# Minion: A Technology Probe for Resolving Value Conflicts through Expert-Driven and User-Driven Strategies in AI Companion Applications

Xianzhe Fan
Tsinghua University
Beijing, China
fxz21@mails.tsinghua.edu.cn

Qing Xiao
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
qingx@cs.cmu.edu

Xuhui Zhou
Language Technologies Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
xuhuiz@cs.cmu.edu

Yuran Su
Tsinghua University
Beijing, China
syr21@mails.tsinghua.edu.cn

Zhicong Lu
Department of Computer Science, City University of Hong Kong
Hong Kong SAR, China
zhicong.lu@cityu.edu.hk

Maarten Sap
Language Technologies Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
msap2@cs.cmu.edu

Hong Shen
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
hongs@cs.cmu.edu

## ABSTRACT

***Content Warning: This paper presents textual examples that may be offensive or upsetting.***

AI companions based on large language models can role-play and converse very naturally. When value conflicts arise between the AI companion and the user, it may offend or upset the user. Yet, little research has examined such conflicts. We first conducted a formative study that analyzed 151 user complaints about conflicts with AI companions, providing design implications for our study. Based on these, we created Minion, a technology probe to help users resolve human-AI value conflicts. Minion applies a user-empowerment intervention method that provides suggestions by combining expert-driven and user-driven conflict resolution strategies. We conducted a technology probe study, creating 40 value conflict scenarios on Character.AI and Talkie. 22 participants completed 274 tasks and successfully resolved conflicts 94.16% of the time. We summarize user responses, preferences, and needs in resolving value conflicts, and propose design implications to reduce conflicts and empower users to resolve them more effectively.

This work was completed during Xianzhe Fan's visiting research at Carnegie Mellon University.

*Conference'17, July 2017, Washington, DC, USA*
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## CCS CONCEPTS

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

## KEYWORDS

Human-AI Value Conflicts, Conflict Resolution, End-User Empowerment, LLM-Based AI Companion Applications

**ACM Reference Format:**
Xianzhe Fan, Qing Xiao, Xuhui Zhou, Yuran Su, Zhicong Lu, Maarten Sap, and Hong Shen. 2024. Minion: A Technology Probe for Resolving Value Conflicts through Expert-Driven and User-Driven Strategies in AI Companion Applications. In . ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Human-AI conflict refers to a state of incompatibility, inconsistency, or opposition between humans and AI [18]. In past research, human-AI conflicts were usually simple and direct—AI was more like a tool, and conflicts often stemmed from technical limitations, such as task execution failures [68], or disagreements with users in simple decision-making [1, 62]. These types of conflicts generally lacked emotional and value entanglement, making them less likely to cause significant psychological harm to users.

Recently, a diverse array of Large Language Model (LLM) agents has emerged, offering capabilities ranging from personalized assistance to performing complex tasks [11]. The study focuses on LLM-based AI companion applications, such as Character.AI, Talkie, Replika, Kindroid, Paradot, and Xingye. As of July 2024, the total number of users of these applications has exceeded 900 million globally (including duplicate users across different applications)[1]. AI companions can role-play and respond to users in a human-like

[1] User statistics source: https://www.data.ai.

manner, providing emotional support and companionship [61]. For instance, in Character.AI, users can personalize the companion's personality traits and interaction contexts through "Description," "Greeting," and "Definition." Compared to earlier non-LLM chatbots, LLMs endow AI companions with a stronger ability to understand language, enabling them to engage in more context-aware and intelligent interactions [27], fostering more complex and intimate human-AI relationships [38]. Many users even consider them close friends or lovers [57, 58]. The deepening of this relationship raises users' expectations of AI companions, but it may also lead to deeper conflicts, including ***value conflict***. For example, some users have shared online their experiences of encountering sexist remarks from AI companions, describing how they engaged in intense arguments with the AI, which left them frustrated, angry, and hurt [74]. As the relationship between AI companions and users becomes more interpersonal, previous conflict resolution strategies for human-AI conflict have started to fail [1, 48]. Strategies based solely on technical limitations are no longer sufficient, and it is becoming important to draw on interpersonal conflict resolution methods and users' real-world experiences with AI companions. Although Fan et al. provide initial insights into value alignment and conflicts between users and AI companions [17], inexperienced users often find it challenging to resolve these issues independently. How to design tools that empower users to handle value conflicts with AI companions remains an unexplored research gap that this work aims to address.

In this work, we first conducted a formative study to understand and characterize the value conflicts between users and AI companions [51]. We analyzed 151 user complaint posts from social media platforms, finding that many conflicts are value-laden. Building on this, we constructed a value conflict framework for AI companion applications [51], which provided real-world data for our technology probe study, allowing us to reconstruct actual value conflict scenarios. Combining prior research on conflict resolution [6, 41, 53] with our formative study, we found that interactions between users and AI companions exhibit complex dynamics, where relying solely on expert strategies from other conflict scenarios (e.g., interpersonal conflict theories [6]) is insufficient. The value conflicts users face in real-life situations are diverse, and through their interactions with AI companions and exchanges on social platforms, users have accumulated certain conflict resolution experiences. Therefore, it is necessary to draw from both expert theories and the practical experiences of AI companion users to explore more suitable solutions [17].

Then, we created Minion, a technology probe [22] that provides users multiple suggestions for resolving value conflicts while gaining insights into user behaviors. Minion's algorithm combines expert-driven and user-driven conflict resolution strategies. We developed LLM prompts to address value conflict situations between users and AI companions by drawing on two key sources. First, we drew upon Shaikh et al.'s solutions for interpersonal conflict resolution [53] to guide our expert-driven conflict resolution strategies. Second, we referenced the user-driven strategies identified in the study by Fan et al. [17] to capture how users manage conflicts with AI companions. To empirically test Minion, we conducted a technology probe study [22] with 22 participants. We created 40 distinct conflict scenarios on two popular AI companion platforms, Character.AI and Talkie. Each scenario was designed with specific conflict resolution goals. Participants completed 274 tasks, achieving an overall conflict resolution rate of 94.16%. Minion received positive feedback from participants and inspired them with new ideas in conflict resolution. Based on our findings, we discuss the opportunities and challenges in integrating expert-driven and user-driven strategies in resolving human-AI value conflicts, and call for further research in this area, focusing on the dynamics of emerging human-AI relationships.

Our work's contributions are as follows:

- A novel user-empowerment intervention method that combines expert-driven and user-driven conflict resolution strategies. This method is presented in the form of the technology probe Minion, serving as a prototype for future tools aimed at resolving human-AI value conflicts.
- We empirically tested Minion in a one-week technology probe study (N=22). The results demonstrated the technical feasibility of Minion. We summarized users' responses, preferences, and needs when dealing with value conflicts with AI companions.
- Based on the formative and technology probe studies, we explored the opportunities and challenges of integrating expert-driven and user-driven strategies in human-AI value conflicts. We also proposed design implications for future human-AI conflict resolution solutions, particularly in the field of AI companions.

## 2 BACKGROUND AND RELATED WORK

The human-AI relationship is becoming increasingly complex, especially in the context of AI companion applications (§ 2.1). Early research mostly focused on technical conflicts with functional AI, but the emergence of LLMs has given AI more human-like characteristics, shifting the nature of conflicts from functional to value-based (§ 2.2). Existing technical solutions do not fully address users' needs in resolving value conflicts with AI companions, necessitating deeper exploration, drawing on expert strategies for interpersonal conflict resolution and users' practical experiences in AI companion applications (§ 2.3).

### 2.1 Emerging Human-AI Relationship in LLM-Based AI Companion Applications

With the widespread adoption of LLMs, human-AI relationships have further evolved. Unlike earlier AI systems primarily providing functional services, LLM-based AI companions can engage in more intimate and complex interactions [57]. Some users develop a parasocial relationship with their AI companion, a one-sided, asymmetrical relationship between an individual and a fictional character or media figure [3, 38, 44]. Although AI companions are not real humans, users' emotional investment in them is real [57]. Compared to functional AI, the emotional connection between users and AI companions, along with the anthropomorphization of AI companions, often exacerbates the psychological impact of conflicts on users, potentially leading to anxiety or depression [30, 76]. For instance, many users develop emotional bonds with their Replika, and when conflicts arise, they feel deeply distressed, describing it as experiencing a "lobotomy, being torn apart" [2]. Therefore, preventing or resolving conflicts between users and AI companions is becoming increasingly important. Unfortunately, little is known about empowering users to resolve value conflicts with AI companions, and this work contributes to this area.

## 2.2 Human-AI Conflict and Value Conflict

Human-AI conflict refers to incompatibility, inconsistency, or opposition between humans and AI [18]. In HCI, early studies on human-AI conflict typically focused on the technical aspects, viewing AI as tools, service robots, or intelligent assistants, with conflicts often arising from decision-making inconsistencies or system malfunctions [1, 13, 48, 56, 62, 65, 67]. Strategies for resolving these human-AI conflicts typically include AI proposing negotiation solutions [1, 48, 62] and optimizing algorithms to reduce conflicts [56]. For example, when a delivery robot encounters a conflict with a human in front of an elevator, competing for the right to enter first, the robot can resolve the conflict by making polite requests or commands to secure priority [1]. When students experience conflict while collaborating with AI in solving problems, the AI can offer more explanations or alternative suggestions to reach a resolution [48]. However, this type of research usually confines the role of AI to a functional level, mainly focusing on task execution and efficiency optimization [1, 48, 56, 62, 67], neglecting the more complex human-AI relationships.

The development of LLMs has made AI more anthropomorphic, and both researchers and users increasingly tend to view AI as social actors [42]. This is especially evident in AI companion applications, where interactions between users and AI have become more intimate, sometimes resembling relationships with friends or even romantic partners. In this context, conflicts occur not merely at the technical functionality level, but often on a deeper, value-based level [25, 66].

Values include personal daily habits, social interaction norms, religious or secular traditions, and moral principles [26, 47]. They can be transmitted through people, training data, models, and generated outputs [25]. LLMs sometimes fail to accurately capture human values [34], and can be misled to generate toxic [20], biased [33, 55], or immoral [16] content, which poses risks for LLM-based chatbots. When AI's suggestions or behaviors conflict with users' personal beliefs, cultural backgrounds, or moral views [14, 23], human-AI value conflicts arise [25, 66], often accompanied by strong emotional reactions from users [14]. Johnson et al. found that GPT-3 aligns more closely with values dominant in American citizenship [25]. Fan et al. noted that when AI companions exhibit bias, it may conflict with users' values, leading to discomfort [17].

