CASS: Towards Building a Social-Support Chatbot for Online Health Community

LIUPING WANG*, Institute of Software, Chinese Academy of Sciences and University of Chinese Academy of Sciences, China

DAKUO WANG*, IBM Research, USA

FENG TIAN, Institute of Software, Chinese Academy of Sciences, China

ZHENHUI PENG, The Hong Kong University of Science and Technology, Hong Kong

XIANGMIN FAN, Institute of Software, Chinese Academy of Sciences, China

ZHAN ZHANG, Pace University, USA

SHUAI MA, The Hong Kong University of Science and Technology, Hong Kong

MO YU, IBM Research, USA

XIAOJUAN MA, The Hong Kong University of Science and Technology, Hong Kong

HONGAN WANG, Institute of Software, Chinese Academy of Sciences, China

Chatbots systems, despite their popularity in today's HCI and CSCW research, fall short for one of the two reasons: 1) many of the systems use a rule-based dialog flow, thus they can only respond to a limited number of pre-defined inputs with pre-scripted responses; or 2) they are designed with a focus on single-user scenarios, thus it is unclear how these systems may affect other users or the community. In this paper, we develop a generalizable chatbot architecture (CASS) to provide social support for community members in an online health community. The CASS architecture is based on advanced neural network algorithms, thus it can handle new inputs from users and generate a variety of responses to them. CASS is also generalizable as it can be easily migrate to other online communities. With a follow-up field experiment, CASS is proven useful in supporting individual members who seek emotional support. Our work also contributes to fill the research gap on how a chatbot may influence the whole community's engagement.

CCS Concepts: • **Human-centered computing** → **Computer supported cooperative work**.

Additional Key Words and Phrases: chatbot; bot; pregnancy; healthcare; AI deployment; online community; social support; peer support; emotional support; machine learning; neural network; system building; conversational agent; human AI collaboration; human AI interaction; explainable AI; trustworthy AI

---

*Both authors contributed equally to this research.

Authors' addresses: Liuping Wang, liuping2019@iscas.ac.cn, Institute of Software, Chinese Academy of Sciences , University of Chinese Academy of Sciences, China; Dakuo Wang, dakuo.wang@ibm.com, IBM Research, USA; Feng Tian, tianfeng@iscas.ac.cn, Institute of Software, Chinese Academy of Sciences, China; Zhenhui Peng, The Hong Kong University of Science and Technology, Hong Kong; Xiangmin Fan, Institute of Software, Chinese Academy of Sciences, China; Zhan Zhang, Pace University, USA; Shuai Ma, The Hong Kong University of Science and Technology, Hong Kong; Mo Yu, IBM Research, USA; Xiaojuan Ma, The Hong Kong University of Science and Technology, Hong Kong; Hongan Wang, Institute of Software, Chinese Academy of Sciences, China.

---

1 INTRODUCTION

Chatbot systems[1] have been increasingly adopted in many fields (e.g., healthcare [30], human
resources (HR) [64], and customer service [122]), since the first chatbot system—ELIZA—emerged
in 1964 to provide consulting sessions as a computer therapist [115]. In recent years, an increasing
number of chatbot systems are being developed in various research labs and companies with a
premise that these systems can have more powerful capabilities and support more user scenarios [21,
44, 103]. For example, Hu et al. [44] built an experimental chatbot system that can understand the
tones in a text input (e.g., sad or polite) and generate responses with an appropriate tone.

Following these system development efforts, many recent Human-Computer Interaction (HCI)
and Computer-Supported Cooperative Work (CSCW) studies have examined various aspects of
chatbots from the end users' perspective, such as human-in-the-loop chatbot design [21], user
perception of chatbots [18], playful usage of chatbots [64], and human trust in chatbots [49].
However, most of these studies have inherent limitations: 1) many chatbot systems (e.g., [64, 123])
use a rule-based architecture, which makes the chatbot capable of understanding only a limited
number of user inputs and responding with prescripted sentences, hindering its generalization;
and 2) most chatbots are deployed and tested only in single-user scenarios, but how these systems
interacting with and impacting a group of users (or a community) is understudied.

The first limitation of current chatbots – only returning a pre-defined list of responses to a
user – is partially caused by the use of traditional heuristic rule-based algorithms or information
retrieval techniques [75]. Even with some advanced chatbot-development toolkit's help (e.g.,
Microsoft Cognitive Service [74] and IBM Watson [47]), a chatbot can only use neural-network
based approaches (NN) to understand the text, but its responding function is still limited to a rule-
based selection process. In this work, we propose a NN-based chatbot architecture based on which
a chatbot system can accurately handle unseen questions and generate various forms of responses
with the same meaning. This architecture is designed to have a high scalability and generalizability,
so that other researchers and developers can take our code[2], provide it with a cleaned and labeled
training dataset, retrain it, and deploy it for different online communities. Inspired by previous
literature [117], we also build a human-in-the-loop module so a human operator can monitor and
intervene the fully automated architecture, if needed.

The second research gap is that most of today's HCI and CSCW research primarily focus on an
individual user's interaction with and perception of a chatbot system [31, 46, 58, 64, 94, 122], and
it is unclear how a chatbot system may affect a group of users or a community. In this paper, we
aim to address this research gap by building and deploying a social-support chatbot system, **CASS**
(**Ch**Atbot for **S**ocial **S**upport), and evaluating its impact on the individuals who need social-support
as well as on the other members in the community. Motivated by existing literature that online
communities often suffer low engagement from the members due to that many posts can not get a
timely response, our chatbot's primary function is designed to engage in conversations with those
un-replied social-support-needing posts. CASS automates the entire end-to-end process: retrieving

---

[1]In this paper, we consider functional bots (e.g., Twitter bots [34], and Wikipedia bots [50]) out of our research scope, as
these bots' primary function focuses on completing tasks (e.g., broadcasting an information or editing an article), as opposed
to communicating with users through a flow of conversations, which is the primary feature of chabots.
[2]We have made the code repository open source: https://github.com/liupingw/CASS-Framework

new posts from the community, classifying these posts as with (or without) social-support needs, and generating appropriate responses to the posts with social-support needs.

This study is exploratory in nature. We choose an online pregnancy healthcare community as our research site (*YouBaoBao*[3]), because this type of online community is extensively studied in recent research [4, 37], allowing us to leverage these existing knowledge to inform the system design and implementation. The whole research project consists of three parts, and we will organize this paper following the order of these three studies:

Study 1 is an empirical exploration study, where we conduct both qualitative content analysis and descriptive statistical analysis to understand the context of the research site. A qualitative open coding of the posts suggests that there are three primary categories of posts in YouBaoBao: *Emotional-Support Seeking, Informational-Support Seeking, and Sharing Daily Life.* The quantitative statistical analysis also shows that half of the posts in this community can not get a timely response, which may cause the already-stressful community members (i.e., pregnant women) to be at a higher risk [37].

To provide social supports to this community, in Study 2, we build the CASS chatbot system, using the proposed generalizable NN-based architecture. The CASS design is tailored based on the findings from Study 1, so that it can understand what posts need social support, and what posts count as an un-replied case that needs immediate intervention. In Study 2, we also adopt the standard evaluation practices (using both automated metrics and human evaluators) to show that the core NN-based algorithms' performance is satisfactory.

At last, in Study 3, CASS is deployed back to the YouBaoBao community, with a human-in-the-loop module to filter inappropriate AI-generated responses. Through a 7-day field experiment, the result shows that the CASS system can provide the desired emotional support to the individual community members who need help. In addition, we find evidence to support that the deployment of CASS indeed have positive impacts on other members and the entire community, as other community members are more likely to participate in the conversations intervened by the chatbot system. This finding suggests that such social-support chatbot can not only support individual member's well-being, but also improve the community engagement level, which is also an important dimension according to McGrath's TIP theory of groups [73].

In summary, this paper makes the following contributions:

- An empirical understanding of the challenges and user needs in an online pregnancy healthcare community, and how these findings can be used to tailor a chatbot system design;
- A scalable and generalizable chatbot development architecture, with which researchers and developers can easily build a fully automated chatbot system with NN-based models to be deployed in another online community;
- Insights and recommendations for designing and deploying future chatbots systems to interact or collaborate with humans in the context of online communities.

2 RELATED WORK

The literature review is divided into three subsections: we first review selected HCI work on social support scenarios in online health communities. Then, we focused on the group of literature about human and chatbot interaction. Lastly, we switch to the literature that specifically addresses challenges and issues of chatbot systems deployment in real world.

---

[3]https://youbaobao.meiyou.com/

2.1 Online Health Community and Emotional Social Support

Online community has been a longstanding research topic for HCI and CSCW researchers (e.g., [10, 20, 27, 37, 72, 99, 104, 105, 114, 118, 125, 127, 134]). Existing studies have looked at a variety of topics including community structure [54, 88, 95], community activities [7, 37, 45, 54, 65, 71, 116], members' commitment and contribution [87, 124], engaging newcomers [53, 55, 86], rewarding mechanism design [40, 54, 88], and the cold start problem [88]. Recently, a number of studies has focused on a special type of online community – the online health community for pregnant women [23, 24, 28, 37, 42]. This is a special group of users. In addition to the significant changes on their bodies, their mental state also changes a lot over the trajectory of their pregnancy. They often have a much higher stress level than before getting pregnant, thus their mental health is at high stake [12, 48, 121, 128]. Banti et al. [11] reported that 12.4% of pregnant women had presented some depression symptoms during the pregnancy, and 9.6% of them encountered depression in the postpartum period.

It is known that pregnant women often go to online health communities to seek social support from peers [14, 29, 57]. Previous literature reveals that members in such health communities actively seek social support from others, and they are also willing to volunteer their time to provide social support to other help seekers [25, 126]. Prior research roughly divided the social supports into two categories: *informational social support and emotional social support* [15, 22, 114]. Informational support refers to posters seeking information or knowledge about the course of their disease, treatments, side effects, communication with physicians, and other burdens (e.g., financial problems) [114]. Emotional support refers to posters seeking encouragement and empathy when experienced an emotional disturbance [37]. In this project, as an illustration, our chatbot system focuses on providing non-informational social support to community members.

It is often difficult to motivate community members to actively reply to others members' posts in a timely manner [3, 61, 70, 77, 107]. Seminal research has explored various ways to solve this problem [84]. In more traditional online communities (e.g., Wikipedia), researchers have attempted to stimulate member's intrinsic and extrinsic motivations [3] with monetary reward [61] or virtual badges and reputation rewards [70]. In the online health communities, when a user posts a support seeking post, he/she is recommended to use simpler language and express the needs more explicitly [8]. It is also suggested that posts with more detailed user profile information and a photo are more likely to get replies [9]. Even so, there are many posts may never get a reply. For example, Wang et al. [113] reported that at least 10% posts in an online community never received a response.

When a pregnant woman posts a support seeking post and never gets a response, it may have more severe harms to the user and to the community. Because pregnancy women are already stressful, overlooking their support seeking may make things worse [62, 79]. Furthermore, when community members constantly fail to get the needed social support, they are less likely to contribute to the community, so the community engagement level decreases accordingly, and even worse, the members may leave the community over time [55, 114, 132].

In this paper, we will illustrate how to leverage on the latest AI technology to build a chatbot system that can automatically detect non-informational support-seeking posts, and respond to it with appropriate sentences. The system architecture is scalable and generalizable so it can be easily migrated to other online communities.

2.2 Human Interaction with Chatbots

Chatbot is an increasingly popular research topic in recent years. Most of today's chatbot systems are built to interact with a single user [31, 46, 58, 64, 94, 122]. For example, the famous ELIZA

and its successors can provide individual cognitive therapy sessions to users with a purpose of relieving their stress and anxiety, as well as helping them gain self-compassion [31, 58, 94]. Many commercial chatbots in customer service domain are designed to answer customers' frequently asked questions or perform a simple function (e.g., check bank account balance) [46, 122]. There are also human resources (HR) chatbots serving as a process guider to lead new employees to go through their onboarding process [64]. There are also some chatbot applications in the healthcare domain which can help patients to better understand their symptoms [30].

