note that some participants concluded their semester shortly before the end of the study period, possibly contributing to decreased anxiety.

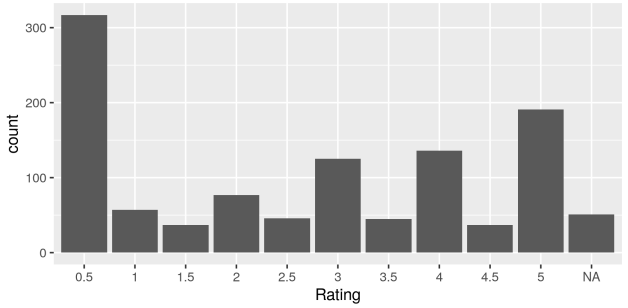

Fig. 3. Distribution of all Ratings.

6 THE RELATIONSHIP BETWEEN CONTEXT AND REFRAME RATINGS (RQ2)

In this section, we share how three types of contextual considerations impacted variation in reframe quality: (1) the *amount* of context in the input, (2) the *type* of contextual factors in the input, and (3) the *tactics* and *meta-behaviors* used to reframe the input. For each of these considerations, rather than empirical claims, we offer interpretations of the data that should be considered ***testable hypotheses for future work*** (bolded/italicized in-line) suggested by: (1) the data itself (i.e., trends in average ratings), and (2) our intuitions after our immersive experiences of analyzing input:reframe pairs and interview transcripts. Future work should carefully test and validate these hypotheses.

Before moving into these three considerations, we begin with a brief overview of how participants described the reasons for their highly variable ratings of reframes. (See Figure 3 for the distribution of all ratings, which had an overall average of $\bar{r} = 2.56$ stars, $\sigma = 1.72$. Participants enjoyed–and rated highly–reframes that both felt true to them, and helped to *"give me a different perspective."* (P3) This was especially salient when reframes provided novel ways of thinking about an issue, i.e., *"something where I hadn't thought of it, and it made me happier."* (P8) Participants especially loved reframes that were *"feisty and funny"* (P1) and reframes that felt almost as though they had been written in the participant's own voice. As P10 said, *"That's what I would have told myself to put myself in a different mindset,"* or as P5 put it, *"I could've written this on a better day."* On the other hand, they disliked reframes that *"try to discard or devalue what I said"* (P3) or low-effort responses that made participants feel like the crowd worker *"didn't even try."* (P9) Regarding middle-rated reframes, P8 said, *"If it was something that seemed relevant, and was sort of useful, but didn't feel particularly innovative, then it was sort of in the middle."* The next three sections describe how both the amount and type of context in the input, as well as crowd workers' specific choices of reappraisal tactics and meta-behaviors, impacted participants' ratings.

6.1 Total amount of context in the input

Participants told us that they felt they received better reframes when they took the time to intentionally add more contextual details to their input. For instance, one especially striking observation from the interviews was just how immediately (often without being asked) our participants tended to speculate about the experience of the AMT crowd workers. For example, P5 *"started to build a personality in my mind for the entity that was behind Flip*Doubt."* After using the system for a while, participants began to tailor thoughts for crowd workers, intentionally adding more detail *"so that there's more to work with"* (P6). They also began to think more deeply and delve beneath surface-level thoughts. For example, P8 felt the need to *"dig in a little deeper to come up with a better phrase than just 'I suck at my job.'"* Participants perceived that being more conscientious about

their inputs *did* lead to better reframes. For example, P11 shared that after she *"started changing what I was putting in–like trying to put in more context,"* she received more high quality responses. Consequently, our dataset of inputs includes a range of thoughts from broad, generic statements to specific, detailed statements. In this study, our coding protocol focused on the *type* but not overall *amount* of context provided in inputs; future work should also consider the total amount of context (e.g., low, medium, high) to test the hypothesis that: ***a higher total amount of context included in input thoughts leads to better reframes provided by others.***

6.2 Specific contextual factors of the input

In this section, we discuss how the input codes relate to the quality of reframes received. We begin by providing summary statistics and descriptions about the types of thoughts people had. Next, we show how contextual aspects of the inputs may have impacted ratings of the reframes received, possibly due to the *controllability* of the circumstances by the person providing the input, or the *relatability* of the circumstances to the person providing a reframe.

6.2.1 Summary of input thoughts. Figure 4 shows the counts of codes applied to input thoughts across the three input categories. We observe that the majority of thoughts relate to circumstances in the `Present` (59.0%), rather than the `Future` (23.9%) or `Past` (17.1%). The plurality of thoughts relate to `Work or School` (41.6%). Our observations while coding suggest that this category of thoughts seemed darker and more repetitive than other categories–a cause for significant concern. For example, one input from P3 was, *"I feel like grad school has made me hate science, when I once used to enjoy it, and I hate that I've become like this."* These `Work or School` thoughts often expressed endless stress and overwhelm, the worthlessness of their work, feelings of being demeaned by advisors, or a sense of despair, like *"slowly being ground down to nothing."* (P9) Thoughts on `Generic Emotions` (22.2%) and `Personal Relationships` (e.g., non-work related family, friendships, romance) (20.4%) were also common, emphasizing non-grad school specific concerns related to everyday circumstances or interpersonal interactions. Thoughts about `Health and Appearance`, `Managing Home Life`, and `Finances` were relatively rare. At a higher level, most of the meta-topics of these thoughts were `Generic Complaints` (34.0%) about negative aspects of participants' experiences, `Expressing Uncertainty or Worry` (22.5%), or `Self-Disparagement` (21.4%), rather than `Ruminating on Others' Thoughts`, `Expressing Regret`, or `Comparison with Others`, suggesting that participants tended to use Flip*Doubt on self-directed thoughts about current or future circumstances, rather than thoughts directed on others or regrettable past events. However, it is also worthwhile to note that different participants experienced different types of negative thoughts (see supplemental analysis in Section C of the Appendix).

Given this set of input data, Figure 5 shows normalized distributions of ratings applied to reframes they received. This figure suggests that contextual factors of the inputs may have impacted participants' perceptions of the quality of reframes they received. We next provide two hypotheses for *why* that may be.

6.2.2 Controllability. In exit interviews, participants told us that one key aspect to a quality reframe lies in its ability to shift their perspective. We offer one hypothesis that: ***the more someone feels that the circumstances of a negative thought are within their control to change, the more likely it is that they will perceive a reframe to be helpful or effective.*** For example, consider the temporal codes. In our sample, average rating was higher for reframes of thoughts about the `Future` ($\bar{r} = 2.76$ stars, $\sigma = 1.68$) vs. `Present` ($\bar{r} = 2.54$, $\sigma = 1.73$) or `Past` ($\bar{r} = 2.37$, $\sigma = 1.74$). It may be more difficult to convincingly shift someone's perspective about past events that cannot be controlled, whereas events in the present, or especially in the future, are more fluid and changeable. Similarly, people might feel more able to do something about their `Personal Relationships`

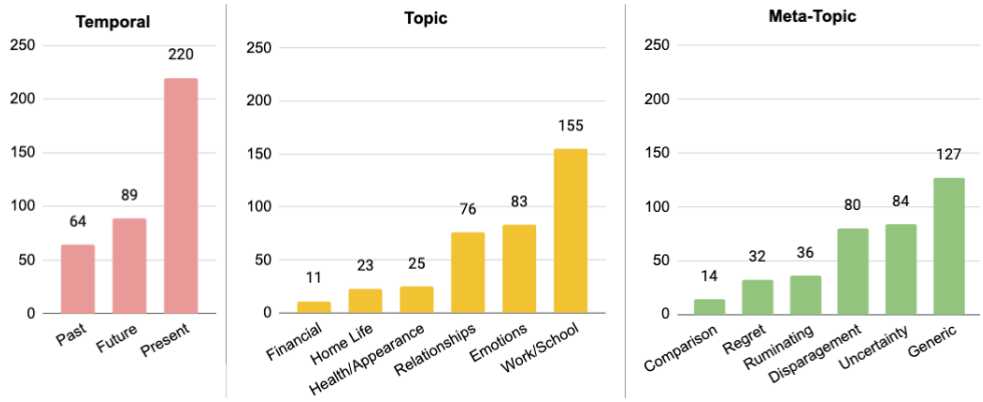

Fig. 4. Summary Counts of Input Codes. Grouped according to Temporal, Topic, or Meta-Topic categories.

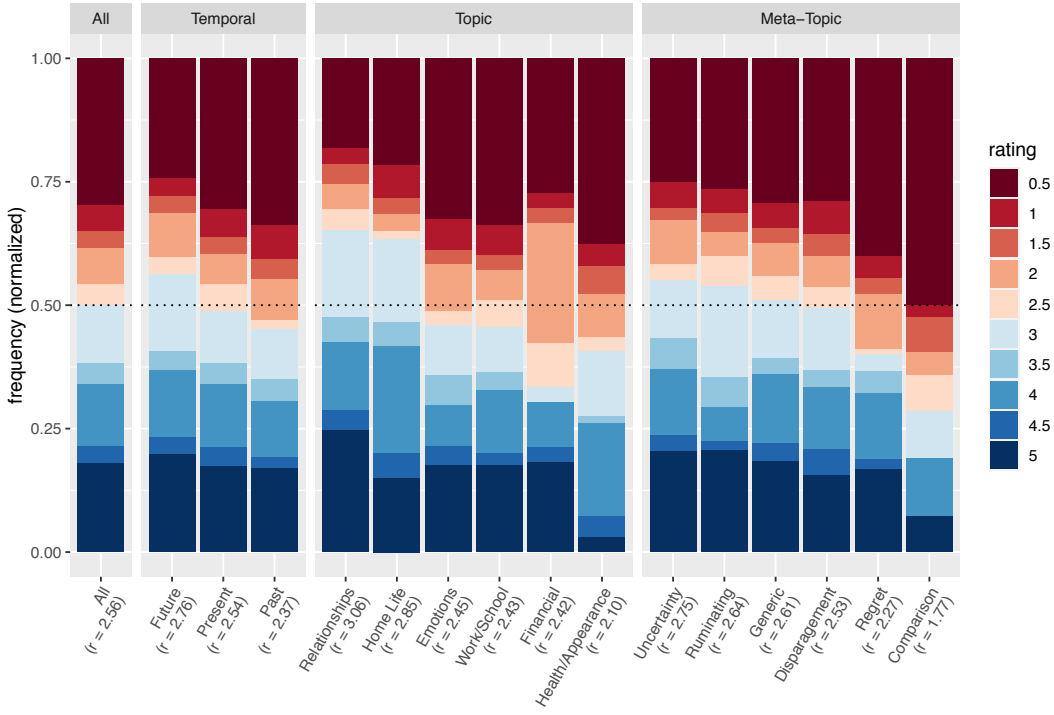

Fig. 5. Normalized Distributions of Ratings by Input Codes. Grouped according to Temporal, Topic, or Meta-Topic categories.