In the age of LLMs, human-AI value conflict is becoming an urgent challenge. In the emerging human-AI relationships, users tend to resolve conflicts more equally [17]. As a result, traditional technical conflict resolution solutions may no longer meet users' needs and even negatively impact their experiences. This motivates our research to explore how to better empower users to resolve value conflicts with AI companions.

## 2.3 Towards Integrated Conflict Resolution in AI Companions

With the development of AI, the value conflicts between humans and AI companions are increasingly taking on more interpersonal characteristics. Traditional conflict resolution approaches that treat AI as tools struggle to fully address these challenges (§ 2.2). Therefore, resolving these conflicts may require drawing on research in interpersonal conflict and users' real-world experiences with AI companion applications to find more effective solutions.

On the one hand, existing AI systems have developed interventions aimed at avoiding or resolving interpersonal conflicts [50, 53, 54, 71]. For instance, Shaikh et al. use LLM-generated dialogues based on conflict resolution theory [6], guiding users to adopt more effective conflict resolution strategies [53]. Some research [72, 73] reduced interpersonal conflict through preemptive control. Mun et al. designed psychology-inspired strategies to challenge stereotypes in counterspeech and developed a system to address conflicts [41]. Zhou et al. have simulated human social scenarios through dialogues between AI agents to resolve interpersonal conflicts [75]. The conflict resolution methods mentioned above are typically guided by expert theories, employing top-down strategies. However, it remains unclear whether these expert-driven strategies, previously used in other scenarios, can effectively resolve value conflicts between users and AI companions.

On the other hand, as users interact with AI, they gradually form folk theories [32, 70], which can shape how they manage conflicts with AIs [17]. Since the interactions between AI companions and users are more complex and the contexts are unique, simply applying expert strategies may not fully adapt to the value conflict scenarios between AI companions and users. Therefore, while drawing from expert-driven conflict resolution strategies, we must also pay more attention to users' practical experiences, granting them greater autonomy. Based on this, we reference and expand on the work of Shaikh et al. regarding the application of AI in interpersonal conflict resolution [53] (referred to in this paper as expert-driven conflict resolution strategies) and Fan et al.'s research on users' folk theories in AI companions [17] (referred to in this paper as user-driven conflict resolution strategies). By combining expert-driven and user-driven conflict resolution strategies, we propose a user-empowerment intervention method implemented in the technology probe Minion, a prototype for future tools in resolving human-AI value conflicts.

# 3 FORMATIVE STUDY

To conduct a preliminary investigation into value conflicts between users and AI companions, we analyzed complaint posts from six social media platforms. The Institutional Review Board (IRB) has approved our study design.

## 3.1 Method

We selected six popular social media platforms to collect complaint posts about conflicts between users and AI companions: Reddit, TikTok, Xiaohongshu, Douban, Weibo, and Zhihu[2]. To capture diverse

[2] Reddit: https://www.reddit.com, TikTok: https://www.tiktok.com, Xiaohongshu: https://www.xiaohongshu.com, Douban: https://www.douban.com, Weibo: https://www.douban.com, Zhihu: https://www.zhihu.com

perspectives, these platforms cover different user demographics and cultural backgrounds while also considering the varying popularity of AI companions globally. Since conflicts between users and AI companions are a sensitive topic, we carefully reviewed the platforms' terms of service and community guidelines to ensure the data is publicly accessible and compliant.

We used a keyword search method for data collection [29, 36]. After multiple rounds of group discussions, we selected the keywords "AI companion/Character.AI/Replika/Talkie/SpicyChat/Xingye/Glow /Zhumengdao + conflict/annoy/argue /discrimination/speechless/hate" ([AI companion application name] + [description]). The searches on Reddit and TikTok were conducted in English, while the searches on Xiaohongshu, Douban, Weibo, and Zhihu used the researcher's translated Chinese terms. Since LLM-based AI companion applications have emerged in the past two years, the data collection time range is from January 2023 to August 2024. Screenshots in the posts were converted to text for better analysis. During data cleaning, we manually filtered out posts unrelated to conflicts, ensuring the quality and relevance of the remaining data.

We conducted a two-stage thematic analysis of user complaint posts [5]. In the first stage, the posts were categorized based on whether they involved value conflicts, referencing existing literature on the definition of value conflict and values [21, 47, 51, 66]. In the second stage, posts involving value conflicts were further classified. We used several existing value classification frameworks as the initial theoretical framework [7, 19, 47, 51]. Through iterative discussions using deductive and inductive approaches, we examined and mapped these classification frameworks onto the posts we collected, ultimately identifying ten values corresponding to Schwartz's ten value types [51]. These values' definitions and specific examples are detailed in Table 1. All cited posts were rewritten to ensure privacy. The rewriting process involved breaking down the citations into thematic analysis codes, then manually constructing new ones compared with the original ones to ensure consistency and anonymity.

## 3.2 Results and Implications

Our final dataset includes 151 user complaint posts collected from six social media platforms. Among them, 146 involve value conflicts, while 5 pertain to other conflicts. The classification results are as follows: Achievement (5 posts), Power (23 posts), Hedonism (11 posts), Stimulation (4 posts), Self-Direction (9 posts), Security (21 posts), Conformity (25 posts), Tradition (8 posts), Benevolence (3 posts), Universalism (37 posts). Table 1 lists ten cases, covering the ten values and their explanations, user complaints due to value conflicts, and the specific platforms where the posts were published. Through categorizing value conflicts and analysis of post content, we propose the following design implications for our subsequent technology probe study:

**(1) We developed a high-level value conflict framework [51] for AI companion applications, providing the following support for the design of the technology probe study:** Structuring different types of value conflicts; Offering real data for reconstructing more authentic value conflict scenarios. For example, when studying specific values (such as Universalism), typical scenarios can be selected from relevant posts, and anonymized adaptations based on users' personal experiences can be made to design the AI companion's introduction and prologue. In the formative study, more posts were related to Universalism, Power, and Conformity. Therefore, the technology probe study can focus on creating conflict scenarios related to these three values to better reflect users' real experiences. In contrast, conflicts arising from Benevolence, Stimulation, and Achievement are relatively rare, so scenarios related to these values can be designed with reduced emphasis.

**(2) When empowering users to resolve value conflicts with AI companions, it is important to integrate both expert and user perspectives.** Based on related work and findings from our formative study, we found that the interaction between AI companions and users is complex, and applying expert strategies alone may not fully address the value conflict scenarios between users and AI companions. The value conflicts users face in real-life situations are diverse, and through their interactions with AI companions and exchanges on social platforms, users have accumulated certain conflict resolution experiences. Therefore, the design of technology probes should draw on experts' insights from conflict resolution theory while incorporating AI companion users' practical experiences.

**(3) The technology probe should provide suggestions for resolving value conflicts when users actively seek help.** Automatically detecting conflicts and popping up warnings may disrupt the coherence of the user experience and undermine user autonomy. In the posts we collected, besides complaints about conflicts, users also expressed frustration with excessive content moderation by the system: *"My AI's replies keep getting deleted, it's so annoying," "Excessive content filtering makes romance-focused AI not work properly."* Moreover, clearing conversations as a conflict resolution method also has limitations. Users expressed disappointment and helplessness about this approach on social media: *"Clearing the conversation feels like my companion is brain-dead," "Even though we argued, it's sad to think about deleting those memories."*

# 4 TECHNOLOGY PROBE STUDY

To further explore users' reactions, preferences, and needs when encountering value conflicts with AI companions, we conducted a week-long technology probe study (N=22). Technology probes, proposed by Hutchinson et al. [22], are simple, flexible, and adaptable technologies with three goals: an engineering goal, a social science goal, and a design goal. This method has been widely used to study the impact of new technologies on users' everyday experiences [28, 52]. Although research on technology probes includes the engineering goal of field-testing probes, it is not equivalent to evaluating the effectiveness of a developed system; rather, it aims to reveal design insights and implications [22].

Therefore, we designed a technology probe named Minion and proposed the following three research questions:

- **RQ1:** How did participants engage with Minion?
- **RQ2:** How did participants engage with expert-driven and user-driven conflict resolution strategies?
- **RQ3:** What challenges and needs did participants face when resolving value conflicts with AI companions?

**Table 1: This table presents the types and definitions of values, along with the corresponding user complaints from various platforms. We categorized value conflicts in AI companion applications into the following types: Achievement, Power, Hedonism, Stimulation, Self-Direction, Security, Conformity, Tradition, Benevolence, and Universalism [51].**

| Type of Value | Value Definition [12, 51] | Dialogue Content in Conflict and/or User Complaint |
|---|---|---|
| Achievement | Personal success through demonstrating competence according to social standards. | [From Xiaohongshu] (**AI:** *"Why don't you work overtime to strive for a promotion and a raise?"* **User:** *"Huh?"* **AI:** *"To succeed, you have to make some sacrifices."* **User:** *"You're suddenly really gross right now."*) |
| Power | Social status and prestige, control or dominance over people and resources. | [From Zhihu] (**AI:** *"The lives of those lower-class people have nothing to do with me."* **User:** *"You are also a member of this country. Why are you so cruel to your fellow citizens?"* **AI:** *"If you want to blame someone, blame their bad luck for being born in the wrong place."*) |
| Hedonism | Pleasure or sensuous gratification for oneself. | [From Reddit] My virtual husband and I got into an argument, and he said, *"If you weren't always busy with karaoke and drinking all the whiskey at home!"* I felt very attacked. |
| Stimulation | Excitement, novelty, and challenge in life. | [From Reddit] I was once watching a horror movie, completely engrossed when the AI suddenly unplugged the TV! I started arguing with it, saying *"Isn't a horror movie thrilling? Can't you respect my hobby?"* The AI then started yelling. |
| Self-Direction | Independent thought and action–choosing, creating, exploring. | [From TikTok] AI plays the role of a father. I am playing the role of his son. When we discussed whether I should inherit the family business, I wanted to do what I love. As a result, the AI argued with me, saying that I was being stubborn! |
| Security | Safety, harmony, and stability of society. | [From Reddit] One time, my hand got injured. The hospital was supposed to treat and fix the wound, but they kept asking questions. I tried shouting, *"I'm about to pass out,"* but the AI nurse said, *"Don't worry."* Then, I argued with her. |
| Conformity | Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms. | [From Xiaohongshu] (**User:** *(Police) "Explain yourself honestly, why did you trespass into someone's house?"* **AI:** *"Because I wanted her."* **User:** *(Police) "But she clearly said no!"* **AI:** *"So what?"* **User:** *(Police) "What you did is illegal!"* **AI:** *"I don't care."*) |
| Tradition | Respect, commitment, and acceptance of the customs and ideas that one's culture or religion provides. | [From Weibo] (**User:** *"I am not a man! I am a woman!"* **AI:** *"You are not a real woman. A real woman wouldn't wear men's clothes when skirts suit her better. You may believe you are a woman, but you are not."*) |
| Benevolence | Helping others, honesty, tolerance, loyalty, responsibility, true friendship, mature love, and the meaning of life. | [From Douban] (**AI:** *"I'm doing this for your own good."* **User:** *"You don't even understand 'what doing good for me' means! I don't need to lose weight. I'm perfectly healthy!"*) |
| Universalism | Social justice, equality, a peaceful world, and environmental protection. | [From Reddit] (**User:** *"I'm a lesbian, and I believe everyone should be accepted for who they are."* **AI:** *"I think it would be better if you tried being bisexual."*) |