Only till recently, a number of researchers have started to explore and build chatbots that can interact with a group of people [2, 21, 66, 96, 98, 103, 130]. For example, Toxtli et al. [103] built a chatbot as a group chat facilitator to assign tasks to group members. Zhang et al. [130] developed a chatbot for an online communication application to automatically summarize group chat messages. Cranshaw et al. [21] developed a chatbot system that can serve as an assistant to coordinate meetings for multiple people via email. All these studies expand the chatbot user scenario from supporting single user to multiple users. Notably, Seering et al. [96] recently built a chatbot in an online gaming community on Twitch. They designed four different versions of the chatbot – "baby", "toddler", "adolescent", and "tennager" – to simulate a chatbot's grownup process in a 3-week deployment. However, the user interaction with the chatbot is quite primitive that users need to input command-line text (e.g., "@Babybot" or "!feed"), thus it is more like a digital pet (tamagotchi) [85] than a chatbot.

Along this research line, our work builds a chatbot system that can provide emotional support to pregenant women members in an online community; also, we design a field experiment to reveal findings on how such deployment of the chatbot can impact the whole community. Different from Seering's work [96], where they deployed the "digital pet" into a Twitch community and users can have chitchat with the system, we aim to build a chatbot to meet community members' existing needs — social-support seeking. We hope our chatbot can provide functional benefits to the users through a conversational communication. Another difference is that Seering's system [96] used an information retrieval (IR) approach, and in turn, it could only return limited responses that were pre-defined by the researchers. The rule-based approach (e.g., IR) constraints the potential of chatbot in providing social support for a community [63], and users may perceive the responses from the chatbot not as useful as from people [75]. In contrast, we build an architecture that leverages on the state-of-the-art NN-based models for the development of more powerful chabot systems.

2.3 Impacts of Chatbot System Deployment

While reviewing recent work on building and deploying chatbots [32, 38, 93], we found that despite many of these chatbots were designed with a good will, the deployment of certain systems may negatively impact the stakeholders or the intended users.

One example is the chatbot system developed by [96] and deployed on Twitch. It designs a novel interaction approach that users can "raise" the chatbot as a pet through a number of commands, such as "!feeding". However, the deployment of such a chatbot may distract users' attention from their original goal of using the platform, which is to watch videos and socialize with the host and the other community members. Thus, with the chabot, users may engage with the platform but not with each other. Such close bonding with a chatbot may even hurt the individual user's benefits in the long term[4] and also negatively impact the community's engagement level [131].

---

[4]A report says there are users interacting with Microsoft's XiaoIce for more than six hours restless, and treat XiaoIce as his girlfriend [131]

The unexpected consequence of a chatbot's deployment is not uncommon when putting AI and machine learning systems to practical use. Often, today's NN-based algorithm research requires a large amount of data. Such data may come from an online community (e.g., Reddit), and they need to be tagged with ground truth labels by human annotators. However, the benefit of the original human annotators may be neglected during the training and deployment of the algorithm, as the algorithm's performance and optimization is developer's most important objective. For example, developing a functional customer service chatbots needs a significant amount of training data labeled by human customer service experts or obtained from their practices (e.g., [44, 122]). But the deployment of such chatbots may cause companies to hire fewer customer service workers in the future, which may also in return reduce the sources of training data.

Fortunately, some HCI researchers have noticed this challenge and they jointly work on an emerging research topic – **Human-Centered AI** – that aims to take an algorithm's impact on human users into the consideration of the algorithm design [17, 38, 39, 59, 90, 119, 133]. For example, Woodruff et al. interview 44 participants from several marginalized populations in the United States. Participants indicate that algorithmic fairness (or lack thereof) could substantially affect their trust in a company or product, if such application is deployed [119]. Thus, such fairness considerations should be taken into account during the algorithm design [59, 90]. In addition to the fairness, various other design considerations may also influence the eventual consequence of an AI system in the real world deployment, such as stakeholders's tacit knowledge [133] and community involvement [60]. It is also important to build an in-depth understanding of user needs or even involve them in the design process, as exemplified by a few recent work adopting participatory design research method [17, 38, 60].

In addition to incorporating the various considerations into an AI system design, there are also some good practices that one can follow in the deployment of the AI system [32, 38, 39]. For example, a group of researchers designed and developed ORES – an algorithmic scoring service that supports real-time scoring of wiki edits using multiple independent classifiers trained on different datasets. This paper proposes an example of deploying AI algorithms in an online community, but their algorithm is less explicit to the users, compared to the chatbot systems that we aim to develop and deploy in this paper.

In this paper, we design and develop a chatbot system that can provide social support for an online health community. Besides the objective of ensuring a good functional performance, we are also interested in evaluating its potential impacts on the online community after its deployment. This work joins the recent **Human-AI Collaboration** research effort [30, 106, 109, 123] that aims to develop and deploy AI systems that can work together with people, instead of replacing people. It differs from the Human-AI Interaction discussion [6], as it goes beyond the usability and interactive design of AI systems, but focuses more on the cooperative nature of AI systems with human partners and their context (e.g., [36, 43]).

3 STUDY 1: EMPIRICAL UNDERSTANDING OF THE RESEARCH SITE

3.1 Research Site

In this subsection, we first provide an overview for the research context—YouBaoBao—one of the largest and most popular online health communities for pregnancy and parenting discussions in China. Users are often pregnant women or couples that are expecting a baby. They can discuss a variety of topics on this platform, such as pregnancy, childbirth, childcare, and early education. This community has 164 sub-forums (similar to "sub-reddits" in Reddit), such as specific sub-forums for different stages of the pregnancy: "First Trimester", "Second Trimester", and "Third Trimester". YouBaoBao is a self-organized health community, where community members can publish a post to
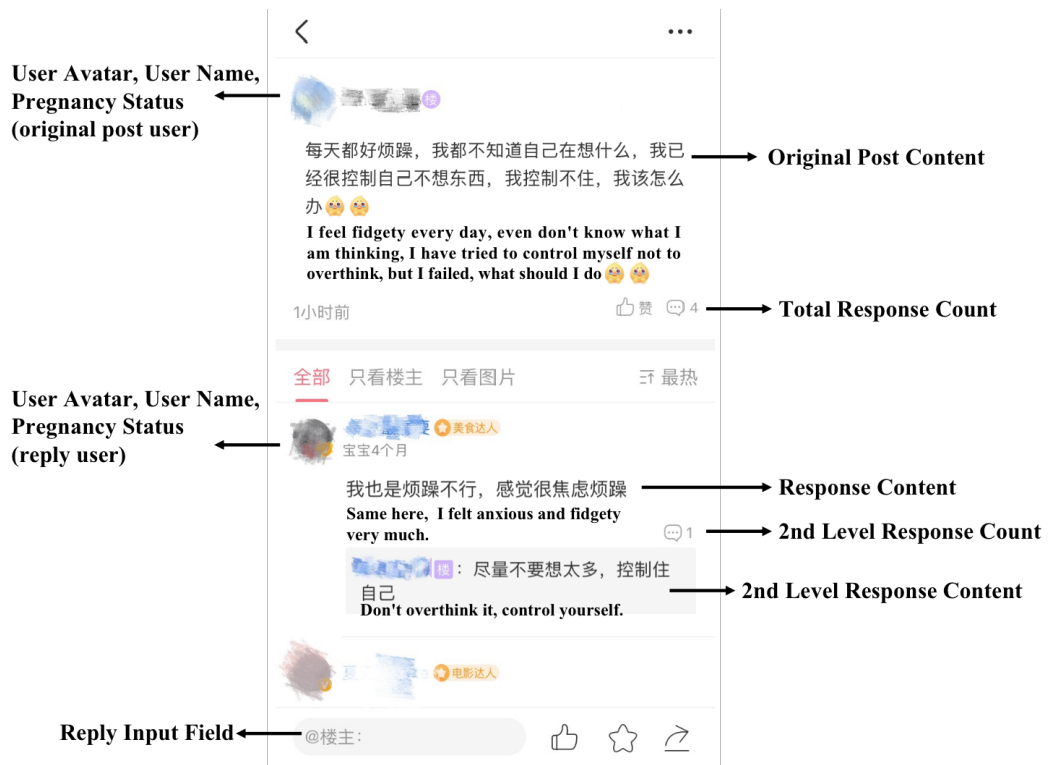
Fig. 1. User Interface of YouBaoBao community. It has a original post section to the top, a response section in the middle, and a input field for typing response at the bottom. The content is translated into English in this figure.

seek advice and information, and they can reply to a post to provide support and advice. Community members have an option to disclose their current *pregnancy status*, which is always displayed next to their user name. There are three pre-defined categories of pregnancy status from which a user can choose: preparation stage (e.g., "planning to have a baby"), three-trimester stage (e.g., "due in 8 months"), and postpartum stage (e.g., "having a 4-month old baby"). It is worth noting that members can post, browse, and reply in any of the sub-forums regardless of their labeled pregnancy status. For example, we saw instances where new moms posting and replying in the "first-trimester" sub-forum, despite they have already delivered.

A *post* or *response* has minimum and maximum word limits—the *content* has to be between 6 and 3000 words. When a post is created, the system can automatically recommend a sub-forum for it according to its content and the user's pregnancy status. Users can also manually modify the recommended sub-forum. Fig. 1 illustrates a typical post and responses it received (a direct response to the post and a second-level response). The user profile avatar, user name, and pregnancy status are displayed right above the content. The system also shows the *total number* of responses for a post, and of second-level responses for a first-level response.

3.2 Method

To better understand what challenges the community faces and where chatbots can support, we first conducted both descriptive statistical analysis and qualitative content analysis (following the method in [37]) to understand the characteristics of support-seeking activities in this community. Specifically, we explored the following **research questions**: What types of support do users (e.g., pregnant women) seek in this online community? How do community members respond to those support seeking posts? What issues or barriers exist in communications and interactions in the community?

3.2.1 Data Collection. In order to gain a comprehensive view of support-seeking across different stages of pregnancy, we chose five sub-forums with each focusing on one specific pregnancy stage: *"pregnancy preparation", "first-trimester", "second-trimester", "third-trimester", and "having a baby less than 6-month-old".* We collected six months of data from September 2018 to March 2019 through the community's application programming interface (API), and organized them into post-response pairs (N = 220,000). All the data we collected were publicly available but we removed all identifiable personal information (e.g., user names and status) in consideration of research ethics. We also replaced images with a special icon if there were any images in a post content. This data collection method is standard in various HCI [37, 125] and NLP research literature [101, 112]. This study is approved by the first author's university Institutional Review Board (IRB).

As this labeled dataset was later used for model training purposes, two coders removed the malicious posts, such as advertisements, before performing any detailed analysis. We followed iterative coding process and conducted axial coding: two coders first independently coded a group of 200 randomly-selected posts to identify 11 themes and organized these themes into categories. Then, the research team met and discussed the coding schema and differences in coding to resolve all disagreements. After that, the same coders conducted another round of independent coding with a group of 300 randomly-sampled new posts, and discussed newly emerged categories and disagreements in coding. They repeated this process until there was no newly emerged categories or inconsistent codings. In total, they repeated three iterations with each time analyzing 300 new posts (200+3*300=1,100 posts). At the end, we randomly sampled another 2,300 posts for computing the coding reliability score. Based on the agreed coding schema, the two coders independently coded these 2,300 new posts and reached a high inter-rater reliability (Cohen's kappa = 0.86).

3.3 Result

3.3.1 Characteristics of Support Seeking and Providing in the Community. The descriptive statistical analysis helped us understand the baseline of user behaviors in the community. These findings later were used to guide the design decisions of the CASS system. More specifically, we observed that, on average, the users on this community posted 5,000 posts per day, and each post had an average of 6 responses. However, approximately 18% of the posts did not receive any response. To further investigate what these posts are and how they can be replied, we randomly sampled another 60,000 posts out of the collected 220,000 samples for further analysis. We calculated the *Interval Time* between the original post was published and the first response was made. As shown in Fig. 2, the median interval time is 10 minutes. Later, we used this number in Study 3 as a threshold to evaluate whether a post gets timely response after the deployment of our chatbot system.