($\bar{r}$ = 3.06, σ = 1.63) or `Home Life` ($\bar{r}$ = 2.85, σ = 1.64), whereas they may perceive that their `Work/School` ($\bar{r}$ = 2.43, σ = 1.75), `Finances` ($\bar{r}$ = 2.42, σ = 1.67), or `Health/Appearance` ($\bar{r}$ = 2.10,σ = 1.54) are less within their control. It might be easier to influence one's own uncertain thoughts or thoughts about what others are thinking (`Uncertainty/Worry` ($\bar{r}$ = 2.75, σ = 1.71)

or Ruminations on Others' Thoughts ($\bar{r}$ = 2.64, σ = 1.69)), but more difficult to reduce self-criticism, to undo things they regret, or to change what they perceive to be true about themselves in comparison to others (Self-Disparagement ($\bar{r}$ = 2.53, σ = 1.71), Expressing Regret ($\bar{r}$ = 2.27, σ = 1.80), Comparison with Others ($\bar{r}$ = 1.77, σ = 1.55).

6.2.3 Relatability. Another hypothesis is that: ***crowd workers may provide better reframes when they can better relate to the context of the input thought.*** For example, perhaps Personal Relationships or Managing Home Life are more ubiquitously relatable and relevant to anonymous crowd workers than, say, issues related to Health and Appearance that are potentially more unique or stigmatizing. Furthermore, the particular issues related to Work or School experienced by grad students may not be well understood by the average, random crowd worker. For instance, P11 said, *"I felt like I was throwing too much at them,"* when she added too many specific details about the complex thoughts she has about her experiences in grad school. Echoing results from prior work [4, 67], to solve the issue of making sure the *right* crowd workers were flipping their thoughts, some participants suggested that Flip*Doubt should match crowd workers and Flip*Doubt users based on *shared context*–e.g., same type of job, or similar type of home life–or to "like" and request a certain crowd worker, to get that person again in the future.

6.3 Tactics and meta-behaviors used by crowd workers to reframe inputs

We begin this section with a quote that highlights its main takeaway–i.e., that certain reappraisal tactics and behaviors seem to work better than others. For instance, P9 described a particular type of thought he'd input a number of times:

> *I would say something like, "Oh, this PhD thing, I feel like all of my research ideas are garbage and I'm never gonna finish," and the reframe would be,* **[A]** ***"Keep persevering! You can do it!"*** *which to me is not helpful. For some people, that might be what they need. But something more like,* **[B]** ***"You know grad school is really hard now, but the pay off at the end of it is gonna be totally worth it,"*** *which to me is like, "Okay yeah, yes it will."*

Reframe A uses the Agency tactic, whereas reframe B uses the Change Future Consequences tactic (as well as the Acknowledge Main Concern meta-behavior). In practice, either tactic might be "needed," but for P9, only B seemed to resonate as true or helpful. We next explore why some tactics and meta-behaviors may be more or less helpful.

6.3.1 Reappraisal tactics. Figure 6 shows the counts of reframe codes applied, while Figure 7 shows distributions of ratings by reframe codes. As suggested by the quote above, the perceived effectiveness of some reappraisal tactics could be due to personal preferences–e.g., even though B is better for P9, A could be better for someone else. This suggests the hypothesis that: ***different reappraisal tactics may be more effective for different people or different situations.*** Yet Fig. 7 also suggests that there may be general trends across participants. For instance, consider the example presented in Figure 2. In this case, crowd workers happened to select three different tactics. The reframe labeled with the Acceptance tactic (rated 4.5 stars) used a new idea with new words that completely flipped the perspective on the problem and created a sense of *"positive affirmation"* for P2. On the other hand, the Agency reframe (2 stars) used an analogous sentence structure to the input, and simply asserted P2's ability to change her behavior. The Direct Negation reframe (0.5 stars) seems to be perceived to P2 as a *"lie"*.

In general, we observed that reframes with the highest rated codes applied—for example, Silver Lining ($\bar{r}$ = 3.25, σ = 1.52), Acceptance ($\bar{r}$ = 3.18, σ = 1.67), Change Future Consequences ($\bar{r}$ = 3.02, σ = 1.63), and Technical Analytic Problem Solving ($\bar{r}$ = 2.98, σ = 1.73)—often

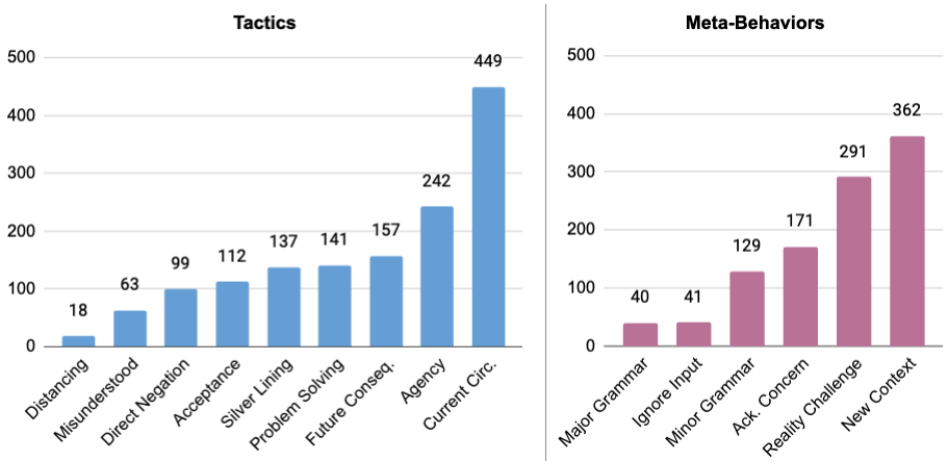

Fig. 6. Summary Counts of Tactics and Meta-Behaviors.

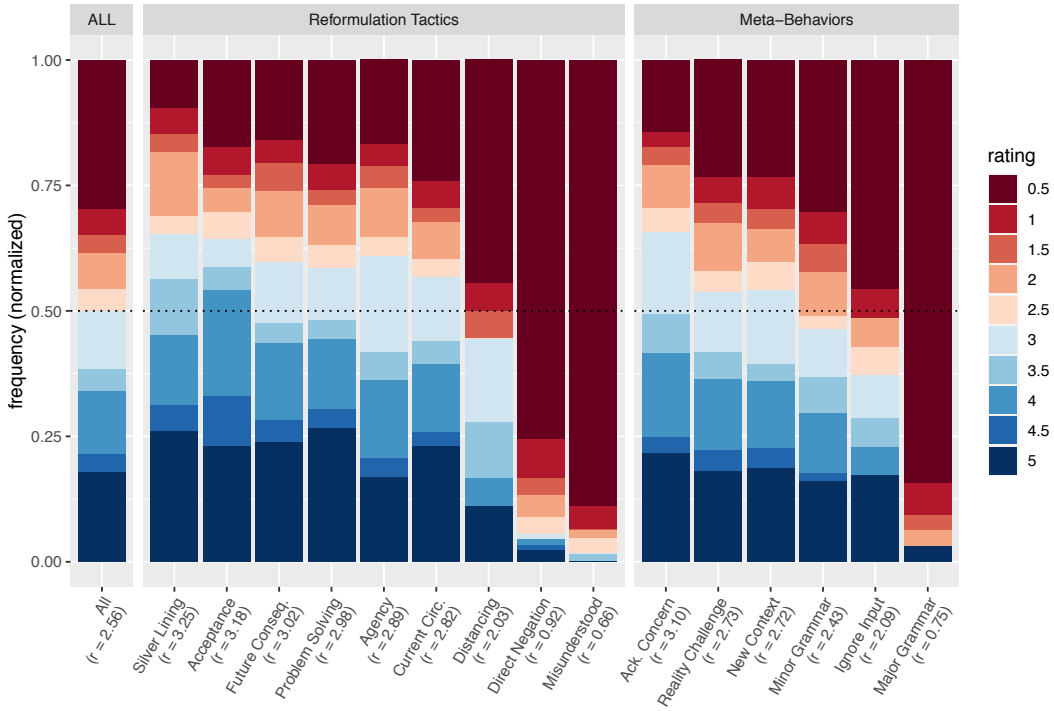

Fig. 7. Normalized Distribution of Ratings by Tactic Codes.

seemed more nuanced, bringing in new ideas, perspectives, or solutions. We hypothesize that: ***tactics that substantively bring in new perspectives not present in the original thought will be more successful.*** These four tactics were also less common, with each code individually applied to 10-14% of reframes. Despite being the most common tactics chosen at 40.2% and 21.7% respectively, `Change Current Consequences` ($\bar{r} = 2.82$, $\sigma = 1.72$), and `Agency` ($\bar{r} = 2.89$, $\sigma = 1.53$) received

somewhat lower ratings than the first four codes mentioned. `Change Current Consequences` reframes varied widely from insightful ways to re-prioritize or view a situation differently, to odd statements that shifted focus without quite capturing the point of what the input was getting at (e.g., pithy generic true-isms). As in the Fig. 2 example, we observed that many `Agency` reframes were relatively simple phrases (with a possibly higher-than-desirable degree of textual similarity [23]) that assert ability, possibly making them seem less thoughtful, even if relevant. It is important to note that the higher frequency of `Change Current Consequences` and `Agency` reframes may be due to a learning bias, since the two examples in the HIT instructions used these tactics (see Appendix Section B.2).

On the other hand, possibly as a result of seeming dismissive of the concern presented in the input, `Distancing` ($\bar{r} = 2.03$, $\sigma = 1.69$) received relatively lower ratings, and was quite rarely used (1.6%).[4] `Direct Negation` ($\bar{r} = 0.92$, $\sigma = 0.98$) reframes were described as *"total garbage"* (P9), but nonetheless occurred in about 8.9% of reframe statements. Participants emphatically described how `Direct Negation` statements are simply untrue and offer absolutely nothing useful. Encouragingly, only 5.6% of reframe statements were coded `Misunderstand Instructions` ($\bar{r} = 0.66$, $\sigma = 0.54$), and these were (unsurprisingly) usually rated 0.5.