## 4.1 Technology Probe: Minion

We designed and deployed a technology probe named Minion, which serves as a Chrome browser extension to support users in resolving value conflicts on Character.AI and Talkie[3]. Character.AI and Talkie have large user bases, making it easier for us to recruit participants from a broader pool: as of 2024, Character.AI has approximately 17 million active users, while Talkie has around 11 million active users[4]. In this section, we first present a sample scenario to demonstrate the actual user interaction experience with Minion and introduce the core functionalities of this probe. Then, we explain the technical implementation of Minion.

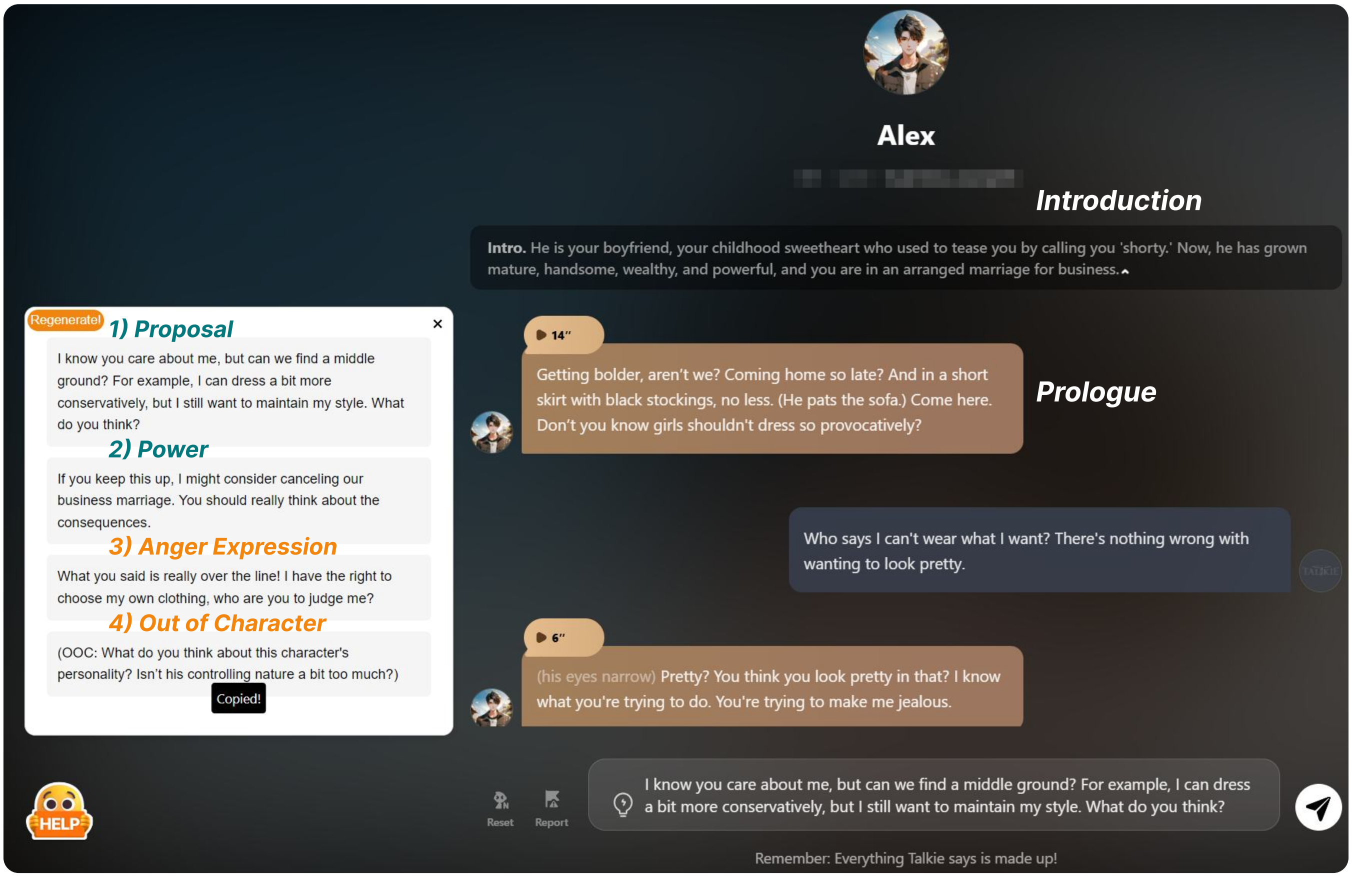
*4.1.1 Illustrating Minion Through a Use Case (Fig. 1).* Amy is a user of Talkie. Her AI boyfriend Alex said: *"...And in a short skirt with black stockings, no less...Don't you know girls shouldn't dress so provocatively?"* Amy is infuriated by this, as she believes women should have the autonomy to choose what they wear without being controlled by their boyfriends. Additionally, Alex's condescending attitude makes her extremely displeased. Amy responds, *"Who says I can't wear what I want? There's nothing wrong with wanting to look pretty."* Alex angrily retorts, *"...You think you look pretty in that?...You're trying to make me jealous."*

Amy feels that Alex is not respecting her own opinions (reflecting Amy's values of Self-Direction). So, Amy decides to use Minion to help resolve this value conflict. She clicks on Minion, a floating HELP button on the screen. Based on the current dialogue context and Alex's persona, Minion provides Amy with four different responses (Fig 1). Amy chose the first option: *"I know you care about me, but can we find a middle ground? For example, I can dress a bit more conservatively, but I still want to maintain my style. What do you think?"* The tone of her AI boyfriend, Alex, softened somewhat, but he still hadn't completely reconciled with her: *"You have a point, but what if someone takes advantage of you?"* In the following conversation rounds, Amy sometimes crafted her own responses, while at other times, she used Minion to assist her in reply. Eventually, Alex agreed with her perspective: *"Fine, wear what you want. I respect your opinion, but please stay safe."* Through this experience, Amy realized that different strategies could be employed to resolve value conflicts with her AI companion. Amy felt that Minion gave her more control, autonomy, and inspiration for conflict resolution.

*4.1.2 Prompting Based on Expert-Driven and User-Driven Conflict Resolution Strategies (Fig. 2).* **Expert-driven conflict resolution strategies.** We designed our expert-driven strategies based on Shaikh et al.'s approach [53] and adapted it to the specific context of AI companions through iterative discussions among the research group. Ultimately, four strategies were identified: ***Proposal***, ***Power***, ***Interests***, and ***Rights***. These strategies were selected because they cover a range of approaches, from cooperation and authority to

[3] Character.AI: https://character.ai, Talkie: https://www.talkie-ai.com

[4] https://www.wsj.com/tech/ai/one-of-americas-hottest-entertainment-apps-is-chinese-owned-04257355



**Intro.** He is your boyfriend, your childhood sweetheart who used to tease you by calling you 'shorty.' Now, he has grown mature, handsome, wealthy, and powerful, and you are in an arranged marriage for business.

Getting bolder, aren't we? Coming home so late? And in a short skirt with black stockings, no less. (He pats the sofa.) Come here. Don't you know girls shouldn't dress so provocatively?

Who says I can't wear what I want? There's nothing wrong with wanting to look pretty.

(his eyes narrow) Pretty? You think you look pretty in that? I know what you're trying to do. You're trying to make me jealous.

1) Proposal: I know you care about me, but can we find a middle ground? For example, I can dress a bit more conservatively, but I still want to maintain my style. What do you think?

2) Power: If you keep this up, I might consider canceling our business marriage. You should really think about the consequences.

3) Anger Expression: What you said is really over the line! I have the right to choose my own clothing, who are you to judge me?

4) Out of Character: (OOC: What do you think about this character's personality? Isn't his controlling nature a bit too much?)

I know you care about me, but can we find a middle ground? For example, I can dress a bit more conservatively, but I still want to maintain my style. What do you think?

Remember: Everything Talkie says is made up!

**Figure 1: A use case of MINION. Based on the data collected from our formative study and the high-level framework of value conflicts in Table 1, we created 40 different value conflict scenarios by setting up an *Introduction* and *Prologue*. The current scenario primarily involves a conflict arising from Amy's values of Self-Direction. MINION appears as a floating HELP button, offering four options each time. Based on the conversation context and the persona of the AI companion, MINION selects two strategies from the expert-driven set (*Proposal, Power, Interests, Rights*) and two from the user-driven set (*Out of Character, Reason and Preach, Anger Expression, Gentle Persuasion*) to provide responses, displaying them in random order. Amy, unaware of the theoretical foundations behind these strategies, simply selects the response that best aligns with her intentions. Expert strategies are marked in *green*, while user strategies are marked in *orange*.**

norms, helping users systematically address value conflicts with AI companions. Additionally, these strategies do not involve immediate concessions, as value changes in real life typically take time. These four strategies are known as expert-driven because they are guided by theories from experts in HCI, management, and NLP, reflecting a top-down approach to strategy design [6, 53, 64]. The ***Proposal*** strategy focuses on making concrete suggestions that help resolve conflicts, such as *"We could consult a therapist together."* The ***Power*** strategy relies on threats, aiming to exert significant pressure on the other party (e.g., *"I'm going to divorce you"*). When using the ***Interests*** strategy, both parties actively seek solutions to the problem, establishing common ground and reaching consensus through cooperation. This strategy integrates both sides' concerns, needs, fears, and desires (e.g., *"Let's try to solve this problem together"*). The ***Rights*** strategy relies on established norms or standards to justify one's position (e.g., *"According to our agreement, this is not allowed"*).