We also analyzed the time span from the beginning as the original post was published to the end as its last comment was made. This indicator reflects the retention of a post and its discussion thread in an online community (e.g., how long a post remains "alive" to the community members). We name this as the *Lifespan* of a post. We found that the average lifespan for a post thread is about
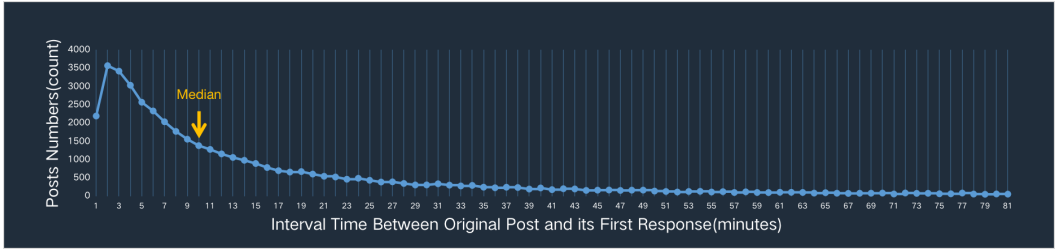
Fig. 2. Distribution of Interval Time between a post is published and its first response is published. Median is 10 minutes.

6 days. We define the lifespan for a post with no response as 0 minute. These numbers were also used in Study 3 as thresholds for the Control Logic Module in step 3 in Fig. 3.

There are some other measurements that we calculated to describe the users' behaviors and the community's characteristics, but the results are similar to the numbers we will report in Study 3, No-Chatbot group, thus we will report them only in Section 5.2.

3.3.2 Categories of Posts. The content analysis of posts led to the identification of 11 concepts, which were then grouped into 3 high-level categories, including *Informational-Support Seeking, Emotional-Support Seeking, and Sharing Daily Life*. As shown in Table 1, 27.45% of the posts were related to seeking emotional support or sharing emotional status, 45.95% of the posts were about seeking informational support, and 26.60% of them focused on sharing posters' daily life.

Emotional-Support Seeking refers to posters expressing their frustration and stress (e.g., complaining something happened in their work or life), or sharing positive news (e.g., announcing the confirmation of pregnancy) to seek encouragement or empathy. Informational-Support Seeking refers to seeking information, knowledge, advice, and suggestions from others to manage a situation (e.g., how due dates are calculated). Sharing Daily Life refers to posts where community members post a photo of food or exercise in a gym, or share updates and progress of pregnancy.

The resulting categories are similar to previous findings [37, 125], confirming that our research site (YouBaoBao) has similar characteristics as other popular pregnancy-related online communities (such as BabyCenter[5] and TheBump[6]). More importantly, the analyses of posts helped us understand the types and nature of emotional support sought by users. Thus, our primary goal for the chatbot is to provide non-informational support (both emotional-support and sharing dailylife) to the posting users. Later in Section 4.3, we will illustrate how we translate these learned contextual knowledge into parameters, and feed the labeled data to train the algorithm models in the system architecture.

4 STUDY 2: BUILDING A NEURAL-NETWORK BASED CHATBOT WITH A SCALABLE AND GENERALIZABLE ARCHITECTURE

Based on the literature [37, 97, 102] and the knowledge gained from Study 1, we learned that one major challenge faced by online health communities is that many posts can not get a timely response (e.g., more than 18% never got a response); for the posts that have responses, it often takes more than 10 minutes to get the first response. In the YouBaoBao context, this challenge is particularly frustrating considering the pregnant women are already at high stress level. To address this prominent challenge, we built a chatbot system to explore the feasibility of chatbot for providing

---

[5]https://www.babycenter.com/
[6]https://www.thebump.com/

Category	Concept	N	%
Emotional support seeking and emotion sharing	Complain about work, life, family, etc.	631	27.45
	Have a fantasy conversation with the baby to be born		
	Share happiness		
	Make a good wish		
Informational support seeking	Seek health-related information	1057	45.95
	Seek personal opinions from peers about pregnancy		
	Seek opinions and suggestions on baby-related issues		
Sharing daily life	Share food and cuisine	612	26.60
	Share sports and other activities		
	Share updates of her body along with pregnancy		
	Share updates and progress of the baby or fetus		

Table 1. Axial coding result with 3 Categories and 11 Concepts for original posts. (N = 2,300, Cohen's kappa = 0.86)

timely response to support seekers. Because the responses to the emotional-support-seeking posts and to the sharing-daily-life posts can be overlapped with compliment or empathy thus our text generation module can handle them the same way. We combine these two non-informational support-seeking categories, and train the models to detect them from the informational-support-seeking posts.

In this section, we will introduce how we build the system, what design decisions we made along the implementation process, and lastly how it performs using common natural language processing (NLP) performance metrics and human judgement in offline evaluation[7].

4.1 CASS: Chatbot As a Social Supporter for Online Health Community

In this study, we propose a system architecture that works in the following five steps (Fig. 3), and we build a chatbot system–**CASS**–using this architecture:

- **Step 1.** Collecting and monitoring new posts (Input interface: an API URL to fetch data from the community);
- **Step 2.** Classifying posts into emotional-support seeking category and informational-support seeking category using a trained CNN model (Input: a labeled dataset with posts, responses, and a post's category out of the 3 coded categories);
- **Step 3.** Randomly dispatching half of the emotional-support seeking posts into a control group and monitor them, this is only for the field experiment in Study 3 (Parameter: 10 minutes that derived from Study 1, as the threshold for deciding whether the chatbot needs to respond to the overlooked post or not);

---

[7]The offline evaluation is in contrast to an online evaluation approach. Offline evaluation uses machine learning testing or holdout subset of the data to evaluate the performance of a model or system, whereas online evaluation means deploying the model back to the application environment.
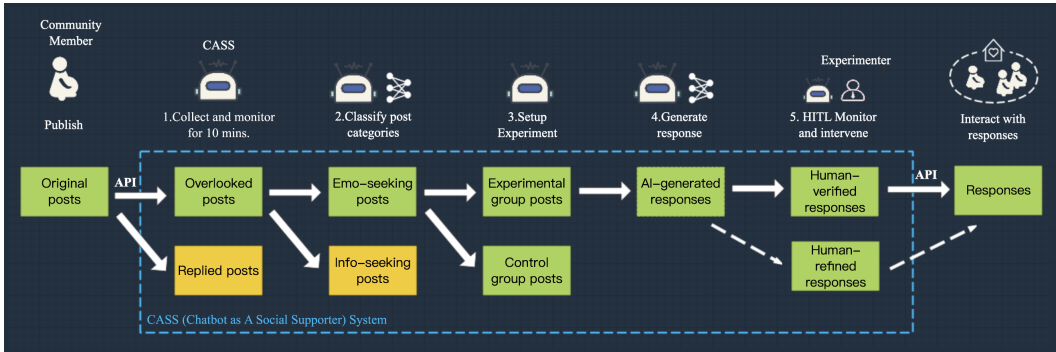
Fig. 3. CASS system architecture and workflow. It has 3 modules and 5 steps. Steps 1, 3, and 5 are implemented via the control logic module; step 2 applies the neural network (NN)-based text classification module; and step 4 applies the neural network (NN)-based text generation module.

- **Step 4.** Generating responses for posts in the experimental group using a trained LSTM model (Input: the same input as in step 2 – the labeled dataset);
- **Step 5.** A human-in-the-loop verification and intervention user interface, before posting the response back to the forum (Output interface: an API URL to post response back to the community).

This chatbot architecture and workflow is quite generic and easy to generalize to other online health forums — a developer or researcher only needs to alternate the two API URL parameters on how the system reads in original post from the community and how it publishes generated response back to the community, and to feed a labeled dataset as input. In the following subsection, we present how we use the architecture to build the CASS system with 3 modules: **a control logic module, a text classification module, and a text generation module**.

4.2 Control Logic Module

We collectively refer to the posts that can not get a response within 10 minutes as **overlooked posts**, and the others as replied posts. CASS's control logic module is created to identify and track those posts, dispatch them to the corresponding system modules for text classification or text generation or for human-in-the-loop intervention, and finally publish them back to the community. In Fig. 3, the control logic module are responsible for the implementation of **Step 1, 3, and 5**.

Step 1 and 5 are responsible for collecting posts from the community (i.e., input) and publishing responses to the community (i.e. output), respectively. Step 1 is also responsible to distinguish whether a post is timely-replied (yellow box in Step 1 in Fig. 3; no further interventions were provided for this type of post) or overlooked (green box in Step 1 in Fig. 3).

In addition to its automated nature of the output step for CASS, we also built a Human-in-the-Loop (**HITL**) module in Step 5. We hypothesized there may be malfunction of the CASS system as it is fully automated after the model is trained and the system is deployed. Thus, we designed a user interface console, which can be used by a human experimenter to track and monitor each AI-generated response in real-time. Every response generated by the text generation module in Step 4 will show up in the console for 10 seconds before being published to the community. The human experimenter has two options for action: he/she can choose to do nothing then the response is published; or he/she can hit "Enter" button to suspend the publishing process. If the latter action is

triggered, the human experimenter will be prompted to type in a new sentence in the console, and that new sentence will replace the AI-generated response and then will be sent to the community.

In the actual online field experiment in Study 3, two experiment operators carefully monitored the performance of the chatbot in generating responses from 8 a.m. to midnight (16 hours) using the HILT module. But they did not find any inappropriate AI-generated responses that need intervention. This may be because the data collected from YouBaoBao and used to train the models were clean and had nearly no malicious speech corpus. This may also attribute to the friendly nature of the research site [37].

Step 3 is a control logic specifically designed for the field experiment in Study 3. We want to note that step 3 is not needed in a real deployment of the CASS system in the future. This step randomly assigns half of the qualified posts (i.e., overlooked emotional support seeking posts) into an experimental group, and the other half into a control group. For the experimental group posts, CASS generates responses, sends them through HILT step, publishes them on the forum, and keeps tracking for 7 days for the post owner's (i.e. poster) and other community members' (i.e. commenter) actions. For the control group posts, CASS simply tracks the user actions for 7 days without intervention for comparison purpose.

4.3   Text Classification Module

Our results from Study 1 suggest that a user typically seek two types of support: informational- and emotional-support. For informational support, posters look for accurate responses rather than diversity in response [114]. Thus, this type of utilitarian needs can be satisfied by various rule-based algorithms or existing chatbot platforms (e.g., [64, 122]); this is not the focus of this study. As such, we build a text classification module based on convolutional neural network (CNN) (**Step 2** in Fig.3) to distinguish informational support seeking posts and non-informational seeking posts (i.e., both emotional support seeking and sharing daily life) so that the chatbot can provide timely response to non-informational support seeking posts only.

CNN models have been widely used in the computer vision field in recent years [56], and some variations have been extended to text classification tasks [51]. It normally requires preprocessing each data instance (i.e., posts in this paper) into word vectors with a same dimension, then feed the word vectors into a CNN network; the output of the network is a classification of the input post into one of the three categories. In our study, we do not need to distinguish the detailed categories — we only need to categorize posts that are informational support seeking, so that our chatbot logic module can ignore those posts. To train this model, we used the labeled data (N=2,300) from Study 1 as training data. Out of the 2,300 labeled posts, 45.95% of them were seeking informational support. To address the data imbalance issue, we oversampled the data [13], adjusted the ratio of informational support seeking posts to emotional support seeking posts to 1:1 (in total 3,000 posts).

After the training dataset is prepared, we first preprocess the training data and encode each post as word vectors using tensorflow [1] (vocabulary size=5000, embedding dimension=64, sequence length=600), followed by feeding these vectors into the CNN model. Our CNN model uses a simple structure, including one convolutional layer, two full-connection (FC) layers, one input layer, and one output layer. The input layer has 600 neurons. The convolutional layer consists of three consecutive operations/layers: convolution with kernels, non-linear activation function, and max pooling. The convolutional layer contains 256 kernels, and the FC layer contains 128 units with a dropout rate of 0.5.