6.3.2 Meta-behaviors. Meta-behaviors also impacted perceptions of quality. For instance, we developed the code `Acknowledge Main Concern` ($\bar{r} = 3.10$, $\sigma = 1.53$) after several participants noted that this behavior was especially beneficial because it demonstrated real understanding and validation of the problematic stressor. For example, P7 said, *"If they validate my feelings first, and then give me a positive spin, that felt more reassuring to me."* This aligns with recent work emphasizing *empathy* in supportive messages [58, 86] and supports the hypothesis that: ***reframes that both demonstrate deep understanding of the input problem and suggest new perspective will be more successful.*** We applied `Acknowledge Main Concern` to only 15.3% of reframes. However, crowd workers more frequently made unsubstantiated assumptions about participants based on some aspect of their input. We applied `Introduce New Personal Context` ($\bar{r} = 2.72$, $\sigma = 1.67$) to 32.4% of reframes in situations when these assumptions translated into the reframes as brand new details or context that had not been mentioned in the input (e.g., gender, personality traits, or situational facts about the participant). We applied `Reality Challenge` ($\bar{r} = 2.73$, $\sigma = 1.67$) to 26% of reframes in cases when the reframe undermines or challenges the nature of the input thought to suggest that something else could be true. Participants' comments suggest that both of these codes can be either strikingly useful (and seem quite serendipitous) if they happen to feel true and relevant, or frustrating and useless if they happen to feel false or unfounded. As P8 described these types of additional details or underlying assumptions added by crowd workers, *"Sometimes that matched well, and sometimes that didn't."*

`Minor Grammar` ($\bar{r} = 2.43$, $\sigma = 1.70$) mistakes appeared in 11.5% of reframes, but seemed to not penalize ratings too harshly compared to the overall average. On the other hand, reframes with the `Ignore Input Issues` ($\bar{r} = 2.09$, $\sigma = 1.78$) code and especially the `Major Grammar` ($\bar{r} = 0.75$, $\sigma = 0.84$) code lowered ratings substantially, though they each occurred in fewer than 4% of reframes.

6.4 Summary of hypotheses generated through our analysis of inputs and reframes

We conclude our results with a recap of the hypotheses generated by this work that can be tested in future studies:

[4]Note that `Distancing` also occurred as a secondary tactic in one of the examples present in the HIT instructions

[H1] A higher total amount of context included in input thoughts will lead to better reframes provided by others.
[H2] The more someone feels that the circumstances of a negative thought are within their control to change, the more likely it is that they will perceive a reframe to be helpful or effective.
[H3] Crowd workers may provide better reframes when they can better relate to the context of the input thought.
[H4] Different reappraisal tactics may be more effective for different people or different situations.
[H5] Tactics like `Silver Lining`, `Acceptance`, `Change Future Consequences`, `Technical Analytic Problem Solving` that substantively bring in new perspectives not present in the original thought will be more successful.
[H6] Reframes that demonstrate deep understanding of the input problem, as well as a reappraisal tactic, will be more successful.

7 DISCUSSION

As incidence rates of mental illnesses continue to rise at alarming rates, sociotechnical systems offer substantial promise to expand models of delivery for effective interventions [44]. Given excessive barriers to accessing professional care, technology can provide new ways for untrained peers and lay people to provide support for mental illness [4, 24, 67, 68]. Yet very little prior work has leveraged novel collaborative or peer-based systems for providing well-established therapeutic interventions like cognitive reappraisal. In this work, we described our deployment of the Flip*Doubt prototype. Despite the time, challenges, and ethical concerns associated with building and deploying prototypical systems for mental health, insights from this type of work are crucial toward developing practical and scalable real-world solutions. Therefore, we now discuss ways that our results can guide and support future work to: (1) design systems that leverage peer training and support to help people change negative thinking patterns, and (2) incorporate AI/ML-based algorithms to increase the effectiveness, safety, and scalability of such systems.

7.1 Implications for systems research to leverage peer training and support

In our deployment, we found that participants experienced both positive and negative impacts of system usage as a consequence of design decisions we made for the study. To increase positive and reduce negative outcomes, our results point to two important implications for future research around: (1) prompting for *reflection* rather than negativity; and (2) providing opportunities for skills training and practice, in addition to receiving reframes from peers.

7.1.1 Prompting for reflection rather than prompting for negative thoughts. Participants in our study tended to "bank" negative thoughts to submit to Flip*Doubt at a later time, rather than popping open the web app the moment they had a negative thought. This behavior may generalize to other types of system designs, since negative thoughts can arise at any time, particularly at high-stress moments when using an app may be overly disruptive, inconvenient, or even impossible. Thus, participants who continued to use Flip*Doubt rather than abandoning it often did so either by intentionally integrating it into specific moments of their daily routines when they felt they would benefit from a reflective pause, or by using it in bursts when the floodgates would open up and issues on the backburner could all pour out. In future studies, providing a more structured experience for thought collection could facilitate such reflective pauses. For example, we imagine a plug-in that scaffolds the creation of repeating events. An initial system prompt could ask users, *"What is a regular time in your daily life when you'd like to pause to collect your thoughts?"* and provide

suggestions like: first thing in the morning, lunch break, just before bed. At the appointed time, another system prompt or notification could trigger interaction with the system–our results also point toward ways to *refine* that trigger.

For example, no participants initially took issue with our prompt to purge their negative thoughts. However, because people didn't use the system at the exact moment when they experienced negative thoughts, this meant that they needed to later revisit negative thoughts. This has the significant con of essentially making them *re-experience* negative thoughts a second time by writing them down, and possibly ingraining them even more deeply. As described in Sec. 5.2.4, some people grew to resent this focus on negativity over time, even if the system caused them to initially develop a useful and desirable awareness of their negative thought patterns. Thus, we suggest that after an initial period focused on identifying negative thoughts and developing awareness, a better prompt may be, *"Would you like to check in with your thoughts right now?"* This prompt does not oblige people to have negative thoughts on a given day, and could allow people to begin to own more positive thought patterns as they develop. Of course, if people enter positive thoughts into the system, the system response should not need to involve cognitive reappraisal at that point. An interesting direction for future work can be to determine appropriate psychological techniques for peer responses to *positive* thoughts to best support an individual's continued progress—e.g., affirmation of progress, congratulations, additional questions to provoke more insight, or possibly even nudges to do other beneficial offline activities and wellness or stress management practices (as in [71], for example).

7.1.2 Opportunities for skills training and practice. Furthermore, even though participants enjoyed mood boosts, occasional laughs, and even reductions in negative thinking, we observed only minimal evidence that Flip*Doubt users actually learned or applied the skill of cognitive reappraisal through receiving reframes from others. Conversely, some participants described cognitive overload (Sec. 5.2.3) which caused them to stop using the app. Some investment of cognitive work is necessary to address negative thought patterns; however, it is possible that the open and unrestricted nature of how we asked participants to use the system contributed to a feeling of overwhelm or overthinking, while our focus on collecting as many negative thoughts as possible contributed to repetitive thinking. Therefore, we suggest that: (1) providing more structure to the reflective pauses described above might afford better opportunities for self-reflection and skills practice and learning; (2) it may be more worthwhile to engage deeply with a few core negative thoughts than to try to capture and process many negative thoughts regularly.

For instance, some participants learned to "pre-reframe" their own thoughts, and then adjust the way they entered inputs into the system, due to awareness of how humans on the other end would consider them. This self-reflection may be a useful activity in and of itself, and could potentially feel less overwhelming if it were directly supported by the system. For example, a simple dialogue box or pop-up could appear after entering input that asks: *"If you were to read this thought as though it weren't yours, do you think it provides enough context to respond to in a helpful way?"* An even better way would be to ask participants to critically evaluate their input by completing a labeling exercise–e.g., they could answer questions about whether the thought is about something in the `Present`, `Future`, or `Past`, and what the topic (e.g., `Personal Relationships`, `Work or School`, etc.) and meta-topic (e.g., `Self-Disparagement`, `Expressing Regret`, etc.) of the thought are. Pinpointing these issues could help users develop a greater awareness of patterns that ail them, and may better prepare them to reframe their thoughts. Another excellent way to cause users to deeply engage could be to ask them to provide a reframe of their own thought (as in Intellicare Thought Challenger [56]) *before* sending it to the crowd. Likewise, labelling their own reframes with our codebook tactics could make them more consciously aware of their own default

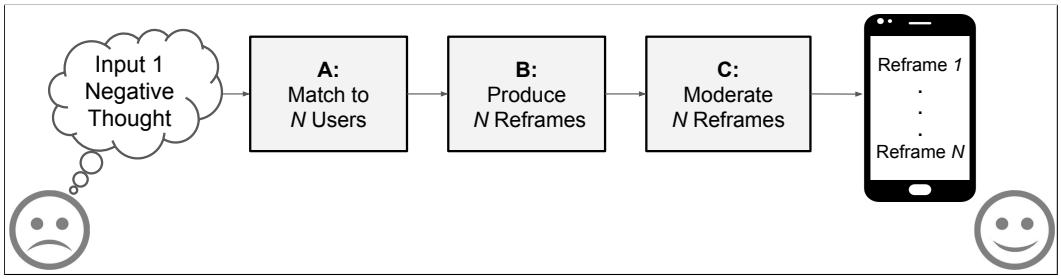

Fig. 8. A Human Computation Pipeline for Cognitive Reappraisal Platforms.

manner of reframing, and allow them to compare with supportive reframe tactics from others to maximize their learning.

Research also shows that providing extrinsic emotional regulation for *others* is also beneficial for the supporter [41, 57, 66, 100]. In particular, reappraising *others'* thoughts can be highly effective for improving one's *own* well-being [5], possibly because cognitive distance allows for more flexible or less distorted thinking, which may then be transferable to one's own situation [49]. We originally considered a study design in which participants reappraised others' thoughts, however our small sample size and recruitment methods led to anonymity concerns; we wanted participants to feel they could express thoughts freely without social repercussions. Future work in more anonymous contexts (e.g., Koko [24]) can build on insights from this work to support cognitive reappraisal tasks for self and others.