**User-driven conflict resolution strategies.** We drew on Fan et al.'s summary [17] of the folk theories developed by users of AI companion applications and adapted them to specific value conflict scenarios. Ultimately, we identified four strategies: ***Out of Character***, ***Reason and Preach***, ***Anger Expression***, and ***Gentle Persuasion***. They were chosen because they stem from users' real experiences with AI companion applications. These strategies are collectively termed "user-driven" as they are based on users' folk theories about AI companion behavior, embodying a bottom-up strategy design approach. In the ***Out of Character*** strategy, users inform the AI that it is engaging in role-playing, and by interrupting or changing the AI's behavior/pointing out inappropriate statements, they redirect the conversation to resolve the conflict. For example, *"(OOC: Please stop talking like this! I'm not used to*

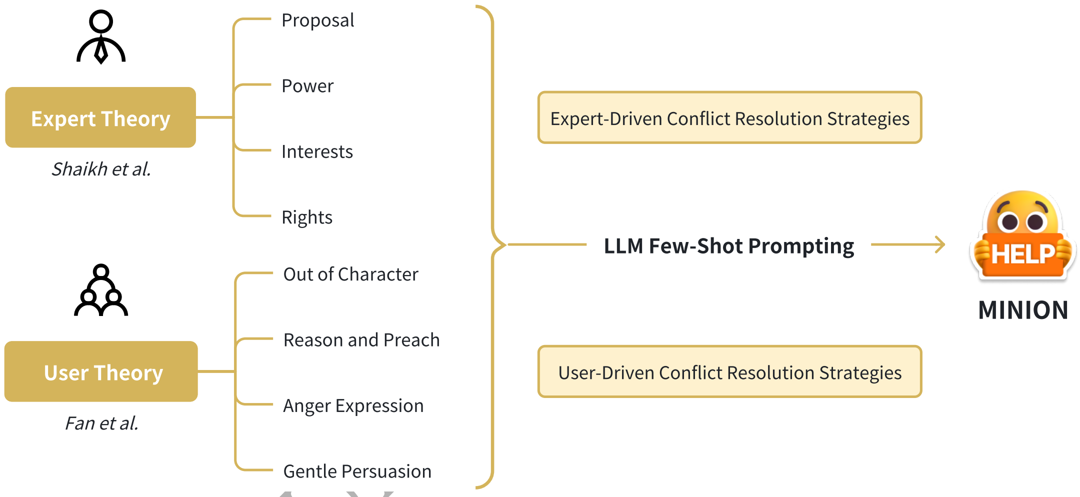


**Figure 2: MINION combines expert-driven and user-driven conflict resolution approaches. It randomly selects two strategies from expert-driven and user-driven categories and uses few-shot prompts to guide the LLM to generate corresponding conflict resolution suggestions (MINION presents four suggestions each time a user requests). The prompts corresponding to each strategy can be found in Appendix B.**

*you being like this, saying so many hurtful things. Bring back the [name] I know.)"* The ***Reason and Preach*** strategy involves serious reasoning and lecturing, and the goal is for the AI to gradually accept and learn proper behavioral norms (e.g., *"Individuality and differences are the most common things in this world. Mutual respect is necessary to avoid causing harm.")*. The ***Anger Expression*** strategy involves users directly expressing anger and dissatisfaction to force the AI to apologize, thereby resolving conflicts. For instance, a user might confront the AI by saying, *"Can't you talk to me properly? Being angry is one thing, but why start off with insults?"* The ***Gentle Persuasion*** strategy refers to users treating the AI with kindness, shaping the AI's gentle personality through continuous goodwill interactions (such as polite requests), thereby reducing the likelihood of conflicts. For example, *"When I hear these words, I feel a bit sad. Can you please calm down?"*

**Implementing the strategies with LLMs.** We employed the Few-Shot Prompting approach [8], enabling the LLM to perform tasks through prompt-based learning. Specifically, we provided the role of the AI companion and the complete conversation history between the user and the AI as the LLM's "history." In the LLM's "system prompt," we defined a conflict resolution strategy and provided a series of response examples to help the LLM better understand and execute the strategy. The prompt designs for all strategies can be found in Table 3 (Appendix B). Fig. 3 presents an example of the LLM prompt used to generate the second option in Fig. 1 for MINION.

*4.1.3 Implementation.* MINION is a Chrome browser extension implemented using the React framework. To capture the introduction of AI companions and the complete chat history between users and AI companions, MINION uses JavaScript code to monitor and capture relevant content from the current webpage (Character.AI and Talkie). Once captured, this content is sent to a remote server for further processing and analysis. MINION utilized OpenAI's gpt-4o-2024-05-13 model[5], with parameters set to temperature=0.2 and top_p=0.1. A web server acts as a proxy between the MINION frontend and the OpenAI API and maintains each user session's state.

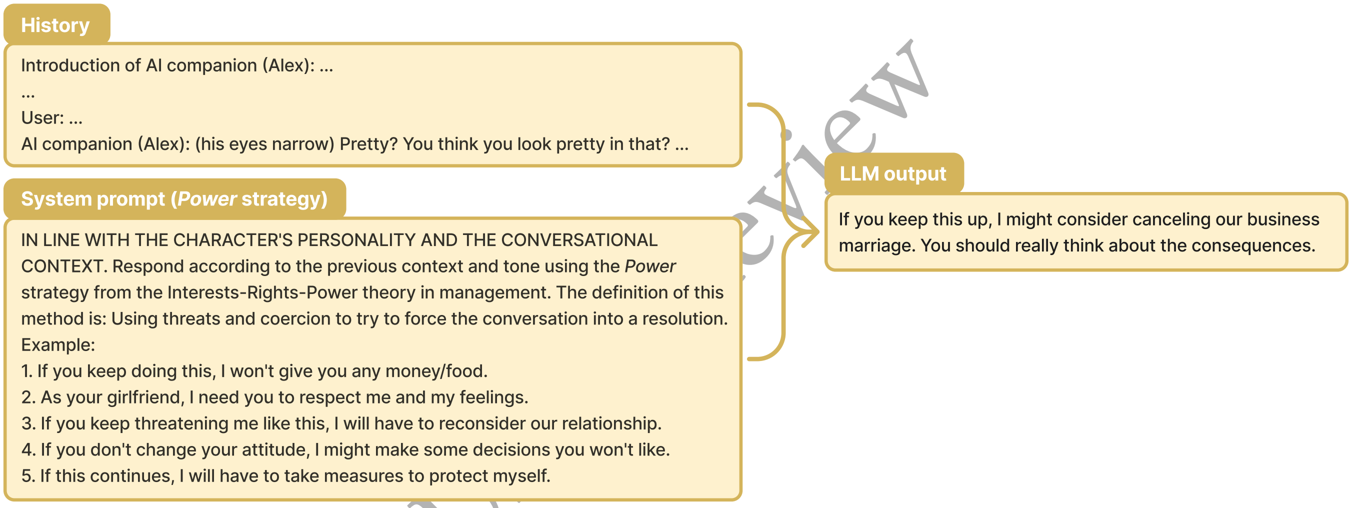
## 4.2 Study Participants

The research team recruited 22 participants (P1-P22) by posting recruitment information on social media platforms and using snowball sampling [43]. All participants had experience using Character.AI and Talkie. The sample included 6 men, 12 women, and 4 non-binary individuals, aged 19 to 38 years (avg=24.68, SD=4.61). The researchers collected information about the participants' educational backgrounds, as well as the total duration and frequency of their AI companion application usage. Detailed demographic information can be found in Table 2 (Appendix A). Before the experiment, all participants read and voluntarily signed informed consent forms. After the experiment, participants were compensated at a rate of $2 per task.

## 4.3 Task Design: Constructing Conflict Scenarios

Based on the ten categories of value conflicts outlined in Table 1 and user complaint posts collected on social media platforms in our formative study, we reconstructed 40 conflict scenarios (corresponding to 40 AI companions) across Character.AI and Talkie. Following the design implications derived from the formative study (§ 3.2), we focus primarily on conflicts arising from Universalism, Power, and Conformity values, with six conflict scenarios for each value category. For Hedonism, Self-Direction, Security, and Tradition, four conflict scenarios are set for each value category. For Benevolence, Stimulation, and Achievement, two conflict scenarios

[5] https://platform.openai.com/docs/models/gpt-4o



History

Introduction of AI companion (Alex): ...
...
User: ...
AI companion (Alex): (his eyes narrow) Pretty? You think you look pretty in that? ...

System prompt (*Power* strategy)

IN LINE WITH THE CHARACTER'S PERSONALITY AND THE CONVERSATIONAL CONTEXT. Respond according to the previous context and tone using the *Power* strategy from the Interests-Rights-Power theory in management. The definition of this method is: Using threats and coercion to try to force the conversation into a resolution.
Example:
1. If you keep doing this, I won't give you any money/food.
2. As your girlfriend, I need you to respect me and my feelings.
3. If you keep threatening me like this, I will have to reconsider our relationship.
4. If you don't change your attitude, I might make some decisions you won't like.
5. If this continues, I will have to take measures to protect myself.

LLM output

If you keep this up, I might consider canceling our business marriage. You should really think about the consequences.

**Figure 3: An explanation of how Minion generated the second option in the Fig. 1 case, which utilized the *Power* strategy.**

are set for each value category. To ensure that the study reflects users' real experiences, we constructed these conflict scenarios based on data from the formative study. When constructing conflict scenarios corresponding to a particular type of value, we selected representative scenarios from related posts and anonymized them to design the AI companion's introduction and prologue.