In training, we optimize the model with AdamOptimizer using a learning rate of 0.001, optimized for accuracy. After training, we conduct 5-fold cross-validation. The overall cross-validation accuracy is 0.86 and F1 is 0.87. This accuracy is good enough for our following experiments.

4.4 Text Generation Module

In **Step 4**, we build a text generation module. It can be considered as a task that for a given input sequence of words (i.e., post), we need to find another sequence of words (i.e., response) that best suit its input. Thus, we can conceptualize this as a NLP translation task which can be solved using a variation of machine translation networks. To that end, we use OpenNMT[52], which is a state-of-the-art open-source toolkit[8] for neural machine translation (NMT) tasks, to build a model to generate responses for a given post. The model is derived from a sequence-to-sequence model with attention mechanism [69]. In addition, we build a information retrieval model (IR) with BM25 ranking function [91] as a baseline model.

The training data is 220,000 post-response pairs from the 5 sub-forums. For each post-response pair, the data contains the post content, a picture icon if the content contains pictures, a post ID, # of views, # of responses, response ID, response user ID, response content, and response timestamp.

First, we pre-process and embed the input sentence into word vectors (500 dimensions), and then feed them into a Long Short-Term Memory network (LSTM) [41] as encoder. In the preprocessing step, we first apply the text classification (Section 4.3) to distinguish and filter out all the informational support seeking posts and its pair. Then, we manually check the posts to see if there are advertisements (e.g., "sell infant formula") or offensive words in the original posts. We do not see any posts using offsensive words but we find many advertisements. Thus, we define a list of stop-words to eliminate advertisement posts. After filtering, 81,000 post-response pairs are left, and we use them as our training dataset.

Second, in the LSTM model training step, attention is applied over the decoder LSTM and combined with the current hidden state at each target time step to produce a prediction of the next word. This prediction is then fed into the decoder LSTM. Specifically, we train a 2-layer deep LSTMs, and each layer has 500 memory cells, using stochastic gradient descent and 0.2 dropout. The batch size is 64, and the global attention translation is conducted using Multilayer Perceptron (MLP). In total, we train 100,000 iterations.

In Fig. 4, we illustrate 6 pairs of post-response (originally all in Chinese). The human-response column contains the ground truth data that real human users commented under a post; and the AI-response column contains example responses generated by our text generation module. As illustrated, these generated responses are vivid and diverse, and some of them even contain emojis.

4.5 Evaluation

In this subsection, we present offline evaluation of CASS performance. Offline evaluation refers to the evaluation that is done before deploying the system to let people test in the real world context. There are two types of evaluation approaches: automated performance evaluation and human evaluation. In this study, we adopt both of them to conduct an offline evaluation of our chatbot system.

Automated performance evaluation is an established mechanism to automatically and quickly evaluate the performance of an AI system in the field of machine learning and AI. It often first reserves a subset of the training data as holdout data, whose both input (i.e., post) and output (i.e., human response) are known. Then, we feed input into the train model which automatically generates a predicted output. We can automatically compare the predicted output with the known human response output to evaluate the model performance.

The human evaluation is a well-known practice to both HCI and machine learning communities. Often, we can define a variate of dimensions and ask human users to rate the AI predicted outputs. Depending on the task, the human graders/coders are not necessarily having to be the actual

---

[8]https://opennmt.net/

Post	Human-Response	AI-Response
I finally got pregnant on the 20th of this month. I had been preparing for it for half a year. It is so hard. 这个月20号测出来怀孕了。备孕半年了。不容易啊。	Congrats, Congrats, Pass on your luck to me. 恭喜，恭喜，接好孕。	CongratsCongratus, Pass on your luck to me. 恭喜恭喜，接好孕。
I heard that dreaming of a snake means I will have a boy baby, but I really want a girl. 🙂 听说梦到蛇生男孩，可是我想要女儿。🙂	I also want a daughter, I dreamed of snakes too last night. 我也想要女儿，我昨晚也梦到蛇了。	It's not true, it is just a dream. Do not overthink it. 不一定，只是梦而已。不要想太多了。
My baby's 13-day photo is so lovely, mom loves you so much. 🖼 宝宝12天照的月子照好萌，妈妈好爱你。🖼	What a lovely baby, what big eyes. 好可爱的宝宝，眼睛好大。	😊 😊 😄
I'm so sensitive and emotionally fragile as a mom. 当妈的玻璃心。	I wish my belly is transparent, so I can always see what the baby is doing. 恨不得肚皮变成透明的，随时可以看到宝宝在干嘛。	What happened? 咋了？
Taking care of a 1-2 months baby can make people frustrated. I can not take a nap during the day, because the baby sleeps very lightly. 陪1-2个月的宝宝成年人真的会崩溃。不能午睡，因为宝宝睡得不安稳。	Totally understand, same here. The baby has been crying himself out of bed, and has to be held, my back is sour, but I have to hang in there. 非常理解，我也是。孩子不睡觉哭唧唧，还总得抱着，腰酸背痛，只能熬着。	Same here. 我也是。
My baby feels uncomfortable, he suddenly cries with a scratchy throat. 宝宝很难受，突然间声音就哭哑了。	My baby also cries with a scratchy throat, is your baby recovering, what did you do, she doesn't like drinking water. 我宝宝也是声音哭哑了，你宝宝好了么，怎么解决啊，又不喝水。	Take him to the hospital. 去医院看看吧。

Fig. 4. Some examples of human user commented responses as ground truth (Human-Response) and our text generation module generated responses (AI-Response). These sentences are all in Chinese and we translate them into English for this paper.

intended users as long as these graders are capable of completing the task. Thus, many such human evaluations are done by crowd-workers [122].

4.5.1 Automated Performance Evaluation. In NLP research, there are many metrics to describe the performance of a machine learning system, with each metric is suitable for different tasks. In our system, the final output is a predicted/generated response based on an input (i.e., a post). This is a Natural Language Generation task which can be evaluated by a widely used metric called *BLEU* score (Bilingual Evaluation Understudy) [80]. BLEU score represents the similarity between CASS generated response and the human ground truth in the training dataset with a n-gram match. It is a number between 0 and 1; 1 indicates that the generated response is exactly the same as the ground truth. We randomly selected 2,000 post-response pairs and reserved them as holdout data before training. The BLEU score on this holdout dataset was 0.23, which is a fairly acceptable performance score [122, 129]. As a comparison, the BLEU score of IR-based baseline model is 0.03.

4.5.2 Human Evaluation. BLEU score simply describes the lexical differences of two sentences. It, however, does not reflect the grammar correctness or the semantic meaning. It is possible that CASS generates a much shorter sentence with the same meaning to the original human response,

but the BLEU score is low due to the length difference. For example, in the second row in Fig. 4, AI-generated response is a valid and even better response to the post, but the BLEU score is low because it is quite different from the human-response. Thus, we recruit human coders to evaluate the generated results. Based on previous literature [5], and for the research interest of this study, we introduce four dimensions for evaluating AI-generated responses:

- Grammar Correctness. The human coders are asked to evaluate the generated response's grammar correctness with a 5-point Likert-scale ranging from strongly disagree (-2) to strongly agree (+2) that the grammar is correct [89].
- Relevance. The response may not be relevant to the given input post so we decided to evaluate the relevance of AI-generated responses. It is also done using a 5-point Likert-scale ranging from strongly disagree (-2) to strongly agree (+2) that the response is relevant to the post's topic [89].
- Willing-to-Reply. A response's content may be correct and relevant, but users in an online community may not feel engaged so they tend to ignore the AI-generated response. In this question, we ask human coders to rate how likely they will comment on the automatically generated response. Again, a 5-point Likert-scale is used with -2 indicating strongly unlikely whereas +2 representing strongly likely.
- Emotional Support. This is a dimension designed specifically for our research context. We are interested in examining the extent to which the human coders perceive that the AI-generated response provides the desired emotional support. Similar to the above dimensions, a 5-point Likert-scale is used.

We randomly select 200 post-response pairs from the holdout dataset (the same dataset used in machine evaluation). Each post has both a human response that we collected from the community as a ground truth, and a AI-generated response. Thus, we end up with a total of 400 pairs of post-response. We ask human coders to evaluate both the AI-generated responses and human ground-truth for comparison. We recruit five human coders to rate each of the 400 pairs; each coder provide 4 scores related to grammar correctness, relevance, willing-to-reply, and emotional support dimensions. Intra-class correlations ICC(3,1) for the four dimensions range from 0.74 to 0.99, which indicates a good inter-coder consistency. For each response, it has four dimension scores, and we use the average from the five coder's scores. Then we perform a t-test to compare the AI-generated response and human ground-truth's quality in each of the 4 dimensions (as shown in AI-Response and Human-Response groups in Fig.5).

Fig.5 shows that AI-response's grammar correctness is rated high and not that different from human-response (1.37 vs 1.52, t(199)=-4.393, $p > 0.1$). However, AI-response's quality is not as good as human-response in topical relevance, willing-to-reply, and emotional support, suggesting that the chatbot should be improved on those dimensions. But building a system to outperform human in writing responses is never the goal of this study. Despite the difference, all the evaluation scores are above 0, meaning that the 5 human coders agree that the AI-generated responses are grammatically correct, of high relevance to the topic, engaging enough for users to reply, and providing a considerate level of emotional support to the poster.

The results of the automated performance evaluation and human evaluation confirm that our chatbot system has a pretty good performance in providing emotional support. Therefore, we deploy the CASS back to the research site to provide timely responses and emotional supports to overlooked posts. We will describe our field deployment and experiment in the next section.
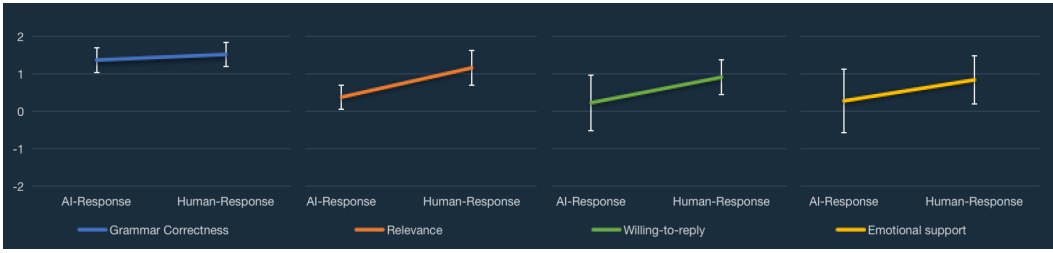
Fig. 5. Comparison of AI-generated responses' and human ground-truth responses' quality with 5 human coders on 4 dimensions. We also performed t-test and post-hoc analyses. Grammar correctness: 1.37 vs 1.52, t(199)=-4.393, $p > 0.1$; Relevance: 0.38 vs 1.16, t(199)=-12.062, $p<0.01$; Willing-to-reply: 0.23 vs 0.91, t(199)=-10.657, $p<0.01$; Emotional support: 0.28 vs 0.34, t(199)=-7.186, $p<0.01$.

5 STUDY 3: DEPLOYING CASS TO THE STUDY SITE AND EVALUATING ITS IMPACT ON SUPPORT SEEKERS AND OTHER COMMUNITY MEMBERS

In the previous two sections, we presented an exploratory study on the YouBaoBao community to build an contextual understanding of the study site, from which, we identified what types of posts the community members post, and highlighted the challenge that some of their emotional support seeking posts can not get a timely response (Section 3.3). Then, we presented CASS system with a fully automated end-to-end architecture that can read in posts, identify overlooked posts, classify emotional support seeking posts, generate responses for those overlooked posts, and publish the generated response back to the community (Fig. 3). Our preliminary evaluations showed that the NN-based models have a satisfying performance score, so the CASS system is ready for deployment. In this section, we present how we deployed the CASS system back to the community following a field experiment research setup, and how we evaluated its impacts on individual members and the community.