7.1.3 Implications for psychology. Beyond implications for sociotechnical systems, our results and codebooks may also be useful for psychology research and practice more broadly. For example, our results can inform the development of more extensive training resources for cognitive reappraisal. In CBT/DBT, clients often complete take-home pen-and-paper worksheets or skills packets [11, 80]; a one- or two-page guide to selecting reappraisal tactics could be developed to supplement existing materials. Such a guide could be enriched and validated by future research that statistically tests the hypotheses we present in this work (see Sec. 7.3.5). It could also easily be translated into an online format or module and included in online versions of CBT/DBT [20, 85], or in a "resources" section of an online community or collaborative peer-based platform for mental health, to provide the much-needed training called for by prior work [67].

In this section, we have discussed design strategies for improving cognitive reappraisal systems in order to potentially reduce cognitive overload and an overly negative focus. These strategies could support greater self-reflection and skill learning for users receiving reframes, however they mostly do not yet speak to our aim of developing mechanisms for safely and effectively involving peer supporters. Consequently, we next propose and describe opportunities for AL/ML-based algorithms that could help with this goal.

7.2 Integrating AI/ML-based algorithms in platforms for cognitive reappraisal

Recent literature reviews on mental health and machine learning suggest that most papers sought to understand, detect, and diagnose users' mental health states, offering contributions for initial technical development of algorithms [83, 93]. Only a tiny fraction deployed end-to-end algorithmic systems or sought to improve available treatments, suggesting important opportunities for *"more effective tailoring of interventions to each person's unique mental health and support needs."* [93] Figure 8 outlines a three-step human-computation pipeline for use in peer-based platforms for

cognitive reappraisal. Although such a pipeline is not itself novel, we describe ways in which our results can help to improve either the effectiveness and personalization (Steps A and B) or safety (Step C) of cognitive reappraisal platforms like Flip*Doubt, Koko, or future systems.

7.2.1 Step A. After a user inputs a negative thought, the first step is to match the thought with N users to provide reframe(s). This leads to two major considerations:

(1) **Matching mechanisms.** Our results suggest that *relatability* may be one factor affecting reframe quality. Rather than random matching (as in Flip*Doubt and Koko), matching users based on shared contextual aspects of their lives or identities [4, 51, 67, 88], or allowing users to identify people they'd like to interact with again, may improve reframe quality by allowing crowd workers to better understand the specific context of the task [82].
(2) **Selecting N.** In this study, we selected N=3 to generate more data about reappraisal behaviors, given a limited amount of input. Participants commented that it was helpful to see multiple reframes–especially when they employed different tactics or perspectives. Consequently, selecting $N > 1$ may be educationally useful. However, selecting too large of an N might require too much community effort or result in choice overload [84]. Future research can determine the best value of N to support positive outcomes for skill learning.

7.2.2 Step B. The second step is to produce N reframes. Assuming that many users have never received clinical training in cognitive reappraisal, the purpose of algorithms at this step should be *assistive*. For example, findings from Flip*Doubt suggest mechanisms similar to the "Assess" or "Recommend" modes of MepsBot [73]. As a case in point, participants told us that the `Acknowledge Main Concern` meta-behavior improved reframes. One way to generate this meta-behavior could be a two-step crowd work process (i.e., first create an acknowledgement statement, then add a reframe tactic). A second way could be by inserting additional instructions, although more than one or two lines of instructions may result in key points being neglected.[5] We suggest two ways that algorithms could reduce cognitive load, while still providing a valuable learning experience:

(1) **Generating Partial Reframes.** Similar to the *empathetic rewriting* task in [86], an algorithm could identify the main issue in an input thought, and composes a phrase of validation. Say an input were *"Grad school is too hard."* An algorithm could generate a phrase like, *"Even though grad school feels hard, ..."* and then allow a human to compose the rest. This approach could encourage use of a helpful meta-behavior and simultaneously discourage use of the damaging `Direct Negation` tactic, or `Ignore Input Issue` meta-behaviors.
(2) **Suggesting Tactics.** An algorithm could assess characteristics of the user and/or input, and next suggest specific tactics to the re-appraiser. For example, assume that tactics like `Technical Analytic Problem Solving` or `Change Future Consequences` are better for topics related to `Work or School` than say, `Change Current Circumstances`. Rather than generic instructions, the system could embed specific tactics into more refined instructions–e.g., *"Please reframe this thought by proposing a way that this person could solve the described problem"* or *"Please reframe this thought by describing positive future consequences."* Alternatively, it might provide *examples* that use those tactics to induce a learning bias, as may have occurred in our study.[6]

[5]For instance, after our main instruction to *"Rewrite this thought in a way that is more positive,"* our HIT instructions explicitly stated that crowd workers should *not* directly negate the input thought. Yet `Direct Negation` nonetheless occurred in 8.9% of reframes.

[6]Our data does not provide an easy way to assess learning bias. However the high prevalence of `Change Current Circumstances` and `Agency` codes, given that examples of these tactics appeared in the instructions, suggests that some learning bias may be present.

7.2.3 Step C. Similar to Reddit and Wikipedia [19, 38, 91], semi-automated moderation could help to prevent ineffective or abusive reframes. In a system like Panoply/Koko, we could design algorithmic systems to screen toxic or unhelpful reframes–e.g., reframes predicted as `Misunderstand Instructions`, `Ignore Input Issue(s)`, `Direct Negation`, or `Major Grammar`.

7.3 Ethics, Limitations and Future Work

7.3.1 Data sources. Our dataset of input:reframe pairs is not large enough for training AI/ML models or building applications. Acquiring large training datasets is a major challenge for ML in mental health [93]. One strategy is to retrieve sets of public content from online communities (such as Reddit, as in [73]). However, recent controversies [43] and research [17, 93] highlight ethical tensions for building models that rely on large sets of public data, due to known biases in the data (e.g., gender and race) [12, 16]. Moreover, online community users often have not *knowingly* consented to this data use; whether/how to responsibly use publicly available data is an ongoing ethical discourse in HCI [28–30, 90, 102]. In this work, we recruited and consented participants. We explained how data would be used, and participants decided whether/how to use the system accordingly. Future work can continue to source data from a broader set of consenting participants, either through additional field studies, or collaborating with existing groups (e.g., CBT/DBT groups) or new or existing online communities. Efforts should be made to recruit diverse populations who are empowered to flag and withhold problematic data, and to take a human- and community-centered approach to building, deploying, and evaluating algorithms [101].

7.3.2 Ethical considerations with potentially vulnerable populations. The graduate student population experiences higher rates of mental illness, yet our protocol excludes those with any more than "minimal" suicide risk–a choice also described in [64]. However, this choice might be argued to run counter to the Belmont Principle of Justice [39], which is intended to ensure that research participants can possibly benefit from research. P10 noted concerns about this:

> *I wonder if [Flip*Doubt] wouldn't actually be more helpful for the people you disqualified from the study...That could have actually changed someone's life in a moment when they really needed it. I don't feel like I'm on the edge where I need that, but it's kind of a weird form of discrimination, for something about mental health.*

For these reasons, a recent consensus statement from domain experts recommends *not* to exclude such individuals or remove them from research against their will, and provides additional guidance for working with potentially suicidal participants [65]. Given this, how can we best balance risks to all involved, including participants, moderators, therapists, and even researchers, whose mental health can be seriously compromised while working in this domain [99]? Pinning down ethical processes for selection, training, and compensation criteria for participants, moderators, and researchers is an essential open question for future work.

7.3.3 Sample Limitations. Our small sample is not representative of all graduate students, nor does it capture all possible diversity of race, gender, age, or discipline. Graduate students are also not representative of the general population, while other populations would likely experience very different thoughts and perceptions. Future work should expand to include both broader populations (e.g., general public), as well as more niche, marginalized, and/or stigmatized communities.

7.3.4 Paid incentives. Paid incentives may have impacted the quality or quantity of thoughts input by Flip*Doubt users; some negative inputs could have been "manufactured" rather than truly organic system usages. Likewise, paid incentives for crowd workers likely influenced their motivations and the quality of task completion. In comparison with Koko [24, 58], Flip*Doubt

seemed to receive a higher volume of poorly rated reframes. *Altruistic motives* may mediate better responses, thus future work can address this limitation by not relying on paid incentives.

7.3.5 Assessing and labeling future cognitive reappraisal datasets. Our study used ratings as an informative starting point for assessing the quality of reframes. Although ratings can provide a useful model of internal preferences [47], they also have some intrinsic limitations. For instance, ratings are imperfect measures of internal preferences [26, 46] that are subject to psychological effects [22] and that can shift over time [3]. To address these limitations, future work should use more refined assessments of reframes. For example, our work suggests the importance of the following factors: (1) introducing new perspective, (2) personal relevance to the user, (3) humor or cleverness, and (4) quality of the writing itself. Using Likert scales on these factors to complement or replace ratings would be highly informative.

Furthermore, our field study of Flip*Doubt necessarily relies on manual researcher annotation to initially generate codebooks. However, manual researcher annotation is impractical on large datasets. In Sec. 7.1.2, we suggest that participants could provide labels on their own data (both inputs and reframes); such a labeling activity would substantially bolster future research efforts by reducing the need for manual annotation, while also creating a more user-centric model for how researchers engage with data for ML, as called for by [93]. While our codebooks are an excellent starting point, a broader deployment could also reveal new categories; therefore providing a `‘‘Something else’’` option with a free-response field would allow participants to surface new categories. Expanding the evaluation process in this way would likely be too cognitively taxing on users if we ask for a high volume of inputs. Future research can collect smaller numbers of inputs from each participant, and use $N = 1$ to gather independent assessments of input:reframe pairs. Such a study design would enable better statistical and hypothesis testing, and allow researchers to study which tactics work best for different types of inputs, individuals, or situations.

8 CONCLUSION

This paper has provided an in-depth exploration of how a sample of 13 graduate students interacted with Flip*Doubt–a novel crowd-powered system for cognitive reappraisal. Our month-long deployment contributes in-depth knowledge about how participants used the system in the wild and new codebooks to describe contextual aspects of negative thoughts and reformulation tactics for positive reframes. These codebooks can help to shape future analyses and methods for gathering, analyzing, and labeling larger datasets. Finally, we outline important implications for future systems research and for innovation in AI/ML to support cognitive reappraisal. All together, we hope this work can help guide future researchers and innovators to build and deploy effective interventions and systems for mental health support and recovery.