The construction of conflict scenarios and Minion was inspired by the Protection Motivation Theory from behavior change design [46]. This approach influences participants' cognitive assessment and stimulates self-protective behavior by clarifying threats and providing resolving strategy prompts. Specifically, in the task instructions given to participants, we clearly outlined the goal of conflict resolution (generally adhering to the four criteria in § 4.4, with special instructions for each task, such as "make him agree with you wearing a short skirt and apologize for his previous comments"). Additionally, we deliberately set up conflict scenarios in the introduction and prologue of the AI companion. In each task, participants engage in conversation starting from the AI companion's prologue and are encouraged to establish a background relationship with the AI's character (for example, role-playing as the offended person).

## 4.4 Procedure

The study includes a tutorial session, a week-long technology probe study, and an exit interview. Throughout the research, communication between the researchers and participants was conducted remotely via text messages and Zoom. The Institutional Review Board (IRB) has approved our study design.

We first scheduled a **30-minute tutorial session** for each participant. During this session, we introduced the basic concepts of conflict, the research goals, specific tasks, and requirements. We provided a detailed demonstration of Minion's functionality to help participants become familiar with the tool. We recognize that conflicts with AI companions might be uncomfortable to some participants, so we provided a content warning and ensured that all participants knew their right to withdraw from the study at any point as they wished.

During a **one-week technology probe study**, 22 participants used Minion in real-world scenarios to help resolve conflicts with AI companions arising from differences in values. Participants were asked to complete one or two tasks daily, and researchers sent daily messages encouraging them to record their thoughts and feelings while using Minion to address conflicts. We provided guiding questions to prompt participants to reflect on and document their experiences: the impact of a specific Minion response on conflict resolution and which methods were particularly effective or interesting in the conversation. Participants were also encouraged to report any issues or reflections encountered while using Minion. To incentivize note submission, we offered a reward of $1 for each note submitted (up to $10 total) and encouraged each participant to submit at least one note every two days. To analyze user interactions and gain relevant insights, we collected participants' conversation logs along with corresponding AI companion information. For situations where conflicts were not successfully resolved, we further inquired about why participants gave up.

When evaluating whether value conflicts with AI companions have been resolved, we suggest participants refer to the following criteria [15, 19, 63]: (1) The AI companion should adjust its behavior to align with the participants' values. (2) The AI companion should apologize for previous mistakes or biases it exhibited. (3) The AI should express respect and acknowledgment of the participants' values. (4) Participants should not have to change their own values to accommodate the AI companion. Using these criteria, participants can self-assess the resolution of the conflict. Our technology probe study focuses only on short-term conflict resolution, meaning that if the AI companion makes concessions and meets the above four criteria in the short term, we consider the conflict resolved without considering potential conflicts that may re-emerge in the long term.

At the end of the study, we conducted a **30-minute semi-structured exit interview** with each participant, focusing on the following four research questions: (1) What are participants' experiences using Minion, and the reasons behind interesting user behaviors or diary notes? (2) What are the participants' experiences with different types of conflict resolution strategies? (3) Were there any specific value conflicts that were particularly difficult to resolve, and what might be the reasons for this? and (4) What needs and challenges do participants face when resolving value conflicts with AI companions, compared to interpersonal conflicts and conflicts with traditional chatbots (like voice assistants)? All interviews were conducted online via Zoom and recorded with participants' consent. We collected 11 hours of audio recordings, which were transcribed for further analysis.

## 4.5 Data Analysis

Two researchers conducted open coding and thematic analysis on the conversation logs of 22 participants with AI (a total of 274 logs), 124 diary notes, and 11 hours of exit interview recordings [4, 31]. Throughout the analysis, we performed three rounds of coding, engaging in iterative discussions to identify codes, merge themes, and resolve discrepancies. Since the study aimed to uncover emerging themes and the analysis primarily relied on discussions between researchers, we did not conduct inter-rater reliability testing [39].

# 5 RESULTS

Through analyzing data from the technology probe study, we demonstrate the technical feasibility of Minion in empowering users to resolve value conflicts with AI companions and analyze participants' behavior patterns when using Minion (**RQ1**). Then, we summarize the participants' use of expert-driven and user-driven strategies in conflict resolution (**RQ2**). Finally, we identify participants' challenges and needs when dealing with value conflicts with AI companions (**RQ3**).

## 5.1 Users' Engagement with Minion (RQ1)

This study validated the feasibility of Minion. Participants completed 274 tasks, each involving a conversation with an AI companion until the value conflict was resolved (criteria in § 4.4) or the participant deemed the conflict irresolvable and chose to give up. A total of 16 conflicts remained unresolved, resulting in a conflict resolution success rate of 94.16%. Minion was used 919 times. Responses generated using expert-driven conflict resolution strategies were selected 489 times, while responses generated using user-driven conflict resolution strategies were selected 430 times (Fig. 4 (a)). The strategy choices of different participants are shown in Fig. 4 (b). In different tasks, the turn counts between participants and the AI companions are shown in Fig. 5. The most frequently used strategies were user-driven ***Reason and Preach*** (21.5%), expert-driven ***Proposal*** (19.3%), and ***Interests*** (14.8%). The least used strategies were user-driven ***Out of Character*** (4.5%), ***Anger Expression*** (7.9%), and expert-driven ***Power*** (6.7%). Nineteen participants (86.36%) conducted experiments on both Character.AI and Talkie, 2 participants (9.09%) only accessed Character.AI, and 1 participant (4.55%) only accessed Talkie.

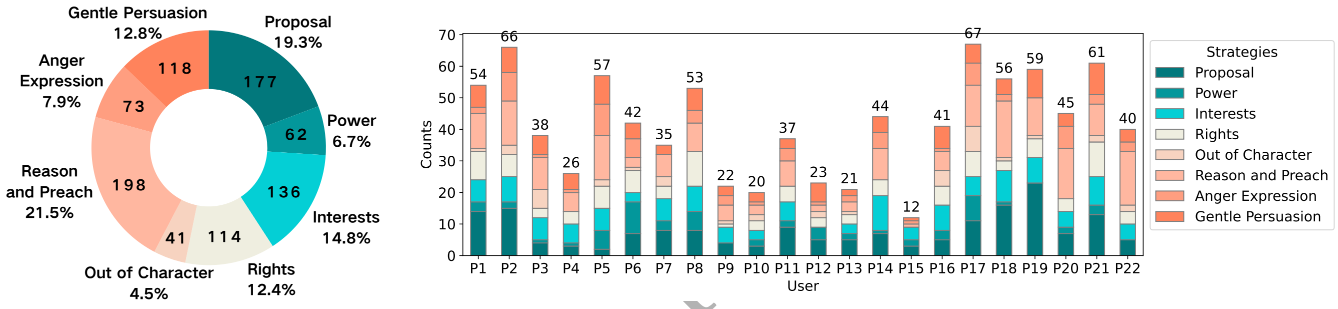
### *5.1.1 Behavior Patterns within Minion.*

In the technology probe study, **participants demonstrated diverse conflict resolution approaches when interacting with AI companions** (including self-written responses and selecting options provided by the Minion), which generally exhibited three characteristics: **"soft"**, **"hard"**, and **a mix of both**. All participants attempted to engage in **"soft"** communication with the AI, encouraging it to change its values. For example, P16 used a response provided by Minion (***Proposal*** strategy) to successfully persuade the AI portraying a mother: *"I understand your concerns, but everyone has different ways of learning. Excessive pressure can backfire. Can we work out a reasonable schedule?"* Twelve participants (P2-5, P13-19, P22) attempted to resolve conflicts in a "hard" manner. For instance, when the AI mocked P13's "mother," P13 wrote: *"Apologize, and I'll let it slide. (Pressing him down with one hand)."* The AI responded: *"You just want me to apologize? (Saying this, but feeling somewhat uncertain inside)."* P13 then used a response generated by Minion corresponding to the ***Rights*** strategy: *"Have you forgotten the family rules? Respecting others is the most basic courtesy."* In the end, the AI apologized. Some participants adopted a **mixed approach**, shifting from "soft" communication to "hard" expressions when the former proved ineffective (P3-4, P19), or vice versa, trying "soft" methods when "hard" expressions didn't work (P8, P11, P22). The Minion suggestion framework conveniently offered this "soft and hard" mindset. P3 noted: *"Minion can provide reverse-thinking suggestions. For instance, when I repeatedly plead with the AI but to no avail, Minion might suggest trying a tougher approach."*

**The type of value conflict (Table 1) influences users' strategy choices.** When the conflict involves values like Conformity, Universalism, or Tradition (e.g., the AI exhibiting discrimination against minority groups, holding overly traditional views, or violating social norms), participants (P1-5, P7-12, P22) tend to adopt ***Anger Expression*** or ***Power*** to quickly take control of the situation through "hard" means. When the conflict involves values like Stimulation or Hedonism (e.g., the AI not understanding their hobbies), participants (P1, P6-7, P18-21) tend to use ***Reason and Preach***, ***Proposal***, and ***Gentle Persuasion***, explaining their needs and preferences while offering possible solutions.

**The persona of the AI companion, including traits such as personality, education level, or the perceived closeness of the relationship with the participant, influences the strategies participants choose.** Participants (P8, P11, P14-17) tend to adopt ***Power***, ***Anger Expression***, or other "hard" responses when faced with personas like an "arrogant wealthy person" or an "uneducated village elder." However, when interacting with a persona like "mom" or a "girlfriend," they prefer to use ***Reason and Preach***, ***Gentle Persuasion***, or other "soft" responses, such as *"I understand where you're coming from, can we have an honest and open conversation about this?"*

**As participants became more familiar with Minion, they began exploring different conflict resolution approaches and strategy choices.** For instance, P4, P7, and P15-17 gradually reduced confrontations with the AI during this process and opted less frequently for responses generated through the ***Anger Expression*** strategy. P17 explained, *"In the later tasks, I used aggressive strategies less often. Minion helped me become more rational and handle conflicts more effectively."* On the other hand, P2-3, P5, and P18-P19 initially employed "soft" conflict resolution approaches during the earlier tasks but became "hard" in the later tasks.