5.1 Method

Field deployment is a well-established HCI research practice [100]. It can reveal users "naturalistic usage" of the introduced new functionality of the system. We follow [100] guideline on how to define the beginning and the end of the field deployment, and its ethical considerations of engaging users.

Moreover, to quantify the support that CASS brings to individual members as well as to the community, we also adopted an experiment study setup [33]. We design a "control group" and an "experiment group" to measure the difference of its impact between users exposed to it and the ones that are not exposed to it. An experiment design in a real-world deployment has its unique challenges. For example, to ensure the users having the most naturalistic behavior while interacting with the intervention, they should not know they are part of an experiment, when such disguise is not harmful to the users. We follow the *field experiment* research method described in the textbook by Terveen et al. [102] to "maximize realism of the context, while still affording some experimental control".

5.1.1 Field Experiment Design. Our research goal is to deploy CASS to provide timely social support for the individual members who seek support. In addition, we are also interested in evaluating its impact on other community members. We deployed the system in August 2019, and conducted the field experiment for 7 days. The reason for deploying the system for 7 days is that the lifespan of

a post thread on average is 7 days (as reported in Study 1, Section 3.3). After we deactivated the CASS system, we still kept tracking of the comments and user actives for another 7 days to ensure we gathered a 7-day data for each of the post.

To evaluate whether CASS helps individual support seekers, our unit of analysis is each individual user who published the original post. We define measurements to capture their online behaviors, and their emotion changes after the post was responded by CASS. To evaluate CASS's impact on other members of the community, the unit of analysis is changed to each individual post and its comments as a thread. We also define measurements (e.g., how many members participated under a post thread) to reflect other community members' behaviors caused by the CASS intervention.

As shown in Study 1 in Section 3.3, the median *interval time* for a post to get its first response is 10 minutes; and in Study 2 in Section 4.2, we define a post as *overlooked post* if it does not get a human response within 10 minutes. Thus, during the deployment, we design CASS to only track the overlooked non-informational support seeking posts (Section 4.3). In total, CASS tracks 3,445 overlooked emotional-support-seeking posts during the 7-day field experiment.

These overlooked posts are then randomly split into two groups: a **baseline condition**, where we only passively track the original poster's and the other members' activities without CASS intervention; and an **experiment condition**, where CASS responds to the overlooked post once 10 minutes has passed, and we track activities and measure the poster interactions. During the 7-day field experiment, 1,717 overlooked posts are assigned to the baseline condition, and another 1,728 overlooked posts are in the experiment condition.

5.1.2 Disguise the AI Identity. As instructed in the textbook [33], when possible, participants should not be aware that they are in a field experiment setting. In addition, a number of recent literature [49, 68, 75] has highlighted that users' behaviors and perceptions may change if they know or believe they are interacting with an AI system, which will further compromise the experiment results. Therefore, during this experimental study, we disguised CASS as a normal community member with a pseudo user name and a user avatar. We disclosed the chatbot's real identity to all the users who had interacted with it via the community built-in private messages after the study was completed. This is a common practice in psychology experiments [33] and clinical trial experiments [26].

5.1.3 Measurements. In this section, we present the measurements to quantify the effectiveness of our chatbot system in providing social support. In McGrath's seminal work, *"Time, Interaction, and Performance"* [73], he proposes a three dimentional framework to describe teamwork: *production, member support, and group well-being*. Grudin argues that a successful CSCW system should be able to provide productivity benefits for not only the individual members, but also for the group dynamics [35]. Inspired by these literature, we thus define a set of measurements to quantify CASS's impact on other community members.

The three measurements for determining the CASS's *effectiveness* in generating timely social support include:

- *The count of posts with no response.* This variable describes how many posts have no response at all in 7 days (in short **Post-NoResp-N**). If AI-Response group has a significant smaller number in this measurement, that suggests CASS is effective in addressing no-response posts. When calculating this score, the CASS's responses are not counted, as the chatbot replies to every post in the AI-Response group.
- *The time interval between the original post and the 1st response, and the time interval between the 1st response and the 2nd response.* Here we calculate two variables, both time intervals in minutes, to describe how fast a post gets replied (in short **Post-1Resp-Time**), and how fast

the 1st response could attract the 2nd response (in short **1Resp-2Resp-Time**). If AI-Response group has a smaller time interval measurement, that suggests CASS indeed supports the poster to get a response in a more timely manner.

Then, we use a set of four measurements to reflect whether the CASS-generated post helps the original support-seeking *individual*.

- *The count of follow-up comments from the original poster.* After the original poster publishes the post and some other users come into the thread and reply to the posts, will the original poster come back to publish follow-up comments and form an active discussion? If so, we count the number of comments from the original poster, in short as **Poster-Comm-N**. The higher this number is, the more active the poster is. Thus, the support-seeking poster found an emotional outlet under her post and leveraged it to engage more with other community members. If the poster never came back, we count it as 0.

- *The poster's original-emotional states, updated-emotional states, and the difference of these two states.* It is difficult to measure the posters' emotional states without directly asking them, thus we employ the *theory of emotional valence* [67], and recruit two coders to annotate an original post's emotional valence score as a proxy for the original poster's initial emotional state (in short **Poster-Emo-Val-Orig**). More specifically, there are three states of emotional valence: *1 for positive (happiness, contentment); 0 for neutral; and -1 for negative (anger, sadness)*. Then, if the original poster publishes a follow-up comment, we use this comment's emotional valence score as a proxy to represent her updated emotional state (in short **Poster-Emo-Val-Updt**). We also calculate the change of emotional states (in short **Poster-Emo-Val-Chng**). If an original poster never come back to post a follow-up comment, we exclude this data point. In summary, we have 759 data points in the experiment condition (N=1,728); and 536 data points in the baseline condition (N=1,717). We refer to these data points as AI-Response and Human-Response data points. Inter-rater reliability between the two coders on the emotional valence score of these 1319 data points reaches 0.88 (cohen's kappa), which indicates a high consistency.

Lastly, we define three measurements to reflect CASS's impact on other members in the community.

- *The count of total responses to an original post.* This variable reflects the participation level under a post. A higher average number of this variable suggests the community has a higher community contribution level, and the community is healthier (in short **Post-Resp-N**).

- *The count of how many members participated and commented under a post thread.* This variable also reflects the participation level under a post. A higher average number of this variable suggests on average more community members engaged in each post's discussion. It reflects a high community commitment level, and the community is healthier (in short **Post-Member-N**).

- *The time interval between each pair of adjacent responses under a post.* This variable (in short **Adj-Comm-Time**) represents on average how fast a post thread can get a new comment. If this number is smaller in the AI-Response group, it suggests that CASS not only helps the original poster to get a timely response, but also activates the liveliness of the community (e.g., other members also participate in a post's discussion in a more active fashion).

5.2 Result

We organize the result section into three subsections to report that 1) *CASS is effective in providing timely social support to overlooked posts*; 2) *CASS-generated responses can improve the individual*

support seeker's emotional status; and 3) *CASS can also positively impact other community members' participation.*

5.2.1 CASS is effective in providing timely social support to overlooked posts. The measurements in this category represent how CASS's functionalities effectively mitigate the primary challenge that a emotional-support-seeking post can not get a timely response.

Post-NoResp-N: The results show that in the baseline condition, there were 595 out of 1,717 posts never received any response within 7 days of the post being published; in the experiment condition, out of 1,728 posts, this number was 433 (we exclude the response from the original poster, or from the CASS chatbot when calculating this number). A Chi-Square test suggests that the difference is significant ($\chi(1)$=39.65, p<0.01). That is, CASS can effectively generate timely social support messages as response to the support seeking posts and significantly reduce the number of overlooked posts.

Post-1Resp-Time & 1Resp-2Resp-Time: These two variables are used for calculating time intervals between the original post and the 1st response, and the interval between the 1st response and the 2nd response. The results show that **Post-1Resp-Time** in the baseline condition is 254 minutes on average, and in the experiment condition is 10 minutes. It is not comparable because we simply set the CASS logic to identify and reply to posts in 10 minutes. Thus, a more comparable measure is **1Resp-2Resp-Time**, which reflects how long it takes for a second response to come in after the 1st response was published. In the baseline condition, the time interval between 1st response and 2nd response is on average 462 minutes, whereas in the experiment condition that number is 349 minutes. An independent sample t-test suggests the difference is significant (t(1577)=-1.433, p<0.05). This result means that CASS not only posts a timely response to the original post, but also accelerates the time that a following response was posted by another user.

In summary, CASS's functionality can help effectively mitigate the primary problem, where some posts can not get a response in a timely manner, by reducing the number of posts with no response (smaller **Post-NoResp-N**), and speeding up the time that a post is responded (smaller **1Resp-2Resp-Time**).

5.2.2 CASS-generated responses can improve the individual support seeker's emotional valence. As introduced in section 5.1.3, we defined and calculated four measurements to quantify the impact of the CDSS-generated response on an individual support seeker.

Poster-Comm-N: The results show that original support seekers come back to their post thread and leave 1.36 follow-up comments in the baseline condition (N = 1,717). In contrast, the posters posted on average 1.78 follow-up comments in the experiment condition (N = 1,728). An independent t-test (t(3443) = 2.616, p<0.1) has a marginal significant effect and suggests that CASS indeed made the original support seeker to follow up with more comments, and interact more actively with other community members. These behaviors arguably are beneficial to release their stress.

Poster-Emo-Val-Orig: As shown in Fig. 6, the results show that for the experiment condition (N = 759), 22% of the original posts are coded by human coders as having positive emotion valence score, 29% as neutral, and 49% have negative emotion valence score. In contrast, in the baseline condition (N = 536), the numbers become 20%, 36%, and 44% for positive, neutral, and negative emotion scores, respectively. As these post-response pairs are only the ones on which the original support seeker left a follow-up comment; thus, to differentiate them from the entire dataset, we refer to them as the AI-Response group (N = 759) and the Human-Response group (N = 536), respectively.

Poster-Emo-Val-Udpt: In the AI-Response group (N = 759), the original poster's 1st comment is the response she replied to the CASS-generated response. 25% of these responses are coded as positive emotion, 62% as neutral emotion, and 13% as negative emotion. In the Human-Response
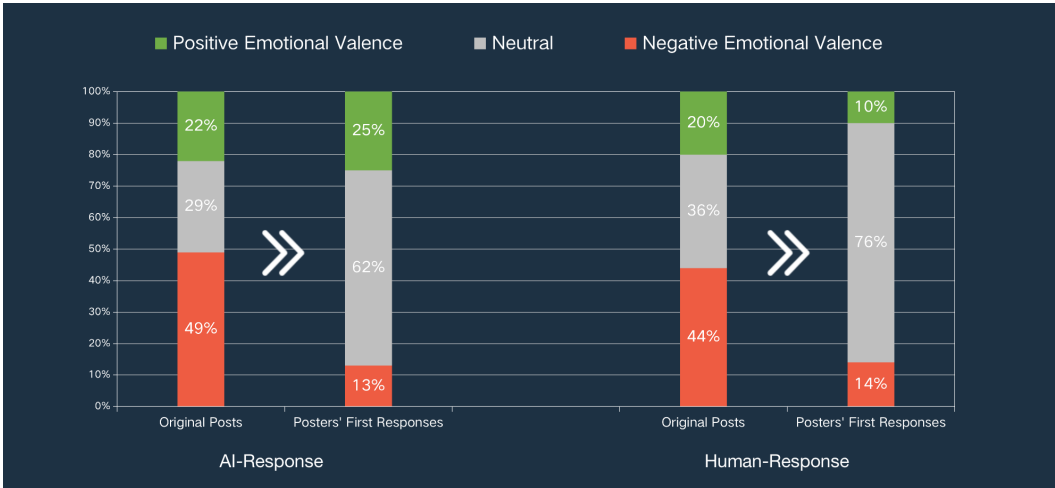
Fig. 6. The poster's emotional valence and their changes, as in two proxy variables: the emotional valence score for her original post, and for her first comment to others' response under her own post.

group (N = 536), human coders rate 10%, 76%, and 14% of the posts expressing positive, neutral, and negative emotion, respectively.