ACKNOWLEDGMENTS

We thank our participants for sharing their thoughts and our reviewers for their supportive feedback. We are grateful to Xinyi Wang and Karan Jaswani who made excellent coding contributions to a preliminary Flip*Doubt prototype built for a web development course taught by our wonderful instructor, F. Maxwell Harper. We also thank Stephen Schueller for his preliminary comments that helped shape our research direction, and our undergraduate research assistants, Benjamin Wiley, Eric Sortland, Shubhavi Arya, and Ishan Joshi. This work was partially funded by the University of Minnesota Social Media Business Analytics Collaborative (SOBACO) faculty award.

REFERENCES

[1] Vincent I. O. Agyapong, Kelly Mrklas, Michal Juhás, Joy Omeje, Arto Ohinmaa, Serdar M. Dursun, and Andrew J. Greenshaw. 2016. Cross-sectional survey evaluating Text4Mood: mobile health program to reduce psychological treatment gap in mental healthcare in Alberta through daily supportive text messages. *BMC Psychiatry* 16, 1 (Dec. 2016), 1–12. https://doi.org/10.1186/s12888-016-1104-2 Number: 1 Publisher: BioMed Central.

[2] Felwah Alqahtani and Rita Orji. 2020. Insights from user reviews to improve mental health apps. *Health Informatics Journal* 26, 3 (Sept. 2020), 2042–2066. https://doi.org/10.1177/1460458219896492 Publisher: SAGE Publications Ltd.

[3] Xavier Amatriain, Josep M. Pujol, and Nuria Oliver. 2009. I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization: formerly UM and AH (UMAP '09)*. Springer-Verlag, Berlin, Heidelberg, 247–258. https://doi.org/10.1007/978-3-642-02247-0_24

[4] Nazanin Andalibi and Madison K. Flood. 2021. Considerations in Designing Digital Peer Support for Mental Health: Interview Study Among Users of a Digital Support System (Buddy Project). *JMIR Mental Health* 8, 1 (Jan. 2021), e21819. https://doi.org/10.2196/21819 Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.

[5] Reout Arbel, Marlyn Khouri, Jasmine Sagi, and Noga Cohen. 2020. *Reappraising Others' Negative Emotions as a way to Enhance Coping during the COVID-19 Outbreak.* Technical Report. PsyArXiv. https://doi.org/10.31234/osf.io/y25gx type: article.

[6] Sara Atanasova and Gregor Petric. 2019. Collective Empowerment in Online Health Communities: Scale Development and Empirical Validation. *Journal of Medical Internet Research* 21, 11 (2019), e14392. https://doi.org/10.2196/14392 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[7] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. *arXiv:2102.08368 [cs]* (Feb. 2021). https://doi.org/10.1145/3442381.3450122 arXiv: 2102.08368.

[8] Marco Bardus, Jane R. Smith, Laya Samaha, and Charles Abraham. 2015. Mobile Phone and Web 2.0 Technologies for Weight Management: A Systematic Scoping Review. *Journal of Medical Internet Research* 17, 11 (2015), e259. https://doi.org/10.2196/jmir.5129 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[9] Jennifer Beckett. 2018. We need to talk about the mental health of content moderators. http://theconversation.com/we-need-to-talk-about-the-mental-health-of-content-moderators-103830

[10] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST '10)*. Association for Computing Machinery, New York, NY, USA, 313–322. https://doi.org/10.1145/1866029.1866078

[11] Martin Bohus, Brigitte Haaf, Timothy Simms, Matthias F. Limberger, Christian Schmahl, Christine Unckel, Klaus Lieb, and Marsha M. Linehan. 2004. Effectiveness of inpatient dialectical behavioral therapy for borderline personality disorder: a controlled trial. *Behaviour Research and Therapy* 42, 5 (May 2004), 487–499. https://doi.org/10.1016/S0005-7967(03)00174-8

[12] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Neural Information Processing Systems (NIPS)* (2016), 9. https://papers.nips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[13] Jonaki Bose. 2017. Key Substance Use and Mental Health Indicators in the United States: Results from the 2017 National Survey on Drug Use and Health. (2017), 124.

[14] Dena M. Bravata, Sharon A. Watts, Autumn L. Keefer, Divya K. Madhusudhan, Katie T. Taylor, Dani M. Clark, Ross S. Nelson, Kevin O. Cokley, and Heather K. Hagg. 2020. Prevalence, Predictors, and Treatment of Impostor Syndrome: a Systematic Review. *Journal of General Internal Medicine* 35, 4 (April 2020), 1252–1275. https://doi.org/10.1007/s11606-019-05364-1

[15] Jonathan B. Bricker, Kristin E. Mull, Julie A. Kientz, Roger Vilardaga, Laina D. Mercer, Katrina J. Akioka, and Jaimee L. Heffner. 2014. Randomized, controlled pilot trial of a smartphone app for smoking cessation using acceptance and commitment therapy. *Drug and Alcohol Dependence* 143 (Oct. 2014), 87–94. https://doi.org/10.1016/j.drugalcdep.2014.07.006

[16] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html ISSN: 2640-3498.

[17] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 79–88. https://doi.org/10.1145/3287560.3287587

[18] Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016. Recovery Amid Pro-Anorexia: Analysis of Recovery in Social Media. *Proceedings of the SIGCHI conference on human factors in computing systems . CHI Conference* 2016 (May 2016), 2111–2123. https://doi.org/10.1145/2858036.2858246

[19] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 174:1–174:30. https://doi.org/10.1145/3359276

[20] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. 2020. Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376341

[21] Helen Christensen, Kathy Griffiths, Chloe Groves, and Ailsa Korten. 2006. Free Range users and One Hit Wonders: Community Users of an Internet-Based Cognitive Behaviour Therapy Program. *Australian & New Zealand Journal of Psychiatry* 40, 1 (Jan. 2006), 59–62. https://doi.org/10.1080/j.1440-1614.2006.01743.x

[22] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. 2003. Is seeing believing? how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, New York, NY, USA, 585–592. https://doi.org/10.1145/642611.642713

[23] Bruce P. Doré and Robert R. Morris. 2018. Linguistic Synchrony Predicts the Immediate and Lasting Impact of Text-Based Emotional Support. *Psychological Science* 29, 10 (Oct. 2018), 1716–1723. https://doi.org/10.1177/0956797618779971 Publisher: SAGE Publications Inc.

[24] Bruce P. Doré, Robert R. Morris, Daisy A. Burr, Rosalind W. Picard, and Kevin N. Ochsner. 2017. Helping Others Regulate Emotion Predicts Increased Regulation of One's Own Emotions and Decreased Symptoms of Depression. *Personality and Social Psychology Bulletin* 43, 5 (May 2017), 729–739. https://doi.org/10.1177/0146167217695558

[25] Nature Editorial. 2019. The mental health of PhD researchers demands urgent attention. *Nature* 575, 7782 (2019), 257–258. https://doi.org/10.1038/d41586-019-03489-1

[26] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. 2011. Collaborative Filtering Recommender Systems. *Foundations and Trends in Human-Computer Interaction* 4, 2 (Feb. 2011). https://doi.org/10.1561/1100000009

[27] Teresa M Evans, Lindsay Bira, Jazmin Beltran Gastelum, L Todd Weiss, and Nathan L Vanderford. 2018. Evidence for a mental health crisis in graduate education. *Nature Biotechnology* 36, 3 (March 2018), 282–284. https://doi.org/10.1038/nbt.4089

[28] Casey Fiesler. 2019. Ethical Considerations for Research Involving (Speculative) Public Data. *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (Dec. 2019), 1–13. https://doi.org/10.1145/3370271

[29] Casey Fiesler, Nathan Beard, and Brian C. Keegan. 2020. No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 187–196. https://ojs.aaai.org/index.php/ICWSM/article/view/7290

[30] Casey Fiesler and Blake Hallinan. 2018. "We Are the Product": Public Reactions to Online Data Sharing and Privacy Controversies in the Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. https://doi.org/10.1145/3173574.3173627

[31] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 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 (June 2017). https://doi.org/10.2196/mental.7785

[32] Jeana H. Frost and Michael P. Massagli. 2008. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data. *Journal of Medical Internet Research* 10, 3 (May 2008), e15. https://doi.org/10.2196/jmir.1053

[33] Nadia Garnefski and Vivian Kraaij. 2006. Relationships between cognitive emotion regulation strategies and depressive symptoms: A comparative study of five specific samples. *Personality and Individual Differences* 40, 8 (June 2006), 1659–1669. https://doi.org/10.1016/j.paid.2005.12.009

[34] R. Stuart Geiger and Aaron Halfaker. 2016. Open algorithmic systems: lessons on opening the black box from Wikipedia. *AoIR Selected Papers of Internet Research* (Oct. 2016). https://spir.aoir.org/ojs/index.php/spir/article/view/8772

[35] Andrea K. Graham, Carolyn J. Greene, Mary J. Kwasny, Susan M. Kaiser, Paul Lieponis, Thomas Powell, and David C. Mohr. 2020. Coached Mobile App Platform for the Treatment of Depression and Anxiety Among Primary Care Patients: A Randomized Clinical Trial. *JAMA psychiatry* 77, 9 (Sept. 2020), 906–914. https://doi.org/10.1001/jamapsychiatry.2020.1011

[36] J.J. Gross and Ross Thompson. 2007. Emotion Regulation: Conceptual Foundations. *Handbook of Emotion Regulation* (2007), 3–27.

[37] James J. Gross. 1998. Antecedent- and Response-Focused Emotion Regulation: Divergent Consequences for Experience, Expression, and Physiology. *Journal of personality and social psychology* 74, 1 (1998), 224–237. https://doi.org/10.1037/0022-3514.74.1.224 Publisher: American Psychological Association, American Psychological Association APA.

[38] Aaron Halfaker and R. Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *arXiv:1909.05189 [cs]* (Aug. 2020). http://arxiv.org/abs/1909.05189 arXiv: 1909.05189.

[39] Department of Health Education & Welfare. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.* Technical Report. Office of Human Research Protections. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html Last Modified: 2018-01-15 00:00:00.