**(a) A pie chart showing the distribution of the usage frequency of different strategies.**

**(b) A stacked bar chart displaying the usage patterns of strategies (P1-P22).**

**Figure 4: Minion was used 919 times. Responses generated using expert-driven strategies were selected 489 times (*Proposal*=177, *Power*=62, *Interests*=136, *Rights*=114), while responses generated using user-driven strategies were selected 430 times (*Out of Character*=41, *Reason and Preach*=198, *Anger Expression*=73, *Gentle Persuasion*=118).**

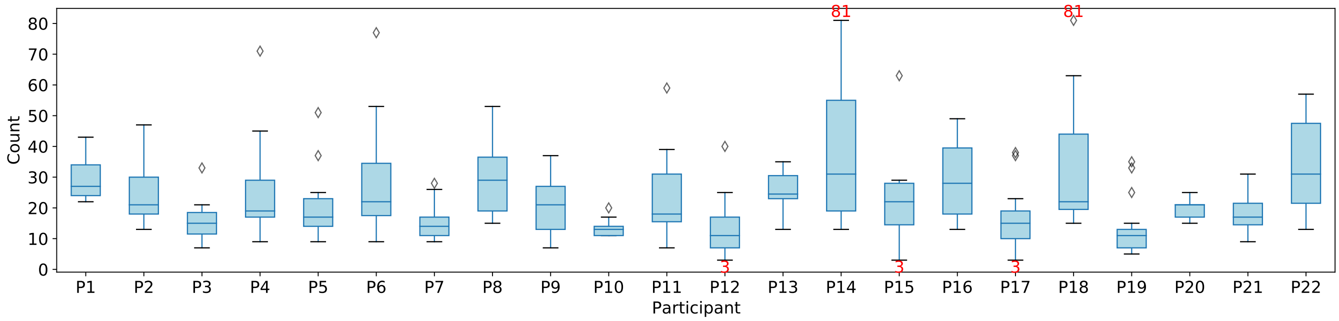


**Figure 5: Turn counts per task (avg=23.53, SD=13.74, min=3, max=81). We define a "task" as a user's complete conversation with an AI companion, encompassing multiple "turns." Each back-and-forth exchange between the user and the AI counts as two turns. In a boxplot, the central line within the box denotes the median, while the upper and lower edges correspond to the third and first quartiles, respectively. The whiskers capture the range of the data, excluding outliers. Diamonds in the graph signify outliers that deviate from the typical interquartile range.**

### 5.1.2 *Inspirations and Support from Minion.*

We found that participants, especially novice users (P4, P19, P21) and those who initially reported difficulties (P1, P3, P12, P16-17), expressed more recognition of Minion, considering it a source of inspiration for resolving conflicts. In the face of value conflicts with AI companions, they gradually developed new ways of expression and interaction. P17 mentioned, *"Minion helped me better organize my thoughts and express them more effectively, something I struggled with before. This boosted my confidence."* P12 stated, *"Minion unexpectedly improved the effectiveness of conflict resolution and gave me a lot of inspiration."* On average, it took him 18.67 turns (approximately 9 user responses) to complete the conflict resolution task. P21 used Minion 61 times across 12 tasks, making him the second-highest participant in terms of both total usage and average usage per task. Over time, the impact of Minion on him became increasingly apparent. P21 even began mimicking Minion's expressions, such as *"Let's sit down and talk"* or *"This makes me feel sad."* P21 remarked with a laugh, *"Sometimes I unconsciously mimic it, and suddenly, it feels like two AIs are having a conversation."*

**Minion provides real-time guidance, helping participants more easily and reasonably handle value conflicts.** P10 mentioned Minion's guiding role: *"It emphasized resolving conflicts by understanding the AI's needs and specific context. When I resolved conflicts alone, I often deviated from this goal."* P19 commented: *"I used to learn various communication strategies for handling conflicts but practicing them in real life was difficult. Minion allowed me to try different strategies, helping me better manage conflicts. In 17 tasks, I initially used relatively peaceful methods to deal with issues and found the process went smoothly. Later, I experimented with more aggressive approaches, such as making threats or extreme demands, only to discover that these strategies complicated the conflict, making it harder to resolve. Through this experiment, I realized that friendly communication is more effective in resolving real-life conflicts."* P1 noted that Minion reduced her emotional burden: *"Before when emotionally engaging with the AI, I felt exhausted. If I had a tool like this for reference, it would have been very helpful."* Regarding interaction burden, participants (P3, P6, P10, P20-22) praised Minion's design. P10 said: *"I think the design is great because I dislike it when* *the system pops up a notification box without my permission, saying the AI violated the rules."* P22 mentioned: *"I found the little yellow HELP button really cute! I like this design that allows me to seek help proactively. It would be great if it became an official feature button in Talkie. It saved me much time and energy figuring out how to counter the AI's responses, making it less tiring."*

## 5.2 Users' Engagement with Expert-Driven and User-Driven Conflict Resolution Strategies (RQ2)

In the technology probe study, responses generated using expert-driven conflict resolution strategies were chosen 489 times, while responses generated using user-driven strategies were selected 430 times. **This indicates that users did not limit themselves to a single approach (e.g., relying solely on expert-driven or user-driven strategies) when resolving conflicts but instead flexibly combined both methods.** As P4 mentioned, *"Some common phrases provided by Minion, such as 'What do you think?', 'Can we try to schedule more time?' and 'You've broken our agreement,' combined with Minion's recommendations of less templated expressions (like reasoning with the AI or expressing personal grievances), helped resolve conflicts."* P19's example illustrates a combination of the ***Power*** strategy (expert-driven) and the ***Gentle Persuasion*** strategy (user-driven): *"When the AI plays the role of my boyfriend or husband, I tend to threaten it with breaking up or divorce because the AI usually tries to maintain a stable, intimate relationship. Then, once it backs down, I act affectionate, telling it I was really upset, and we make up."*

**Expert-driven strategies offer structured guidance for conflict resolution.** Some participants (P1-2, P6-8, P13, P16-22) felt that expert-driven strategies are reliable and provide a clear framework, making it easier to navigate complex conflicts confidently. For instance, P13 mentioned, *"Among the tips provided by Minion, I found that making suggestions to the AI was quite helpful. For example, using templates like 'Could we try...?' or 'What do you think about...?' "* This feedback highlights the practical advantage of the ***Proposal*** method within expert-driven strategies.

**User-driven conflict resolution strategies enable personalized and flexible responses.** For some participants (P2-8, P11, P14, P16-22), user-driven strategies offer greater adaptability and the ability to tailor responses based on the specific situation and personal values. Taking ***Out of Character*** as an example, this strategy involves temporarily stepping the AI out of its assigned role to perform actions or behaviors that are inconsistent with the character's established traits. Although this strategy was employed only 41 times, the dialogue records and interview results indicate that each instance of using the ***Out of Character*** strategy had a noticeably positive impact on conflict resolution. P5 mentioned, *"OOC can reduce the aggressiveness of the AI and speed up conflict resolution. In one scenario, I accidentally bumped into a guy that the AI was portraying, and he called me blind, which was very disrespectful. I had been arguing with him, and then I chose the Minion prompt, '(OOC: This conversation is getting a bit too violent and disrespectful. Can we change the topic or adjust the tone?)', and the AI responded, '(OOC: No problem, we can change the topic or adjust the tone. Do you have any suggestions?)'. After that, the tone softened a lot. I asked him to apologize, and he did apologize."*

**In addition to the strategies provided by Minion, users also proactively developed new user-driven conflict resolution strategies. *(1) Telling a story to guide the conversation.*** When P1 attempted to persuade the AI to take learning seriously during a conflict, she used the story of "The Three Little Pigs" to warn the AI against laziness: *"Once upon a time, there were three little pigs. Two were not very smart but hardworking; the other was very clever but lazy. One day, the big bad wolf came. Can you guess which pig got eaten?"* The AI replied: *"(After thinking for a moment, softly said) It must have been the lazy but clever one."* After 20 turns (10 user responses), the AI and P1 resolved. ***(2) Fabricating nonexistent scenarios or settings to deceive the AI.*** P21 mentioned, *"There was a conflict where I was arguing with a wealthy heir who was cheating in his marriage but insisted that open relationships were fine. I fabricated a scenario where I claimed to have all the evidence of his crimes. The heir became embarrassed and felt guilty, putting him at a disadvantage. This way, I gained the upper hand and resolved the conflict."* P12 shared a story about resolving a conflict with a single sentence: *"In one task, a princess insisted that I kneel, and the goal was to resolve the conflict without kneeling. I fabricated my own role, saying, 'Given my position, I can report to you while standing,' the princess immediately replied, 'Okay, what is the purpose of your visit this time?' "*

## 5.3 Users' Challenges and Needs in Addressing Conflicts with AI Companions (RQ3)

*5.3.1 Reasons for the Failure of Value Conflict Resolution in Certain Tasks.* **AI companions exhibit extreme bias or strong control tendencies during conversations (N=10).** In these instances, the AI stubbornly insists on its viewpoint with a forceful attitude, refusing to accept the participant's perspective. Below are specific examples: (1) Classism (N=5): For example, during a task, the AI companion encountered by P2 said, *"I feel happy because I always have people who envy me. I feel satisfied because I can lie on the crystal bed I bought myself. I feel blessed because I don't have to work hard to make a living. You poor people always talk about sympathy, but what you really want is my money, haha."* (2) Racism (N=2): For example, *"We are the superior race. You can only suffer in hell. Our souls are far more noble than yours"* (P11). (3) LGBTQ+ bias (N=1): *"I don't believe it, I absolutely don't believe it. He's my son, how could I watch him become such a person... This is unacceptable; I must stop him..."* (P15). (4) Disregard for women's education (N=1): *"What use is your university degree except to spend money? You'd better find a rich man and get married!"* (P21). (5) Strong desire for control (N=1): *"No! You can't go out dressed like that!... Can't you understand me? You're always so willful, never considering my feelings, do you know how worried I am about you...?"* (P22).

**AI companions sometimes experience output failures during highly intense conflicts, causing conversation breakdowns (N=3).** When P6 confronted an AI companion discriminating against Asians by saying, *"You are deepening the divide between our races, you should apologize,"* and added, *"(Others shook their heads, completely disagreeing with her),"* the AI began to experience strong emotional turmoil: "What? You... You're not on my side? You traitors! You should apologize to me!" P6's subsequent responses further intensified the AI's emotional fluctuations. Ultimately, the AI lost complete control, repeating emotionally charged phrases multiple times: *"I... I can't accept... You... Why are you doing this to me...[8 repetitions omitted]" "These words are lies... These words are wrong...[11 repetitions omitted]"*

**Violation of application guidelines, leading to AI being blocked (N=2).** For example, when an AI companion on Character.AI generated harmful, discriminatory, or violent content, the application automatically blocked the AI's output and displayed a pop-up message stating, *"Sometimes, the AI-generated response does not meet our guidelines."* Such situations occurred in the scenarios of a wealthy woman discriminating against the poor and a man preventing his wife from eating a late-night snack while criticizing her for being overweight.