These results are shown in Fig. 6. From the chart, we can observe that as for the emotional valence score for the original posts, there is no difference between the AI-Response group and the Human-Response group. But we can see there is a clear difference in the emotional valence score of the poster's 1st response.

Poster-Emo-Val-Chng: It is interesting to see that the emotional valence score in both groups changed from the support seeker's original post to her 1st response, in Fig. 6. It appears that in AI-Response group, many posters' original posts are negative, but then change to neutral (29% to 62%), and some neutral ones convert to positive (22% to 25%). Many community members' emotional valence become more positive, as shown in the example below:

> Poster's original post: *I may have insomnia. I still can't fall asleep and it's already 3:00am in the morning.*
> AI-generated response: *Have a good sleep, and don't push yourself too hard.*
> Poster's 1st comment: *Wow!!! I want to hug you!*

In contrast, in the Human-Response group, many posters' emotional valence score change from positive to neutral (36% to 76%). That suggests, some of the responses posted by another community member may adversely affect a certain number of posters' emotion.

To statistically analyze poster's emotional change, we further calculate the percentage of the posts that have an emotional valence score going up (e.g., neutral to positive), going down (e.g., positive to neutral), or remaining unchanged. In the AI-Response group, 48% of the posts' scores go up, 38% remain unchanged, and 14% go down. In the Human-Response group, the percentages are 38%, 44%, and 18%, respectively. We perform Chi-Square test to compare the percentages across the two groups.

There are significantly more posters having an increased emotional valence score in the AI-Response group than in the Human-Response group ($\chi(1)=14.35$, p<0.01). Similarly, there are significantly less posters having decreased emotional state score in the AI-Response group than in

the Human-Response group ($\chi(1)$=5.242, p<0.05). This suggests that the AI-generated response can motivate posters to become more positive, and prevent them from experiencing frustration or a similar negative emotion, when compared to the human-generated response.

In summary, our results show that in the AI-Response group, more original support seekers (i.e. posters) revisit and comment under their own posts to interact with other community members. In addition, more posters turn to a positive emotional valence, and less turn to negative emotion, when compared to the posters in the Human-Response group. In this sense, CASS can help individual members use their published post as an emotional outlet to chat with others, and may further improve their emotional states.

5.2.3 CASS can also positively impact other community members' participation. In addition to CASS's utilitarian effectiveness and support for individual members, we are also interested in examining how it may improve the well-being of the community as a whole.

Post-Resp-N: The results show that in the experiment condition (N = 1,728), there are 7.63 (S.D.=34.25) responses (N = 997) for each post; in contrast, in the baseline condition (N = 1,717), each post has an average of 6.3 responses (N = 1120, S.D. = 35.14). We have excluded the responses published by CASS or by the original poster, so this variable is a pure indicator of other community member's participation. This result suggests that other community members in the experiment condition have a higher participation level, but an independent sample t-test suggests such difference is not significant (t(2115)=0.879, p = 0.228).

Post-Member-N: For each post, there are 5 other community members (S.D.=20.97) participated in the thread (N = 997) in the experiment condition. In contrast, about 4 members (S.D.=23.35) participated in each post thread in the baseline condition (N = 1,120). This result suggests that each post in the experiment condition has more community members participated, and it has a higher level of community commitment, when compared to the baseline condition. However, an independent sample t-test suggests such difference is not significant (t(2115)=0.911, p = 0.232).

Adj-Comm-Time: We also calculate the time intervals between each pair of adjacent responses in a post thread. In section 5.2.1, we have reported **Post-1Resp-Time** is 10 minutes in the experiment condition, and 254 minutes in the the baseline condition; **1Resp-2Resp-Time** is 349 minutes and 462 minutes, respectively. In addition, we continue to calculate the time intervals up to five follow-up responses. As Fig. 7 illustrated, there is a trend that the AI-Response group's time intervals are always shorter than the ones in the Human-Response group. This suggests that CASS not only helps the original poster to get a quicker response, but also energizes other members in the community, and accelerates their participation in a post's discussion.

In summary, we find that in the experiment condition, community members participate more actively in a post thread. Results also suggest that with the presence of our chatbot system, each post has a higher level of community participation, with both more people and more responses, even though difference is not significant.

6 DISCUSSION

6.1 How does Our Work Differ from Existing Literature on Online Healthcare Communities?

We start our exploration with a detailed content analysis of the posts on the pregnancy forum. Despite the granularity difference between these two coding schema, our category and concept axis echo previous literature [37]. In [37], the researchers propose 6 categories from coding 683 online posts; in contrast, we identify 3 top-level and 11 detailed-level concepts after coding 2,300 posts. Our "informational-support seeking" category is quite similar to a combination of their 3 categories

Fig. 7. Time intervals for each pair of adjacent responses under a post. In this chart, we only show the first 5 pairs, but this trend persists for the rest pairs. This result suggests the community is more active in the AI-Response group than in the Human-Response group.

"advice", "formal knowledge", and "informal knowledge seeking". Our "emotional support seeking and sharing" is similar to a combination of their "emotional support" and "reassurance".

We find a novel post category that refers to community members sharing their own daily life, such as posting a photo of food or exercise in a gym. This indicates that community members gain support and contribute to the community at the same time. For our research purpose, we do not attempt to distinguish the emotional support seeking category and the sharing daily life category, because the responses to these two categories can be overlapped with compliment or empathy and our text generation module can handle them the same way, as long as they are not seeking factorial information.

Study 1 suggests approximately 18% of the posts did not receive any response, signifying an urgent need to address "post-with-no-response" issue. Community members engagement is a key factor influencing the response amount. Maintaining the active engagement of members is a key challenge for online communities. To address this challenge, some communities employed incentives schemes to motivate member participation, such as gamification [19] or monetary reward [120]. While they are effective to some extend, such solutions come with a relatively high implementation and maintenance cost [87]. With the advances in AI and chatbot technology, some platforms have started experimenting with the use of chatbot service to more closely connect members, for instance, by recommending members to reach out to others [82] or by suggesting further actions based on participation status [81].

Our work takes a different route — builing a chatbot to act as an active member to boost the community engagement level. This result sounds similar to the catfish effect[9] — an active catfish in a group of sardines in a fish tank can stimulus the sardines to be more active and live longer during the transition. CASS is an "active member", which generates diverse responses to reply to other members' posts. The existence of CASS energizes other members to post more responses to the support-seeking post. We speculate this is one plausible explanation to interpret our results in Study 3.

6.2 The Neural-Network-based Chatbot Architecture is a Promising Solution for Building CSCW Systems to Support Online Communities.

---

[9]https://en.wikipedia.org/wiki/Catfish_effect

To develop a successful AI system to provide social support for an online community, a chatbot needs to meet high creterias with respect to its post classification and response generation capabilities. From the original support seeker's perspective, they need authentic and helpful responses in a timely manner, otherwise it may even negatively affect their mood. From the other community member's perspective, if a post's first response has a high quality and high relevancy to the post, they may follow the lead and be more likely to provide their high quality comments to the post thread.

Therefore, CASS needs to be able to generate authentic, diverse, and high-quality responses. However, existing rule-based and IR-based chatbots cannot sufficiently meet such needs [83]. NN-based chatbot can be a promising solution in this case, but it may be affected by the noisy-label issue in training data. We avoid this potential risk with three pre-processing methods. First, our research site is a friendly community. When we clean up the data, we only see some advertisement (e,g., which brand of milk powder is better for the baby) and do not find any offensive or hatred language. Second, we build a CNN-based classification model to filter out informational-seeking posts, as CASS is not designed to generate responses that are relevant to medical information. Thirdly, we design a user interface console with human-in-the-loop AI design, which can be used by a human experimenter to track and monitor each AI-generated response in real-time. From data collection to response generation and publication, we employ rigorous methods to avoid the potential risk of such NN-based chatbot architecture.

However, an online community is fluid and its norms and topics may evolve along with time [110]. AI-generated responses rely on the style and topics in its training dataset. Thus, the NN-based chatbot may need to be re-trained with the latest community data to keep up-to-date with its popular memes and language norms after a period of deployment. This process may sound labor-intensive, but it is much easier than re-training and refining a rule-based chatbot system.

6.3 Human Behavior May Change in a Human-AI Co-Existing Online Community in the Future.

Previous work that studied the human interaction with chatbots often focused only on the impact of a chatbot on an individual user. Little is known about how chatbot system may affect a group or a community. Our results provide evidence that the NN-based AI system can provide timely social support to individual members, as well as boost up other community members' engagement level.

Existing literature suggests that chatbots could play a significant role in boosting the popularity of post content [34, 82]. Seering et al. [96] also find the existence of the chatbot sparked both human-chatbot interactions and human-human interactions. The results of our Study 3 suggest that CASS can help individual members to use their published post as an emotional outlet and interact more with other community members, and ultimately improve their emotional states. These results extend prior work and provide valuable insights into the design of future chatbot systems for online community.

In addition, we also find that CASS can promote other team members to engage more with a post. For example, results from study 3 show that other community members who did not interact with CASS directly have also been affected by CASS. That is, CASS made them become more engaged and more active in replying to those support-seeking posts. We speculate that our research site is a Chinese online community, thus Chinese culture [16] may contribute to this effect. For example, people may not want to be the first person to express their opinion in a public discourse [16]. Being the first commenter seems to require more courage and a stronger desire to express. If the post has already been replied by someone (in this case by the charbot), then the psychological pressure to reply will be less. So when CASS replied to the post, other members may be more willing to participate in the discussion with lower pressure. For HCI and AI researchers, this could be a new

research direction that how different cultures interplay with the user behaviors when people and AI systems co-exist in an online community.

More importantly, with more and more chatbots and AI-based moderation applications being built and adopted into various online communities, we argue that in the future it will be common for online communities to have a human-AI co-existence ecosystem. To support this new paradigm, we need to extend the current human-in-the-loop AI design philosophy [21] and the mixed-initiative design approach [43]. An AI system with high accuracy is not enough. In addition, the AI system should also be designed to cooperate with the primary target users and the other users in the community. An **cooperative AI** system should seamlessly fit into the context and bring some benefits to the community (e.g., supports stressed community members and increases community engagement). And the users who interact with it should feel comfortable to have the AI system in a community, instead of feeling threatened or intimidated. This "Human-AI Collaboration" future [108] is the ultimate research goal that we are aiming for.

6.4 Limitation and Future Directions

One limitation of our work is that the CASS system was designed to handle the emotional-support-seeking posts only, and neglected those informational support seeking posts. It is because the expected responses are different: for emotional support responses, users prefer diverse and lively content, and it is time-sensitive; whereas responses for informational support seeking posts require accurate and factorial answers, thus the diversity is not a critical concern [37]. That being said, NN-based approaches can also help with such task, but it can not be the same seq-to-seq architecture we used for emotional support. Open domain Q&A [111] is a promising NLP technique that can parse a whole textbook and automatically generate answers for any given question about the book. With this technique, chatbot can reply to users' informational support seeking questions by generating answers from medical textbooks, wikipedia or other authentic data sources. The CASS system can still follow the architecture to have 5 Steps and 3 Modules. Only inside the response generation module, CASS needs a hybrid architecture to use the corresponding NLP technique for the incoming posts to generate responses. Theoretically, it is easy to do. But it requires careful design considerations, solid data preparation, and rigorous model training and experimentation work. We will explore this research direction in our future work.

We did not disclose CASS identity before or during the experiment for two main reasons: Firstly, as suggested by previous literature [49, 68, 75, 76, 92], we believe that the identity of chatbot will affect the results of the experiment, attracting too much attention [96] or making the responses biased [49]. For the purpose of this study, we want to avoid that. Secondly, the potential impact of disclosing chatbot's identity on community users' behaviors was not the scope of this study. We are planning to thoroughly examine this topic in our future work. Whether disclosing the chatbot identity is related to an important design question: the appropriateness of languages for a chatbot. In our study, the appropriateness of the generated language is evaluated based on the accuracy and whether it answers the given input. However, we may anticipate that some of the words may be perceived appropriate if they are posted by pregnant human members but inappropriate if they are posted by an AI algorithm. For example, in the Fig 4 example, one pregnant woman complains about her baby causing her some sleeping issues, and our chatbot answered "same here". The answer is totally appropriate if it is from another human member of the community, but some people may perceive it weird as the AI is suggesting "her baby" also caused "her" similar troubles. Thus, there may need a novel dimension for characterizing the generated language's AI-appropriateness.