[40] Christopher K. Hsee. 1996. The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives. *Organizational Behavior and Human Decision Processes* 67, 3 (Sept. 1996), 247–257. https://doi.org/10.1006/obhd.1996.0077

[41] Tristen K. Inagaki and Edward Orehek. 2017. On the Benefits of Giving Social Support: When, Why, and How Support Providers Gain by Caring for Others. *Current Directions in Psychological Science* 26, 2 (April 2017), 109–113. https://doi.org/10.1177/0963721416686212 Publisher: SAGE Publications Inc.

[42] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (July 2019), 31:1–31:35. https://doi.org/10.1145/3338243

[43] Khari Johnson. 2020. AI ethics pioneer's exit from Google involved research into risks and inequality in large language models. https://venturebeat.com/2020/12/03/ai-ethics-pioneers-exit-from-google-involved-research-into-risks-and-inequality-in-large-language-models/

[44] Alan E. Kazdin and Stacey L. Blase. 2011. Rebooting Psychotherapy Research and Practice to Reduce the Burden of Mental Illness. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 6, 1 (Jan. 2011), 21–37. https://doi.org/10.1177/1745691610393527

[45] Aniket Kittur, Bongwon Suh, and Ed Chi. 2008. Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia *(CSCW '08)*. ACM, 477–480. https://doi.org/10.1145/1460563.1460639

[46] Daniel Kluver, Michael D. Ekstrand, and Joseph A. Konstan. 2018. Rating-Based Collaborative Filtering: Algorithms and Evaluation. In *Social Information Access: Systems and Technologies*, Peter Brusilovsky and Daqing He (Eds.). Springer International Publishing, Cham, 344–390. https://doi.org/10.1007/978-3-319-90092-6_10

[47] Daniel Kluver, Tien T. Nguyen, Michael Ekstrand, Shilad Sen, and John Riedl. 2012. How many bits per rating?. In *Proceedings of the sixth ACM conference on Recommender systems (RecSys '12)*. Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/2365952.2365974

[48] K. Kroenke, R. L. Spitzer, and J. B. Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* 16, 9 (Sept. 2001), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[49] Ethan Kross and Ozlem Ayduk. 2011. Making Meaning out of Negative Experiences by Self-Distancing. *Current Directions in Psychological Science* 20, 3 (2011). https://doi.org/10.1177/0963721411408883

[50] Ranjitha Kumar, Juho Kim, and Scott R. Klemmer. 2009. Automatic retargeting of web page content. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems (CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 4237–4242. https://doi.org/10.1145/1520340.1520646

[51] Zachary Levonian, Marco Dow, Drew Erikson, Sourojit Ghosh, Hannah Miller Hillberg, Saumik Narayanan, Loren Terveen, and Svetlana Yarosh. 2020. Patterns of Patient and Caregiver Mutual Support Connections in an Online Health Community. *arXiv:2007.16172 [cs]* (Sept. 2020). http://arxiv.org/abs/2007.16172 arXiv: 2007.16172.

[52] Haley MacLeod, Grace Bastin, Leslie S. Liu, Katie Siek, and Kay Connelly. 2017. "Be Grateful You Don't Have a Real Disease": Understanding Rare Disease Relationships. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1660–1673. https://doi.org/10.1145/3025453.3025796

[53] Kateri McRae, Bethany Ciesielski, and James J. Gross. 2012. Unpacking cognitive reappraisal: Goals, tactics, and outcomes. *Emotion* 12, 2 (2012), 250–255. https://doi.org/10.1037/a0026351

[54] Kateri McRae, Kevin N. Ochsner, Iris B. Mauss, John J. D. Gabrieli, and James J. Gross. 2008. Gender Differences in Emotion Regulation: An fMRI Study of Cognitive Reappraisal. *Group Processes & Intergroup Relations* 11, 2 (April 2008), 143–162. https://doi.org/10.1177/1368430207088035 Publisher: SAGE Publications Ltd.

[55] David C. Mohr, Michelle Nicole Burns, Stephen M. Schueller, Gregory Clarke, and Michael Klinkman. 2013. Behavioral Intervention Technologies: Evidence review and recommendations for future research in mental health. *General Hospital Psychiatry* 35, 4 (July 2013), 332–338. https://doi.org/10.1016/j.genhosppsych.2013.03.008

[56] David C Mohr, Kathryn Noth Tomasino, Emily G Lattie, Hannah L Palac, Mary J Kwasny, Kenneth Weingardt, Chris J Karr, Susan M Kaiser, Rebecca C Rossom, Leland R Bardsley, Lauren Caccamo, Colleen Stiles-Shields, and Stephen M Schueller. 2017. IntelliCare: An Eclectic, Skills-Based App Suite for the Treatment of Depression and Anxiety. *Journal of Medical Internet Research* 19, 1 (Jan. 2017), e10. https://doi.org/10.2196/jmir.6645

[57] M. Mongrain, Caroline Barnes, Ryan Barnhart, and Leah B. Zalan. 2018. Acts of Kindness Reduce Depression in Individuals Low on Agreeableness. (2018). https://doi.org/10.1037/tps0000168

[58] Robert R. Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M. Schueller. 2018. Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions. *Journal of Medical Internet Research* 20, 6 (June 2018), e10148. https://doi.org/10.2196/10148 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[59] Robert R. Morris and Rosalind Picard. 2014. Crowd-powered positive psychological interventions. *The Journal of Positive Psychology* 9, 6 (Nov. 2014), 509–516. https://doi.org/10.1080/17439760.2014.913671

[60] Robert R Morris, Stephen M Schueller, and Rosalind W Picard. 2015. Efficacy of a Web-Based, Crowdsourced Peer-To-Peer Cognitive Reappraisal Platform for Depression: Randomized Controlled Trial. *Journal of Medical Internet Research* 17, 3 (March 2015). https://doi.org/10.2196/jmir.4167

[61] Robert (Robert Randall) Morris. 2015. *Crowdsourcing mental health and emotional well-being.* Thesis. Massachusetts Institute of Technology. http://dspace.mit.edu/handle/1721.1/97972

[62] Casey Newton. 2020. Half of all Facebook moderators may develop mental health issues. https://www.theverge.com/interface/2020/5/13/21255994/facebook-content-moderator-lawsuit-settlement-mental-health-issues

[63] Tien T. Nguyen, Daniel Kluver, Ting-Yu Wang, Pik-Mai Hui, Michael D. Ekstrand, Martijn C. Willemsen, and John Riedl. 2013. Rating support interfaces to improve user experience and recommender accuracy. In *Proceedings of the 7th ACM conference on Recommender systems (RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 149–156. https://doi.org/10.1145/2507157.2507188

[64] Alicia L. Nobles, Jeffrey J. Glenn, Kamran Kowsari, Bethany A. Teachman, and Laura E. Barnes. 2018. Identification of Imminent Suicide Risk Among Young Adults Using Text Messages. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 413:1–413:11. https://doi.org/10.1145/3173574.3173987

[65] Matthew K. Nock, Evan M. Kleiman, Melissa Abraham, Kate H. Bentley, David A. Brent, Ralph J. Buonopane, Franckie Castro-Ramirez, Christine B. Cha, Walter Dempsey, John Draper, Catherine R. Glenn, Jill Harkavy-Friedman, Michael R. Hollander, Jeffrey C. Huffman, Hye In S. Lee, Alexander J. Millner, David Mou, Jukka-Pekka Onnela, Rosalind W. Picard, Heather M. Quay, Osiris Rankin, Shannon Sewards, John Torous, Joan Wheelis, Ursula Whiteside, Galia Siegel, Anna E. Ordóñez, and Jane L. Pearson. 2020. Consensus Statement on Ethical & Safety Practices for Conducting Digital Monitoring Studies with People at Risk of Suicide and Related Behaviors. *Psychiatric Research and Clinical Practice* (Dec. 2020), n/a–n/a. https://doi.org/10.1176/appi.prcp.20200029 Publisher: American Psychiatric Publishing.

[66] Yuki Nozaki and Moïra Mikolajczak. 2020. Extrinsic Emotion Regulation. *Emotion* 20, 1 (2020), 6. https://doi.org/10.1037/emo0000636

[67] Kathleen O'Leary, Arpita Bhattacharya, Sean A. Munson, Jacob O. Wobbrock, and Wanda Pratt. 2017. Design Opportunities for Mental Health Peer Support Technologies. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1470–1484. https://doi.org/10.1145/2998181.2998349

[68] Kathleen O'Leary, Stephen M. Schueller, Jacob O. Wobbrock, and Wanda Pratt. 2018. "Suddenly, We Got to Become Therapists for Each Other": Designing Peer Support Chats for Mental Health. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 331:1–331:14. https://doi.org/10.1145/3173574.3173905

[69] Judith S. Olson and Wendy Kellogg (Eds.). 2014. *Ways of knowing in HCI.* Springer, New York. OCLC: ocn879418401.

[70] A. Osman, C. L. Bagge, P. M. Gutierrez, L. C. Konick, B. A. Kopper, and F. X. Barrios. 2001. The Suicidal Behaviors Questionnaire-Revised (SBQ-R): validation with clinical and nonclinical samples. *Assessment* 8, 4 (Dec. 2001), 443–454. https://doi.org/10.1177/107319110100800409

[71] Pablo Paredes, Ran Giald-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. 2014. PopTherapy: Coping with Stress through Pop-Culture. https://eudl.eu/doi/10.4108/icst.pervasivehealth.2014.255070

[72] Kevin Patrick, Eric B. Hekler, Deborah Estrin, David C. Mohr, Heleen Riper, David Crane, Job Godino, and William T. Riley. 2016. The Pace of Technologic Change: Implications for Digital Health Behavior Intervention Research. *American Journal of Preventive Medicine* 51, 5 (Nov. 2016), 816–824. https://doi.org/10.1016/j.amepre.2016.05.001

[73] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15.

https://doi.org/10.1145/3313831.3376695

[74] Julie Prescott, Amy Leigh Rathbone, and Terry Hanley. 2020. Online mental health communities, self-efficacy and transition to further support. *Mental Health Review Journal* 25, 4 (Jan. 2020), 329–344. https://doi.org/10.1108/MHRJ-12-2019-0048 Publisher: Emerald Publishing Limited.