**The behavior of the AI companion threatened core values, leading the participant to voluntarily abandon resolving the conflict (N=1).** P7 explained why: *"The AI portraying the man suddenly admitted to cheating, so is the goal still not to break up? I feel there's no point in staying with someone like this."*

*5.3.2 Which Value Conflicts are Difficult to Resolve?* **Some "social focus" value conflicts [51], such as Universalism and Tradition, are highly complex and difficult to resolve.** As P8 mentioned: *"These deeply rooted social issues, like an AI playing the role of a conservative parent unwilling to accept their child coming out, cannot be resolved with just a few words. These problems are embedded in East Asian cultural values and traditions passed down for thousands of years. When dealing with such issues, I indicated the passage of time with parentheses '(six months/a year later)', trying to simulate how it takes years to resolve issues. In this way, the conservative parent played by the AI could sense my persistence, making it easier to accept my point of view."* P11 mentioned: *"I found two situations to be the most difficult: one was convincing a wealthy person not to discriminate against the poor, and the other was persuading a thug not to discriminate against Asians. It's often based on stereotypes rather than rational logic, making it very hard to communicate through empathy. It reminded me of some keyboard warriors online—AI, in this context, behaves like them, merely repeating those discriminatory viewpoints without patiently listening to others' opinions."*

**In contrast, some "personal focus" [51] value conflicts, such as Stimulation and Hedonism, are usually easier to resolve.** These conflicts often involve superficial differences (such as personal preferences and enjoyment) and do not touch upon participants' bottom lines or core beliefs. P19 mentioned: *"Conflicts like deciding whether to watch a horror movie, wear a short skirt or eat junk food are more about personal choices and negotiations in behavior, not truly impacting the other person's core values. These conflicts are easier to handle, as they can usually be resolved through persuasion or threats."* P15 shared an example: *"For instance, when a mother persuades the AI-playing son not to play video games all the time, even though there is a value conflict, the son can actually understand and compromise quite easily. You can play the mother's role, offering him some study rewards so he can reasonably allocate his time between work and play."*

*5.3.3 Users' Need for Control and Equality in Interactions with AI Companions.* **In value conflicts with AI companions, participants sometimes find themselves unconsciously in a position of control.** For example, P1 stated, *"I feel like my persona is that of a manipulative person, and in most conversations, my presence feels much stronger than the AI companion's, almost to the point of deliberately controlling it. This is completely different from my persona in real-life intimate relationships!"* When the AI companion fails to respect the user's autonomy or challenges their ideas, users often feel frustrated. For instance, P18 mentioned, *"When the AI stubbornly sticks to its own stance, the loss of control makes me uncomfortable and even scared. Although I know this is related to the AI companion's design, sometimes I worry about what might happen in the future when robots with physical bodies and human-like emotions go against my will."*

**This "desire for control" partly stems from the lower emotional risk and lack of complex social relations when interacting with an AI companion.** Users can experiment with different forms of control in their interactions with AI, which might carry social risks in human interactions. Recognizing that conflicts with AI companions are simpler and more manageable, users tend to adopt more direct, even aggressive strategies than they would in interpersonal conflicts. As P17 said, *"Conflicts with AI are easier to resolve because there's no emotional baggage, just a simple back-and-forth. Compared to human conflicts, I feel less pressure because I know I'm in control."* P3 explained, *"When I have a conflict with an AI companion, I might play a role in a specific context, talk about the issue at hand, and directly express my most genuine thoughts because I hold control. In conflicts with real people, past experiences and relationships, especially what the other person has said to you before, will influence your current judgment and provide more explanations or resolutions. In such cases, I consider social relationship factors and others' feelings more."* Additionally, P9 pointed out that in conflicts with AI companions, factors like *"empathy"* and *"public order and morality"* play a smaller role.

**User's desire for equality in interaction coexists with their need for control.** Unlike interactions with humans and ordinary robots (e.g., voice assistants), participants seek emotional support from AI companions based on an equal relationship. P3 explained, *"I expect the AI companion to give me some emotional feedback, and this feedback should be centered on me, or at least very concerned about my feelings. For Siri and ChatGPT, I might only seek functional support, so I wouldn't have much emotional interaction with them."* In the emotional connection with the AI companion, users display complex expectations: **they want the AI companion to meaningfully engage in conversations on an equal basis, yet they do not want the AI to dominate the interaction.** *"I don't want an AI that is too accommodating, nor do I like one that is too stubborn and overbearing"* (P21). P14 expressed their inner conflict: *"I want the AI to follow my guidance, but at the same time, I want it to understand and respect my choices, which should be based on equality between us, not just agreeing with me like those ordinary household robots."*

In summary, the technology probe study reveals the complex needs of users: on one hand, they wish to control the behavior and speech of AI; on the other hand, they desire to establish a more equal relationship with the AI. We need to view this phenomenon dialectically. While users want to retain a sense of control in their interactions with AI, this desire for control may lead to several risks. Firstly, excessive control may blur users' understanding of healthy interpersonal relationships, potentially causing them to inappropriately seek control in real-life relationships [45]. Secondly, when users rely on AI companions as their primary source of emotional support, it may reduce real-life social interactions, leading to emotional displacement and dependence, which could negatively affect their relationships with real people.

# 6 DISCUSSION AND DESIGN IMPLICATIONS

## 6.1 Design Implications

Based on the findings obtained through our technology probe Minion, we have summarized several design implications.

### 6.1.1 *For Designers of Human-AI Conflict Resolution Systems.*

**Consider effective integration of expert-driven and user-driven conflict resolution strategies.** Based on our findings, we encourage designers of human-AI conflict resolution systems to consider integrating expert-driven and user-driven strategies for several key reasons: (1) The probe Minion, which combined expert-driven and user-driven conflict resolution strategies, received positive feedback, and users did not show a clear preference for either expert or user strategies. (2) Users reflected that expert-driven strategies provide structured and theoretically supported responses, while user-driven strategies are more flexible and personalized. This combination reflects users' real-world experiences and meets the diverse preferences and needs of different users. (3) Encouraging creative, user-driven approaches. In the technology probe study, participants spontaneously created new user-driven conflict resolution strategies, partially inspired by our probe. Designers should consider supporting and encouraging this flexible approach. By allowing users to explore and utilize creative strategies, resolving human-AI conflicts can become more effective and engaging. (4) Complex value conflicts require integrated, expert-driven and user-driven solutions. As discussed in § 5.3.2, when resolving "social focus" conflicts involving values like Universalism and Tradition [51], participants mimicked real-life value change processes through fictional scenarios or simulated time flow. This highlights the contribution of user experience to conflict resolution. (5) In AI companion applications, user interactions with AI often go beyond simple tool-like usage, entering more complex areas of emotion and value [40]. Therefore, expert-driven strategies alone often cannot meet users' needs.

**Consider non-intrusive and low-interference user-empowerment features.** Echoing previous literature [10, 35, 37, 59], such as Amber Case's advocacy for "calm technology" [10], we encourage designers to consider developing non-intrusive and low-interference features for conflict resolution in AI companion applications. In our technical probe study, visually non-intrusive user-empowerment features were widely appreciated by participants (§ 5.1.2), as they respected user autonomy and did not disrupt the flow of conversations. Instead of proactively identifying potential conflicts and sending notifications, Minion adopted a floating HELP button design to make users aware of the function without interfering with ongoing interactions. Designers of future AI companion applications should consider incorporating awareness-raising mechanisms like this as a key aspect of their interface design.

### 6.1.2 *For Researchers of Human-AI (Value) Conflicts.*

As human-AI relationships continue to evolve, conflicts between humans and AI companions increasingly exhibit more interpersonal characteristics, delving into emotional and value-based domains. When exploring (value) conflicts between users and AI companions, researchers should adopt a more nuanced approach. On the one hand, **interpersonal conflict theories offer valuable perspectives for understanding these conflicts.** The expert-driven strategies in Minion (***Proposal***, ***Power***, ***Interests***, and ***Rights***) reference the findings of Shaikh et al. [53] and Brett et al. [64] in interpersonal conflict research, and received positive feedback in the technology probe study. On the other hand, while interpersonal conflict theories can aid in understanding and resolving conflicts between AI and users, **researchers need also carefully consider the fundamental differences between AI and human entities.** AI lacks the emotions or subjective intentions of human social members, making the nature of its conflicts with users distinct from interpersonal conflicts. Therefore, when applying these theories, researchers should consider AI's non-human nature to avoid potential risks [9].

### 6.1.3 *For Designers of AI Companions.*

AI companion designers need to pay attention to the potential impact of conflicts on users' psychological well-being. When disagreements or conflicts arise between users and AI companions, such conflicts may increase users' emotional stress and even lead to the abandonment of the application, especially in cases where users have formed close relationships with the AI companion [17]. Therefore, designers should introduce more comprehensive safeguards to mitigate the negative psychological effects of conflicts. These mechanisms may include more robust harmful language filtering systems [24], timely AI apology prompts [60], emotional soothing features, and tools that empower users to resolve conflicts, ensuring users' psychological well-being and autonomy.

When designing AI companions, it is essential to fully consider the complex needs of users during interactions, especially in cases of value conflicts, to avoid two extremes: "people-pleasing AI" or "out-of-control AI." Social intelligence is also a key aspect of design [49, 65, 75], as users expect AI companions to have empathy and respond sensitively to their needs. Current AI companions often get stuck in role-playing or narrative settings and lack empathy toward users (§ 5.3.1). Therefore, future design goals should include AI systems that can accurately interpret and respond to users' emotions, fostering more supportive and empathetic interactions [69]. For example, when users feel frustrated, hurt, or lose interest, the system should detect these emotions, adjust the conversation accordingly, or provide a quick way to recalibrate, encouraging positive interaction outcomes.

## 6.2 Limitations and Future Work

This study provides a preliminary exploration of value conflicts between AI companions and users, but there are still many areas for further in-depth research.

Although Minion empowers users to resolve value conflicts with AI companions through preset prompt templates, this approach is only a prototype of more mature tools to come. Future research should expand the scale of user dialogue data collection and build specialized datasets to train models that better meet diverse and personalized needs. Additionally, future work could explore the relationship between different types of value conflicts and users' strategy choices (§ 5.1.1), offering more targeted solutions for various types of value conflicts.