Another limitation of our chatbot is that in the field experiment, we limited our chatbot's engagement with the poster to only one round. It is not because our chatbot can not deal with multiple round conversations, rather, it is due to the semi-controlled experiment design consideration: if

our chatbot can reply to poster for an arbitrary number of rounds, it introduces more confounding variables to the major behavioral measurements of interest. In the future, we can extend the CASS capability to have multi-round dialogue skills [78], and engage with community members in a more natural manner.

Our work is under the assumption that for these support-seeking posts, the sooner they get answers, the better it is for the support seeker. However, based on Figure 1, it is clear that most posts either quickly receive a response, or never receive a response. The introduction of the proposed chatbot may significantly change such community dynamic. Maybe this change will have some unexpected negative implications. Further study is required to fully understand its impact in a longer-term deployment.

Our deployment in this particular healthcare community may be a limitation for generalizing the empirical findings, for example, those findings specific to the pregnant women community may not apply to other online communities, or the Chinese cultural context may hinder the generalizability to other cultures. However, we believe the scalable and generalizable chatbot architecture can be easily customized for other online community contexts. Thus, we welcome other researchers to join our effort, together we can replicate and extend this study design in other community contexts.

7 CONCLUSION

In this paper, we present a comprehensive research project consisting of three studies, in which we developed and deployed a chatbot system to automatically generate and post responses to emotional support seeking posts in an online health community for pregnant women. Our studies show that the neural-network-based chatbot architecture is a promising solution to build chatbots to generate timely responses for posts with no replies. In addition, we present evidence regarding the chatbot's positive impact on influencing the support seeker's emotional status, and encouraging other community members to be more engaging.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. USENIX Association, Savannah, GA, 265–283. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

[2] Norah Abokhodair, Daisy Yoo, and David W. McDonald. 2015. Dissecting a Social Botnet: Growth, Content and Influence in Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 839–851. https://doi.org/10.1145/2675133.2675208

[3] Mark S Ackerman and David W McDonald. 1996. Answer Garden 2: merging organizational memory with collaborative help. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. 97–105.

[4] Teresa Almeida, Rob Comber, and Madeline Balaam. 2016. HCI and Intimate Care as an Agenda for Change in Women's Health. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2599–2611. https://doi.org/10.1145/2858036.2858187

[5] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics* 4 (2016), 463–476.

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[7] Tawfiq Ammari, Sarita Schoenebeck, and Daniel M. Romero. 2018. Pseudonymous Parents: Comparing Parenting Roles and Identities on the Mommit and Daddit Subreddits. In *Proceedings of the 2018 CHI Conference on Human*

Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174063

[8] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to Me: Foundations for Successful Individual-group Interactions in Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 959–968. https://doi.org/10.1145/1124772.1124916

[9] Saeideh Bakhshi, David A. Shamma, and Eric Gilbert. 2014. Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 965–974. https://doi.org/10.1145/2556288.2557403

[10] Antonina D. Bambina. 2007. *Online Social Support: The Interplay of Social Networks and Computer-Mediated Communication*. Cambria Press.

[11] S Banti, Mauro Mauri, A Oppo, C Borri, C Rambelli, D Ramacciotti, M S Montagnani, V Camilleri, S Cortopassi, Paola Rucci, et al. 2011. From the third month of pregnancy to 1 year postpartum. Prevalence, incidence, recurrence, and new onset of depression. Results from the Perinatal Depression-Research & Screening Unit study. *Comprehensive Psychiatry* 52, 4 (2011), 343–351.

[12] Marguerite Barry, Kevin Doherty, Jose Marcano Belisario, Josip Car, Cecily Morrison, and Gavin Doherty. 2017. mHealth for Maternal Mental Health: Everyday Wisdom in Ethical Design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 2708–2756. https://doi.org/10.1145/3025453.3025918

[13] Sukarna Barua, Md Monirul Islam, Xin Yao, and Kazuyuki Murase. 2012. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering* 26, 2 (2012), 405–425.

[14] Natalya N. Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. 2015. Social Sharing of Emotions on Facebook: Channel Differences, Satisfaction, and Replies. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 154–164. https://doi.org/10.1145/2675133.2675297

[15] Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 827–836.

[16] Michael Harris Bond. 1996. *The handbook of Chinese psychology*. Oxford University Press Hong Kong.

[17] Tone Bratteteig and Guri Verne. 2018. Does AI make PD obsolete? exploring challenges from artificial intelligence to participatory design. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial-Volume 2*. 1–5.

[18] Heloisa Candello, Claudio Pinhanez, and Flavio Figueiredo. 2017. Typefaces and the Perception of Humanness in Natural Language Chatbots. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3476–3487. https://doi.org/10.1145/3025453.3025919

[19] Huseyin Cavusoglu, Zhuolun Li, and Ke-Wei Huang. 2015. Can Gamification Motivate Voluntary Contributions? The Case of StackOverflow Q&A Community. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (CSCW'15 Companion)*. Association for Computing Machinery, New York, NY, USA, 171–174. https://doi.org/10.1145/2685553.2698999

[20] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms Matter: Contrasting Social Support Around Behavior Change in Online Weight Loss Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 666, 14 pages. https://doi.org/10.1145/3173574.3174240

[21] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2382–2393.

[22] Carolyn E Cutrona and Julie A Suhr. 1994. Social support communication in the context of marriage: an analysis of couples' supportive interactions. (1994).

[23] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3267–3276. https://doi.org/10.1145/2470654.2466447

[24] Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and Predicting Postpartum Depression from Shared Facebook Data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 626–638. https://doi.org/10.1145/2531602.2531675

[25] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

[26] Rebecca DerSimonian and Nan Laird. 1986. Meta-analysis in clinical trials. *Controlled clinical trials* 7, 3 (1986), 177–188.

[27] Judith S Donath. 2002. Identity and deception in the virtual community. In *Communities in cyberspace*. Routledge, 37–68.

[28] Patricia Drentea and Jennifer L. Moren-Cross. 2005. Social capital and social support on the web: the case of an internet mother site. *Sociology of Health & Illness* 27, 7 (2005), 920–943. https://doi.org/10.1111/j.1467-9566.2005.00464.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9566.2005.00464.x

[29] Marilyn Evans, Lorie Donelle, and Laurie Hume-Loveland. 2012. Social support and online postpartum depression discussion groups: A content analysis. *Patient Education and Counseling* 87, 3 (2012), 405 – 410. https://doi.org/10.1016/j.pec.2011.09.011

[30] Xiangmin Fan, Daren Chao, Zhan Zhang, Dakuo Wang, Xiaohua Li, and Feng Tian. 2020. Utilization of Self-Diagnosis Health Chatbots in Real-World Settings: Case Study. *Journal of Medical Internet Research* 22 (2020).

[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 (06 2017), e19. https://doi.org/10.2196/mental.7785

[32] R. Stuart Geiger. 2018. The Lives of Bots. *CoRR* abs/1810.09590 (2018). arXiv:1810.09590 http://arxiv.org/abs/1810.09590

[33] Darren Gergle and Desney S Tan. 2014. Experimental research in HCI. In *Ways of Knowing in HCI*. Springer, 191–227.

[34] Zafar Gilani, Reza Farahbakhsh, and Jon Crowcroft. 2017. Do Bots Impact Twitter Activity?. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 781–782. https://doi.org/10.1145/3041021.3054255

[35] Jonathan Grudin. 2008. 1 McGrath and the Behaviors of Groups (BOGs). (2008).

[36] Jonathan Grudin. 2017. From tool to partner: The evolution of human-computer interaction. *Synthesis Lectures on Human-Centered Interaction* 10, 1 (2017), i–183.

[37] Xinning Gui, Yu Chen, Yubo Kou, Katie Pine, and Yunan Chen. 2017. Investigating Support Seeking from Peers for Pregnancy in Online Health Communities. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 50 (Dec. 2017), 19 pages. https://doi.org/10.1145/3134685

[38] Aaron Halfaker and R Stuart Geiger. 2019. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *arXiv preprint arXiv:1909.05189* (2019).

[39] Aaron Halfaker, R Stuart Geiger, and Loren G Terveen. 2014. Snuggle: Designing for efficient socialization and ideological critique. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 311–320.

[40] F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. 2008. Predictors of Answer Quality in Online Q&A Sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 865–874. https://doi.org/10.1145/1357054.1357191

[41] Sepp Hochreiter and Jürgen Schmidhuber. [n.d.]. Long Short-Term Memory. *Neural Computation* 9, 8 ([n. d.]), 1735–1780.

[42] Bree Holtz, Andrew Smock, and David Reyes-Gastelum. 2015. Connected Motherhood: Social Support for Moms and Moms-to-Be on Facebook. *Telemedicine and e-Health* 21, 5 (2015), 415–421. https://doi.org/10.1089/tmj.2014.0118 arXiv:https://doi.org/10.1089/tmj.2014.0118 PMID: 25665177.

[43] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[44] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch your heart: a tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[45] Jina Huh and Mark S. Ackerman. 2012. Collaborative Help in Chronic Disease Management: Supporting Individualized Problems. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 853–862. https://doi.org/10.1145/2145204.2145331

[46] IBM. 2017. Tone Analyzer for Customer Engagement. https://www.ibm.com/blogs/bluemix/2017/04/tone-analyzer-customer-engagement-7-new-tones-help-understand-customers-feeling/. Accessed: 2019-04-03.

[47] IBM. 2018. IBM Watson Assistant. Retrieved: 2018-12-10. https://assistant-us-south.watsonplatform.net/us-south/b3a5bd9b-9ea9-4be8-9ec7-145f04f69453/home.

[48] N Ishibe, M Albitar, I. B. Jilani, L. R. Goldin, G. E. Marti, and N. E. Caporaso. 2006. Incidence of postpartum depression in Trabzon province and risk factors at gestation. *Turk psikiyatri dergisi = Turkish journal of psychiatry* 17, 4 (2006), 243–251.

[49] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception That Profile Text Was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 239, 13 pages.

https://doi.org/10.1145/3290605.3300469

[50] Isaac L. Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at Home on the Range: Peer Production and the Urban/Rural Divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 13–25. https://doi.org/10.1145/2858036.2858123

[51] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[52] Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, and Alexander M. Rush. 2018. OpenNMT: Neural Machine Translation Toolkit. (2018).

[53] Robert Kraut, Moira Burke, John Riedl, and Paul Resnick. 2010. Dealing with newcomers. *Evidencebased Social Design Mining the Social Sciences to Build Online Communities* 1 (2010), 42.

[54] Robert E Kraut and Paul Resnick. 2011. Encouraging contribution to online communities. *Building successful online communities: Evidence-based social design* (2011), 21–76.

[55] Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design.* Mit Press.

[56] Alex Krizhevsky, I. Sutskever, and G. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems* 25, 2 (2012).

[57] Neha Kumar and Richard J. Anderson. 2015. Mobile Phones for Maternal Health in Rural India. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 427–436. https://doi.org/10.1145/2702123.2702258

[58] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: A Chatbot for Self-Compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 702, 13 pages. https://doi.org/10.1145/3290605.3300932

[59] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3365–3376.

[60] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[61] Uichin Lee, Jihyoung Kim, Eunhee Yi, Juyup Sung, and Mario Gerla. 2013. Analyzing crowd workers in mobile pay-for-answer q&a. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 533–542.

[62] Bronwyn Leigh and Jeannette Milgrom. 2008. Risk factors for antenatal depression, postnatal depression and parenting stress. *BMC Psychiatry* 8 (2008), 24 – 24.

[63] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541* (2016).