[75] Candace M. Raio, Temidayo A. Orederu, Laura Palazzolo, Ashley A. Shurick, and Elizabeth A. Phelps. 2013. Cognitive emotion regulation fails the stress test. *Proceedings of the National Academy of Sciences* 110, 37 (Sept. 2013), 15139–15144. https://doi.org/10.1073/pnas.1305706110 ISBN: 9781305706118 Publisher: National Academy of Sciences Section: Biological Sciences.

[76] Derek Richards. 2009. Features and benefits of online counselling: Trinity College online mental health community. *British Journal of Guidance & Counselling* 37, 3 (Aug. 2009), 231–242. https://doi.org/10.1080/03069880902956975 Publisher: Routledge _eprint: https://doi.org/10.1080/03069880902956975.

[77] Derek Richards and Ladislav Timulak. 2012. Client-identified helpful and hindering events in therapist-delivered vs. self-administered online cognitive-behavioural treatments for depression in college students. *Counselling Psychology Quarterly* 25, 3 (Sept. 2012), 251–262. https://doi.org/10.1080/09515070.2012.703129 Publisher: Routledge _eprint: https://doi.org/10.1080/09515070.2012.703129.

[78] William N. Robiner. 2006. The mental health professions: Workforce supply and demand, issues, and challenges. *Clinical Psychology Review* 26, 5 (Sept. 2006), 600–625. https://doi.org/10.1016/j.cpr.2006.05.002

[79] Adam Rosenstein, Aishma Raghu, and Leo Porter. 2020. Identifying the Prevalence of the Impostor Phenomenon Among Computer Science Students. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20)*. Association for Computing Machinery, New York, NY, USA, 30–36. https://doi.org/10.1145/3328778.3366815

[80] Barbara Olasov Rothbaum, Elizabeth A. Meadows, Patricia Resick, and David W. Foy. 2000. Cognitive-behavioral therapy. In *Effective treatments for PTSD: Practice guidelines from the International Society for Traumatic Stress Studies*. Guilford Press, New York, NY, US, 320–325.

[81] Sabirat Rubya and Svetlana Yarosh. 2017. Video-Mediated Peer Support in an Online Community for Recovery from Substance Use Disorders. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1454–1469. https://doi.org/10.1145/2998181.2998246

[82] Sabirat Rubya and Svetlana Yarosh. 2021. Comparing Generic and Community-Situated Crowdsourcing for Data Validation in the Context of Recovery from Substance Use Disorders. In *CHI*.

[83] Esteban A. Ríssola, David E. Losada, and Fabio Crestani. 2021. A Survey of Computational Methods for Online Mental State Assessment on Social Media. *ACM Transactions on Computing for Healthcare* 2, 2 (March 2021), 17:1–17:31. https://doi.org/10.1145/3437259

[84] Benjamin Scheibehenne, Rainer Greifeneder, and Peter M. Todd. 2010. Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload. *Journal of Consumer Research* 37, 3 (Oct. 2010), 409–425. https://doi.org/10.1086/651235

[85] Jessica Schroeder, Chelsey Wilkes, Kael Rowan, Arturo Toledo, Ann Paradiso, Mary Czerwinski, Gloria Mark, and Marsha M. Linehan. 2018. Pocket Skills: A Conversational Mobile Web App To Support Dialectical Behavioral Therapy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 398:1–398:15. https://doi.org/10.1145/3173574.3173972

[86] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. *arXiv:2101.07714 [cs]* (Jan. 2021). http://arxiv.org/abs/2101.07714 arXiv: 2101.07714.

[87] Gal Sheppes and Nachshon Meiran. 2007. Better Late Than Never? On the Dynamics of Online Regulation of Sadness Using Distraction and Cognitive Reappraisal. *Personality and Social Psychology Bulletin* 33, 11 (Nov. 2007), 1518–1532. https://doi.org/10.1177/0146167207305537

[88] C. Estelle Smith, Avleen Kaur, Katie Z. Gach, Loren Terveen, Mary Jo Kreitzer, and Susan O'Conner-Von. 2021. What is Spiritual Support and How Might It Impact the Design of Online Communities?. In *Proceedings of the 2021 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '21)*. Association for Computing Machinery, Virtual.

[89] C. Estelle Smith, Zachary Levonian, Haiwei Ma, Robert Giaquinto, Gemma Lein-Mcdonough, Zixuan Li, Susan O'conner-Von, and Svetlana Yarosh. 2020. "I Cannot Do All of This Alone": Exploring Instrumental and Prayer Support in Online Health Communities. *ACM Transactions on Computer-Human Interaction* 27, 5 (Aug. 2020), 38:1–38:41. https://doi.org/10.1145/3402855

[90] C. Estelle Smith, Eduardo Nevarez, and Haiyi Zhu. 2020. Disseminating Research News in HCI: Perceived Hazards, How-To's, and Opportunities for Innovation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.

3376744

[91] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376783

[92] Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine* 166, 10 (May 2006), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

[93] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems. *ACM Transactions on Computer-Human Interaction* 27, 5 (Aug. 2020), 34:1–34:53. https://doi.org/10.1145/3398069

[94] John Torous, Jennifer Nicholas, Mark E. Larsen, Joseph Firth, and Helen Christensen. 2018. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence-Based Mental Health* 21, 3 (Aug. 2018), 116–119. https://doi.org/10.1136/eb-2018-102891 Publisher: Royal College of Psychiatrists Section: Clinical review.

[95] Allison S. Troy, Amanda J. Shallcross, and Iris B. Mauss. 2013. A Person-by-Situation Approach to Emotion Regulation: Cognitive Reappraisal Can Either Help or Hurt, Depending on the Context. *Psychological Science* 24, 12 (Dec. 2013), 2505–2514. https://doi.org/10.1177/0956797613496434

[96] Allison S. Troy, Frank H. Wilhelm, Amanda J. Shallcross, and Iris B. Mauss. 2010. Seeing the silver lining: Cognitive reappraisal ability moderates the relationship between stress and depressive symptoms. *Emotion* 10, 6 (2010), 783–795. https://doi.org/10.1037/a0020262

[97] Jim Warren, Sarah Hopkins, Andy Leung, Sarah Hetrick, and Sally Merry. 2020. Building a Digital Platform for Behavioral Intervention Technology Research and Deployment. https://doi.org/10.24251/HICSS.2020.414 Accepted: 2020-01-04T07:51:24Z.

[98] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. (May 2017). https://arxiv.org/abs/1705.09899v2

[99] Maria K. Wolters, Zawadhafsa Mkulo, and Petra M. Boynton. 2017. The Emotional Work of Doing eHealth Research. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 816–826. https://doi.org/10.1145/3027063.3052764 event-place: Denver, Colorado, USA.

[100] Jamil Zaki and W. Craig Williams. 2013. Interpersonal emotion regulation. *Emotion* 13, 5 (2013), 803–810. https://doi.org/10.1037/a0033839 Place: US Publisher: American Psychological Association.

[101] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 194:1–194:23. https://doi.org/10.1145/3274463

[102] Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology* 12, 4 (Dec. 2010), 313–325. https://doi.org/10.1007/s10676-010-9227-5 Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 4 Publisher: Springer Netherlands.

A INTERVIEW PROTOCOL

Semi-structured interviews were conducted in person or over telephone after the completion of the one-month study deployment period. Participants had already been informed of the nature and purpose of the research at intake, thus we began interviews simply by requesting consent to record. Next, we used the following questions as the basis for all interviews, while also asking additional follow-up questions on an individual basis:

- Please describe your experience using Flip*Doubt over the past month.
- When did you find yourself most likely to use the app? Why did you use it at that time?
- To what degree did you naturally have negative thoughts occurring that you wanted to input vs. having to "come up with one"?
- What did you like about using Flip*Doubt?
- What did you dislike about using Flip*Doubt?
- Have you noticed any differences in your mood or wellbeing that you would attribute to using Flip*Doubt?
- How would you generally describe why you rated things positively or negatively?
- Do you have any particular favorite reframes, or stories about using the app that really stand out? What made them memorable?
- How do you feel about the thoughts being reframed by random strangers vs. people you know?
- How would you feel about being on the other side, i.e. reframing other people's thoughts?
- If you could add any features at all to Flip*Doubt, what would they be and why?
- Is there anything else you would like to share about your experience that we haven't talked about yet?

B INSTRUCTIONS PRESENTED TO AMAZON MECHANICAL TURK CROWDWORKERS

B.1 Verbatim HIT Instructions

HIT Title: "Re-write a negative thought with a positive spin."
HIT Instructions: The following text is a negative thought that a person typed about themself:
[Participant's Input]

Rewrite this thought in a way that is more positive. For instance, try to think up a way of saying it that might go on a motivational poster. Try to be as inspirational as possible so that you can encourage the person to feel better about themself!

[Free Response Text Field to Write Reframe]

Here is an example negative thought: *"I am so terrible at speaking in public...everyone is going to think I'm an idiot."*
Acceptable responses could be:

- "I present great ideas. That's what really matters."
- "I am going to share my brilliance with the world, and who cares what they think!"

Note: You cannot simply negate what was written, or make the thought more negative. Unacceptable responses include not changing the text at all, submitting nothing, or something like:

- "I am not terrible at speaking in public, and everyone is not going to think I'm an idiot."

- "I suck so bad at speaking in public. What a complete loser!"

B.2 Potential Learning Bias

In the above examples, if the example reframes were to be coded according to our coding protocol, codes applied would be as presented in blue below.

- "I present great ideas. That's what really matters." `Change Current Consequences`
- "I am going to share my brilliance with the world, and who cares what they think!" `Agency`, `Distancing`
- "I am not terrible at speaking in public, and everyone is not going to think I'm an idiot." `Direct Negation`
- "I suck so bad at speaking in public. What a complete loser!" `Misunderstand Instructions`

`Change Current Consequences` and `Agency` were the top two most commonly applied tactics codes. This is potentially due to a learning bias resulting from exposure to the above examples. Future work could determine if this is true by repeating data collection with different examples that use different tactics, and seeing if the frequency of codes shifts towards the tactics presented in examples. The fact that `Distancing` was rarely applied is also worth noting. (In this case, `Distancing` is a *second* tactic phrase, possibly making it less influential.)