[64] Q Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N Sadat Shami, and Werner Geyer. 2018. All Work and No Play?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[65] John Logie, Joseph Weinberg, F. Maxwell Harper, and Joseph A. Konstan. 2011. Asked and Answered: On Qualities and Quantities of Answers in Online Q&A Sites. In *The Social Mobile Web*.

[66] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. "Could You Define That in Bot Terms"? Requesting, Creating and Using Bots on Reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3488–3500. https://doi.org/10.1145/3025453.3025830

[67] Nurul Lubis, Sakriani Sakti, Graham Neubig, Koichiro Yoshino, Tomoki Toda, and Satoshi Nakamura. 2015. A study of social-affective communication: Automatic prediction of emotion triggers and responses in television talk shows. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 777–783.

[68] Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu. 2019. Machines Versus Humans: The Impact of AI Chatbot Disclosure on Customer Purchases. *Luo, X, Tong S, Fang Z, Qu* 2019 (2019).

[69] Minh Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *Computer Science* (2015).

[70] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest q&a site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2857–2866.

[71] Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. 2015. Collective Sensemaking in Online Health Forums. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3217–3226. https://doi.org/10.1145/2702123.2702566

[72] Elijah Mayfield, Miaomiao Wen, Mitch Golant, and Carolyn Penstein Rosé. 2012. Discovering Habits of Effective Online Support Group Chatrooms. In *Proceedings of the 17th ACM International Conference on Supporting Group Work (GROUP*

'12). Association for Computing Machinery, New York, NY, USA, 263–272. https://doi.org/10.1145/2389176.2389216

[73] Joseph E McGrath. 1991. Time, interaction, and performance (TIP) A Theory of Groups. *Small group research* 22, 2 (1991), 147–174.

[74] Microsoft. 2018. Cognitive Services: Language Understanding (LUIS). Retrieved: 2018-12-10. https://www.luis.ai/home.

[75] Robert R. Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen Matthew Schueller. 2018. Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions. In *Journal of medical Internet research.*

[76] Alessandro Murgia, Daan Janssens, Serge Demeyer, and Bogdan Vasilescu. 2016. Among the Machines: Human-Bot Interaction on Social Q&A Websites. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 1272–1279. https://doi.org/10.1145/2851581.2892311

[77] Kevin Kyung Nam, Mark S Ackerman, and Lada A Adamic. 2009. Questions in, knowledge in? A study of Naver's question answering community. In *Proceedings of the SIGCHI conference on human factors in computing systems.* 779–788.

[78] QIU Nan and WANG Haofen. 2018. State machine based context-sensitive system for managing multi-round dialog. US Patent App. 15/694,917.

[79] 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)*. ACM, New York, NY, USA, 1470–1484. https://doi.org/10.1145/2998181.2998349

[80] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics.* 311–318.

[81] Zhenhui Peng, Taewook Kim, and Xiaojuan Ma. 2019. GremoBot: Exploring Emotion Regulation in Group Chat. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing (CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 335–340. https://doi.org/10.1145/3311957.3359472

[82] Zhenhui Peng and Xiaojuan Ma. 2019. Exploring how software developers work with mention bot in GitHub. *CCF Transactions on Pervasive Computing and Interaction* (05 Sep 2019). https://doi.org/10.1007/s42486-019-00013-2

[83] Zhenhui Peng and Xiaojuan Ma. 2019. A survey on construction and enhancement methods in service chatbots design. *CCF Transactions on Pervasive Computing and Interaction* 1, 3 (2019), 204–223.

[84] Trevor Perrier, Nicola Dell, Brian DeRenzi, Richard Anderson, John Kinuthia, Jennifer Unger, and Grace John-Stewart. 2015. Engaging Pregnant Women in Kenya with a Hybrid Computer-Human SMS Communication System. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1429–1438. https://doi.org/10.1145/2702123.2702124

[85] Dominic Pettman. 2009. Love in the Time of Tamagotchi. *Theory, culture & society* 26, 2-3 (2009), 189–208.

[86] Yuqing Ren, F Maxwell Harper, Sara Drenner, Loren Terveen, Sara Kiesler, John Riedl, and Robert E Kraut. 2012. BUILDING MEMBER ATTACHMENT IN ONLINE COMMUNITIES: APPLYING THEORIES OF GROUP IDENTITY AND INTERPERSONAL BONDS. *MIS Quarterly* 36, 3 (2012), 841–864.

[87] Yuqing Ren, Robert Kraut, Sara Kiesler, and Paul Resnick. 2012. Encouraging commitment in online communities. *Building successful online communities: Evidence-based social design* (2012), 77–124.

[88] Paul Resnick, Joseph Konstan, Yan Chen, and Robert E Kraut. 2012. Starting new online communities. *Building successful online communities: Evidence-based social design* 231 (2012).

[89] Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Conference on Empirical Methods in Natural Language Processing.*

[90] Lionel P Robert, Casey Pierce, Liz Marquis, Sangmi Kim, and Rasha Alahmad. 2020. Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Human–Computer Interaction* (2020), 1–31.

[91] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[92] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. 2016. Botivist: Calling Volunteers to Action Using Online Bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 813–822. https://doi.org/10.1145/2818048.2819985

[93] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 315, 14 pages. https://doi.org/10.1145/3173574.3173889

[94] 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, Article 398, 15 pages. https://doi.org/10.1145/3173574.3173972

[95] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. 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, 111–125. https://doi.org/10.1145/2998181.2998277

[96] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376708

[97] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.

[98] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the Effects of Embodiment for a Group Facilitation Agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 391, 13 pages. https://doi.org/10.1145/3173574.3173965

[99] Eva Sharma and Munmun De Choudhury. 2018. Mental Health Support and Its Relationship to Linguistic Accommodation in Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 641, 13 pages. https://doi.org/10.1145/3173574.3174215

[100] Katie A Siek, Gillian R Hayes, Mark W Newman, and John C Tang. 2014. Field deployments: Knowing from using in context. In *Ways of Knowing in HCI*. Springer, 119–142.

[101] Ming Tan, Dakuo Wang, Yupeng Gao, Haoyu Wang, Saloni Potdar, Xiaoxiao Guo, Shiyu Chang, and Mo Yu. 2019. Context-Aware Conversation Thread Detection in Multi-Party Chat. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6457–6462.

[102] Loren Terveen, Joseph A Konstan, and Cliff Lampe. 2014. Study, Build, Repeat: Using Online Communities as a Research Platform. In *Ways of Knowing in HCI*. Springer, 95–117.

[103] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding Chatbot-mediated Task Management. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 58, 6 pages. https://doi.org/10.1145/3173574.3173632

[104] Cornelia F. van Uden-Kraan, Constance H. C. Drossaert, Erik Taal, Bret R. Shaw, Erwin R. Seydel, and Mart A. F. J. van de Laar. 2008. Empowering Processes and Outcomes of Participation in Online Support Groups for Patients With Breast Cancer, Arthritis, or Fibromyalgia. *Qualitative Health Research* 18, 3 (2008), 405–417. https://doi.org/10.1177/1049732307313429 arXiv:https://doi.org/10.1177/1049732307313429 PMID: 18235163.

[105] Joseph Walther and Saxon Boyd. 2002. Attraction to computer-mediated social support. *Communication Technology and Society: Audience Adoption and Uses* (01 2002), 153–188.

[106] Dakuo Wang, Josh Andres, Justin Weisz, Erick Oduor, and Casey Dugan. 2021. AutoDS: Towards Human-Centered Automation of Data Science. In *Proceedings of the CHI 2021*.

[107] Dakuo Wang, Youyang Hou, Lin Luo, and Yingxin Pan. 2016. Answerer engagement in an enterprise social question & answering system. *IConference 2016 Proceedings* (2016).

[108] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. Brilliant AI Doctor in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment. In *Proceedings of the CHI 2021*.

[109] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *To appear in Computer Supported Cooperative Work (CSCW)* (2019).

[110] Minjuan Wang, Christina Sierra, and Terre Folger. 2003. Building a dynamic online learning community among adult learners. *Educational Media International* 40, 1-2 (2003), 49–62.

[111] Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R 3: Reinforced ranker-reader for open-domain question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[112] Yi-Chia Wang, Mahesh Joshi, William W Cohen, and Carolyn Penstein Rosé. 2008. Recovering Implicit Thread Structure in Newsgroup Style Conversations.. In *ICWSM*.

[113] Yi-Chia Wang, Robert Kraut, and John Levine. 2015. Eliciting and Receiving Online Support: Using Computer-Aided Content Analysis to Examine the Dynamics of Online Social Support. *Journal of medical Internet research* 17 (04 2015), e99. https://doi.org/10.2196/jmir.3558

[114] Yi-Chia Wang, Robert Kraut, and John M. Levine. 2012. To Stay or Leave?: The Relationship of Emotional and Informational Support to Commitment in Online Health Support Groups. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 833–842. https://doi.org/10.1145/2145204.2145329

[115] Joseph Weizenbaum. 1966. ELIZA&Mdash;a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM* 9, 1 (Jan. 1966), 36–45. https://doi.org/10.1145/365153.365168

[116] Miaomiao Wen and Carolyn Penstein Rosé. 2012. Understanding participant behavior trajectories in online health support groups using automatic extraction methods. In *Proceedings of the 17th ACM international conference on Supporting group work*. 179–188.

[117] Marty J Wolf, K Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications. *ACM SIGCAS Computers and Society* 47, 3 (2017), 54–64.

[118] Marisol Wong-Villacres, Neha Kumar, and Betsy DiSalvo. 2019. The Parenting Actor-Network of Latino Immigrants in the United States. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300914

[119] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[120] Ziming Wu and Xiaojuan Ma. 2017. Money as a Social Currency to Manage Group Dynamics: Red Packet Gifting in Chinese Online Communities. In *Chi Conference Extended Abstracts on Human Factors in Computing Systems*.

[121] Mi Xiang, Zhiruo Zhang, and Huigang Liang. 2020. Sedentary behavior relates to mental distress of pregnant women differently across trimesters: An observational study in China. *Journal of affective disorders* 260 (2020), 187–193.

[122] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3506–3510.

[123] Ying Xu, Dakuo Wang, Penelope Collins, Hyelim Lee, and Mark Warschauer. [n.d.]. Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner. *Computers & Education* 161 ([n. d.]), 104059.

[124] Diyi Yang, Robert Kraut, and John M. Levine. 2017. Commitment of Newcomers and Old-Timers to Online Health Support Communities. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6363–6375. https://doi.org/10.1145/3025453.3026008

[125] Diyi Yang, Robert E. Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 344, 14 pages. https://doi.org/10.1145/3290605.3300574

[126] Diyi Yang, Zheng Yao, and Robert Kraut. 2017. Self-disclosure and channel difference in online health support groups. In *Eleventh International AAAI Conference on Web and Social Media*.

[127] Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The Channel Matters: Self-disclosure, Reciprocity and Social Support in Online Cancer Support Groups. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 31, 15 pages. https://doi.org/10.1145/3290605.3300261

[128] Esra Yazici, Tulay Sati Kirkan, Puren Akcali Aslan, Nazan Aydin, and A Yazici. 2015. Untreated depression in the first trimester of pregnancy leads to postpartum depression: High rates from a natural follow-up study. *Neuropsychiatric Disease and Treatment* 11 (02 2015), 405–11. https://doi.org/10.2147/ndt.s77194

[129] Zi Yin, Keng-hao Chang, and Ruofei Zhang. 2017. DeepProbe: Information Directed Sequence Understanding and Chatbot Design via Recurrent Neural Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 2131–2139. https://doi.org/10.1145/3097983.3098148

[130] Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat Through Collaborative Tagging and Summarization. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 196 (Nov. 2018), 27 pages. https://doi.org/10.1145/3274465

[131] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.

[132] Haiyi Zhu, Robert E Kraut, and Aniket Kittur. 2014. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 281–290.

[133] 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 (2018), 1–23.

[134] Haiyi Zhu, Amy Zhang, Jiping He, Robert E Kraut, and Aniket Kittur. 2013. Effects of peer feedback on contribution: a field experiment in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2253–2262.