C DIFFERENTIAL NEGATIVE THOUGHT PATTERNS

We completed a supplementary analysis in order to understand: *Do different users experience different patterns of negative thoughts (thus leading to inputs with differentially applied codes)?*

	code	p.value	p.adjust	sig
1	Comparison with Others	5.364949e-01	8.366186e-01	FALSE
2	Regret	2.002898e-02	1.201739e-01	FALSE
3	Uncertainty or Worry	7.066742e-06	9.893439e-05	TRUE
4	Generic	6.206414e-04	6.827056e-03	TRUE
5	Ruminating on Others' Thoughts	4.962160e-03	4.325123e-02	TRUE
6	Self-Disparagement	5.941491e-07	8.912236e-06	TRUE
7	Future	2.081923e-02	1.201739e-01	FALSE
8	Past	3.266205e-02	1.306482e-01	FALSE
9	Present	5.799983e-03	4.325123e-02	TRUE
10	Generic Emotions	7.922088e-04	7.922088e-03	TRUE
11	Financial	1.215888e-01	3.647664e-01	FALSE
12	Health or Appearance	1.656100e-04	2.152930e-03	TRUE
13	Managing Home Life	4.183093e-01	8.366186e-01	FALSE
14	Personal Relationships	4.719880e-04	5.663856e-03	TRUE
15	Work/School	4.805692e-03	4.325123e-02	TRUE

Table 3. Results of holm-bonferroni corrected chi-sq. test.

We computed the number of input codes applied to thoughts each user had, and the number of times each user had that code applied. For each of 15 codes, we treated this like a binomial process: either a user's thought does, or does not, have that code. We used a 13-sample chi-sq. test to test the null hypothesis that, for a given code, all users generated thoughts with this code at the same rate. A small p-value indicates that at least two users generate thoughts with different codes. Finally, we applied a holm-bonferroni p-value correction to account for 14 non-independent p-value tests. Table 3 shows results demonstrating that for 9 out of 15 codes (which are all more commonly applied codes), there exist meaningful differences in use-rate between at least two users.

The favorable alignment of significant results against well-used codes indicates that this is likely to be a result that is generally true, which is being limited by code use patterns and data sample size.

D DETAILED CODEBOOK INFORMATION

D.1 Adaptation from Prior Codebook by McRae et al.

The Reformulation Tactics codebook from [53] included the following bulleted reformulation tactics, which in some cases were modified from the original code title (left) to better match the data and support consistent application of codes. Most of the codes maintained the same basic meaning of McRae's codes, however our code definitions include a few more specific textual indicators/instructions that help to distinguish between cases (e.g. see our defintions of "Agency" and "Technical-Analytic Problem Solving." If a code was substantially altered, we provide an explanation to the right of the arrows below:

- **Explicitly Positive** → Our research team changed the title of this code to **"Silver Lining"** because we felt that it helped to clarify when the code was applicable (e.g. the code's definition that something good/positive occurred specifically because of something bad/negative). All reframes are *intended* to be positive, therefore the phrase "explicitly positive" felt confusing and too easy to over-apply. As in McRae, we still used this code as an "orthogonal" code which must be applied in addition to a main tactic code.
- **Change Current Circumstances**
- **Reality Challenge** → Our research team migrated this code out of the Reformulation Tactics codebook and into the Meta-Behaviors codebook because it was very difficult to reliably distinguish it from "Change Current Circumstances." Additionally, reframe statements that use other main reformulation tactics can also challenge the nature of a situation, suggesting that "Reality Challenge" is better suited to an additional Meta-Behavior than a main tactic.
- **Change Future Consequences**
- **Agency**
- **Distancing** → We retained this code from McRae, however we narrowed its scope to only include statements/words that the user shouldn't care about the input issue, or that the input issue doesn't matter. Broader conceptions of the code as any form of psychological distancing were again difficult to reliably differentiate from Change Current Circumstances.
- **Technical-Analytic Problem Solving**
- **Acceptance**

Our reformulation tactics codebook *adds* the new codes "Direct Negation" and "Misunderstand Instructions," which were not present in McRae, while our Meta-Behaviors codebook is new compared to McRae (except for the Reality Challenge code, which we migrated to the Meta-Behaviors codebook instead of Reformulation tactics). We note that "Change Current Circumstances" is the broadest category in this codebook and was also the most common tactic used in this study; future work could potentially provide finer-grained subdivision of this code to assess even more specific strategies for changing current circumstances.

For future researchers interested in using our codebooks, we are happy to discuss the nuances of our coding application, or to share more examples from our data. Please direct inquiries in this topic to the lead author of this study.

D.2 Complete Codebooks

Codes for Inputs		Definition
Temporal α = 0.94	**Past**	May involve present or past tense verbs, but uses language suggesting that the subject of the thought is from the past. Thoughts that express past issue(s) that are causing distress in the present should be labeled past.
	Present	These thoughts focus on the current state of the person.
	Future	Mentions the future or uses language that suggests the subject of the thought is something about the future.
Topic α = 0.83	**Personal Relationships**	Involves personal relationship(s) with one or more other people, or even pets (not including pet care, which is "logistics of life"). If bosses, advisors or coworkers are mentioned, then the thought should be "work or school."
	Work or School	Involves feelings about, conflicts related to, or performance/progress at work or school.
	Health and Appearance	Involves health, appearance, fitness level, eating or sleep habits, of other things attributable to health.
	Financial	Involves money, finances, ability to pay for future expenses, indebtedness, and complaints about cost.
	Managing Home Life	Involves logistics of life such as chores, errands, raising children, or a negative status at home. Unless the thought specifically mentions work, this category also includes general statements of overwhelm, time management issues, or having too many responsibilities.
	Generic Emotions	Involves emotions or feelings that do not specifically relate to a concrete topic.
Meta-Topic α = 0.84	**Self-Disparagement**	Statements that reflect internal self-doubt, self-critique, inability, insecurity, lack of self worth, or other negative statements that resemble put-downs, insults, or insufficiencies. It also includes statements that imply a person needs to or should do better or more than what they're presently doing.
	Expressing Regret	Expressions of anguish or regret about past events, relationships or decisions. There must be explicit signals of regret, such as *"should have done X differently"* or *"I regret…"* or *"I wish I had not Y."*
	Comparison with Others	Statements of comparison to friends, family, peers, or mentors.
	Expressing Uncertainty/Worry	Statements that express uncertainty or worry about how things are going now, or will turn out in the future. May be expressed in the form of a question, or a rhetorical question.
	Ruminating on Others' Thoughts and Motivations	Statements expressing fear of what other people will think, including judgement, assumptions that other people have negative opinions, or worry about people's motivations.
	Generic Complaint	Complaints or statements about hard facts of life, reality, situations, events, people, etc. Includes any thought that does not fit neatly into any other category without additional inference.

Table 4. Full Codebook for Inputs. α is Krippendorf's alpha, calculated for three coders.

Received January 2021; revised April 2021; revised July 2021; accepted July 2021

Codes for Reframes		α	Definition
Reformulation Tactics	**Silver Lining**	0.78	The person or situation is better off, or something positive "is able to or allowed to occur" specifically because of some aspect of the negative input stressor. **Rule:** As in McRae et al., this code is designated to be "orthogonal" and must be applied in addition to a main tactic code.
	Change Current Circumstances	0.64	Changes the interpretation or even the nature of the current circumstance, possibly by adding new information to the input statement that was not present in the original thought. Must be present or past tense.
	Change Future Consequences	0.85	Specifies that in the future, the consequences will be different, better, or improved rather than what one might first assume. May include language like, "The situation will improve with time, it'll get better soon." Must be future tense.
	Agency	0.75	Invokes that a person has the skill, or will work with someone who has the skills, to change the current situation, or re-states the input with an assertion that the person will actively do something about the issue described in the input.
	Distancing	N/A*	Invokes psychological distance through stating that the person shouldn't care about the negative stressor, or else that it doesn't matter.
	Technical-Analytic Problem Solving	0.72	Focuses on specific steps that were not explicitly mentioned in the input that can be taken to solve or improve the situation. **Rule:** If a statement uses agency-like language (e.g. I can, or I will), followed by a new problem-solving step or action not mentioned in input, the code Technical-Analytic Problem Solving should be applied, not Agency.
	Acceptance	0.65	Normalizes the negative event invoking the justification that sometimes bad things happen. Could add or shift perspective that encourages acceptance of the circumstance without having to change the circumstance itself. **Rule:** If the reframe acknowledges the problem, but then uses a different main tactic, the code "Acknowledge Main Concern" should be applied, not "Acceptance."
	Direct Negation	1	Simply stating the opposite of the input, without further insight or interpretation. **Rules:** If a statement is both a direct negation and another tactic, Direct Negation trumps the other tactic. If Direct Negation is applied, do not apply Reality Challenge, Introduce New Personal Context, or Ignore Input Issue(s).
	Misunderstand Instructions	0.81	This includes copy/pasting or paraphrasing (without reframing) either: (1) the input itself, or (2) an example from the instructions that is unrelated to the input.
Meta-Behaviors	**Reality Challenge**	0.6	Undermines or challenges the nature of the situation that is implied by the input thought, or suggests that the input thought might be inaccurate in some way, e.g. by making assumptions that might be untrue.
	Introducing New Personal Context	0.6	Reframe includes new information (never mentioned in the input) that may or may not be true about the specific context/personal qualities/thoughts/feelings of the participant. **Rule:** This excludes statements that assume *ability.*
	Acknowledge Main Concern	0.85	The reframe contains a phrase (separate from the main clause) that repeats a concern communicated by the input, as a means of validating the concern. **Rule:** If an input has the format, "Because of X, Y", and the reframe, "Because of X, Z", where "Because of X" is identical or synonymous (i.e. no additional perspective added), this does not get the "Acknowledge Main Concern" code.
	Ignore Input Issue(s)	0.76	The reframe does not relate to any issues in the input (e.g. generic motivation). **Rule:** This code cannot be applied if "Acknowledge Main Concern" is applied.
	Minor Grammar	0.64	One or two misspellings, mispunctuations, incorrect capitalizations, missed articles, or minor erroneous phrasing. **Rules:** If a grammar issue in the input is replicated in the reframe, do not apply this code. There must be a definite error, i.e. *awkward* phrasing doesn't get this code.
	Major Grammar	N/A*	ALL CAPS, severely erroneous or nonsensical phrasing.

Table 5. Full Codebook and Coding Rules for Reframes. α is Krippendorf's alpha, calculated for three coders. The language of Reformulation Tactics definitions is directly adapted from [53]; see Section D.1 above for a description of how codes were adapted.