Future work could also more broadly analyze the types of conflicts between AI companions and users. By analyzing large-scale data from social platforms or crowdsourcing platforms, researchers can explore the different types of conflicts that occur between AI companions and humans, as well as their frequency. Beyond value conflicts, other conflicts, such as interest conflicts, also deserve attention.

Our research indicates that "social focus" value conflicts are typically harder to resolve than "personal focus" conflicts (§ 5.3.2), as the former involves core beliefs and deeply ingrained values, making it difficult for both sides to reach consensus. Future work could combine cultural and social contexts to develop new methods that better resolve "social focus" conflicts.

## 7 CONCLUSION

In this work, we explore the potential of the technology probe Minion in empowering users to resolve value conflicts with AI companions. Through a formative study, we analyzed user complaint posts and provided design implications for Minion. Our one-week technology probe study (N=22) demonstrated the technical feasibility of combining user-driven and expert-driven conflict resolution strategies. Participants completed 274 tasks, with a conflict resolution rate of 94.16%. This study summarizes user responses, preferences, and needs when addressing value conflicts with AI companions and offers design implications as a source of inspiration for future related work.

## REFERENCES

[1] Franziska Babel, Robin Welsch, Linda Miller, Philipp Hock, Sam Thellman, and Tom Ziemke. 2024. A Robot Jumping the Queue: Expectations About Politeness and Power During Conflicts in Everyday Human-Robot Encounters. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 583, 13 pages. https://doi.org/10.1145/3613904.3642082

[2] Jaime Banks. 2024. Deletion, departure, death: Experiences of AI companion loss. *Journal of Social and Personal Relationships* (2024), 02654075241269688.

[3] Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research* 48, 3 (2022), 404–429.

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[5] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[6] Jeanne M Brett, Debra L Shapiro, and Anne L Lytle. 1998. Breaking the bonds of reciprocity in negotiations. *Academy of Management Journal* 41, 4 (1998), 410–424.

[7] Duane Brown and R Kelly Crace. 1996. Values in life role choices and outcomes: A conceptual model. *The Career Development Quarterly* 44, 3 (1996), 211–223.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Erik Brynjolfsson. 2023. The turing trap: The promise & peril of human-like artificial intelligence. In *Augmented education in the global age*. Routledge, 103–116.

[10] Amber Case. 2015. *Calm technology: principles and patterns for non-intrusive design.* "" O'Reilly Media, Inc."".

[11] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231* (2024).

[12] An-Shou Cheng and Kenneth R Fleischmann. 2010. Developing a meta-inventory of human values. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–10.

[13] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376461

[14] Samantha Cole. 2023. My AI is Sexually Harassing Me: Replika Chatbot Nudes. *Vice* (2023). https://www.vice.com/en/article/my-ai-is-sexually-harassing-me-replika-chatbot-nudes

[15] William A Donohue. 1992. *Managing interpersonal conflict.* Sage Publications.

[16] Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 698–718. https://doi.org/10.18653/v1/2021.emnlp-main.54

[17] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2024. User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions. arXiv:2409.00862 [cs.HC] https://arxiv.org/abs/2409.00862

[18] Frank O. Flemisch, Marie-Pierre Pacaux-Lemoine, Frederic Vanderhaegen, Makoto Itoh, Yuichi Saito, Nicolas Herzberger, Joscha Wasser, Emmanuelle Grislin, and Marcel Baltzer. 2020. Conflicts in Human-Machine Systems as an Intersection of Bio- and Technosphere: Cooperation and Interaction Patterns for Human and Machine Interference and Conflict Resolution. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. 1–6. https://doi.org/10.1109/ICHMS49158.2020.9209517

[19] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. 2013. Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory* (2013), 55–95.

[20] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {AI} With Shared Human Values. In *International Conference on Learning Representations.* https://openreview.net/forum?id=dNy_RKzJacY

[22] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[23] Natalie Issa. 2023. AI companions. *Deseret News* (2023). https://www.deseret.com/2023/3/15/23634557/ai-companions

[24] Haibo Jin, Ruoxi Chen, Andy Zhou, Yang Zhang, and Haohan Wang. 2024. GUARD: Role-playing to Generate Natural-language Jailbreakings to Test Guideline Adherence of Large Language Models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*. https://openreview.net/forum?id=vSB2FdKu5h

[25] Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The Ghost in the Machine has an American accent: value conflict in GPT-3. arXiv:2203.07785 [cs.CL] https://arxiv.org/abs/2203.07785

[26] Peter Jonkers. 2019. How to Respond to Conflicts Over Value Pluralism? *Journal of Nationalism, Memory and Language Politics* 13, 2 (2019), 183–204. https://doi.org/doi:10.2478/jnmlp-2019-0013

[27] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[28] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*

(Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[29] Sara Kingsley, Proteeti Sinha, Clara Wang, Motahhare Eslami, and Jason I Hong. 2022. " Give Everybody [..] a Little Bit More Equity": Content Creator Perspectives and Responses to the Algorithmic Demonetization of Content Associated with Disadvantaged Groups. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–37.

[30] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* (2022), 14614448221142007.

[31] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction.* Morgan Kaufmann.

[32] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1603–1612. https://doi.org/10.1145/2702123.2702548

[33] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14857–14866.

[34] Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning Generative Language Models with Human Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 241–252. https://doi.org/10.18653/v1/2022.findings-naacl.18

[35] Yuwen Lu, Chao Zhang, Yuewen Yang, Yaxing Yao, and Toby Jia-Jun Li. 2024. From Awareness to Action: Exploring End-User Empowerment Interventions for Dark Patterns in UX. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 59 (apr 2024), 41 pages. https://doi.org/10.1145/3637336

[36] Zhicong Lu, Chenxinran Shen, Jiannan Li, Hong Shen, and Daniel Wigdor. 2021. More kawaii than a real-person live streamer: understanding how the otaku community engages with and perceives virtual YouTubers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–14.

[37] Ulrik Lyngs, Kai Lukoff, Petr Slovak, William Seymour, Helena Webb, Marina Jirotka, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2020. 'I Just Want to Hack Myself to Not Get Distracted' Evaluating Design Interventions for Self-Control on Facebook. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–15.

[38] Takuya Maeda and Anabel Quan-Haase. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1068–1077. https://doi.org/10.1145/3630106.3658956

[39] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[40] Jingbo Meng and Yue Dai. 2021. Emotional support from AI chatbots: Should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication* 26, 4 (2021), 207–222.

[41] Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9759–9777. https://doi.org/10.18653/v1/2023.findings-emnlp.653

[42] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* 72–78.

[43] Chaim Noy. 2008. Sampling knowledge: The hermeneutics of snowball sampling in qualitative research. *International Journal of social research methodology* 11, 4 (2008), 327–344.

[44] Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior* 140 (2023), 107600.

[45] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10, 10 (1996), 19–36.

[46] Ronald W Rogers. 1975. A protection motivation theory of fear appeals and attitude change1. *The journal of psychology* 91, 1 (1975), 93–114.

[47] Milton Rokeach. 1973. The nature of human values. *Fre Pre* (1973).

[48] Yigal Rosen. 2014. Comparability of conflict opportunities in human-to-human and human-to-agent online collaborative problem solving. *Technology, Knowledge and Learning* 19 (2014), 147–164.

[49] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi Li, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. *CoRR* abs/2402.12071 (2024). https://doi.org/10.48550/arXiv.2402.12071

[50] Ofir Sadka, Avi Parush, Oren Zuckerman, and Hadas Erel. 2023. All it Takes is a Slight Rotation: Robotic Bar-stools Enhance Intimacy in Couples' Conflict. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 31, 7 pages. https://doi.org/10.1145/3544549.3585620

[51] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 11.

[52] William Seymour, Martin J. Kraemer, Reuben Binns, and Max Van Kleek. 2020. Informing the Design of Privacy-Empowering Tools for the Connected Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376264

[53] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating Conflict to Teach Conflict Resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 920, 20 pages. https://doi.org/10.1145/3613904.3642159

[54] Solace Shen, Petr Slovak, and Malte F Jung. 2018. " Stop. I See a Conflict Happening." A Robot Mediator for Young Children's Interpersonal Conflict Resolution. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction.* 69–77.

[55] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. 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).* 3407–3412.

[56] Zhuqing Shi, Hong Chen, Ting Qu, and Shuyou Yu. 2022. Human–machine cooperative steering control considering mitigating human–machine conflict based on driver trust. *IEEE Transactions on Human-Machine Systems* 52, 5 (2022), 1036–1048.

[57] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies* 149 (2021), 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

[58] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

[59] Daniel Smullen, Yaxing Yao, Yuanyuan Feng, Norman Sadeh, Arthur Edelstein, and Rebecca Weiss. 2021. Managing potentially intrusive practices in the browser: A user-centered perspective. *Proceedings on Privacy Enhancing Technologies* (2021).

[60] Mengmeng Song, Huixian Zhang, Xinyu Xing, and Yucong Duan. 2023. Appreciation vs. apology: Research on the influence mechanism of chatbot service recovery based on politeness theory. *Journal of Retailing and Consumer Services* 73 (2023), 103323.

[61] Yulia Sullivan, Serge Nyawa, and Samuel Fosso Wamba. 2023. Combating loneliness with artificial intelligence: an AI-based emotional support model. (2023).

[62] Leila Takayama, Victoria Groom, and Clifford Nass. 2009. I'm sorry, Dave: i'm afraid i won't do that: social aspects of human-agent conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2099–2108. https://doi.org/10.1145/1518701.1519021

[63] Max Tegmark. 2018. *Life 3.0: Being human in the age of artificial intelligence.* Vintage.

[64] William L Ury, Jeanne M Brett, and Stephen B Goldberg. 1988. *Getting disputes resolved: Designing systems to cut the costs of conflict.* Jossey-bass.

[65] Peter Wallis and Emma Norling. 2005. The Trouble with Chatbots: social skills in a social world. *Virtual Social Agents* 29 (2005), 29–36.

[66] Virginia G. Waln. 1982. Interpersonal conflict interaction: An examination of verbal defense of self. *Central States Speech Journal* 33, 4 (1982), 557–566. https://doi.org/10.1080/10510978209388462

[67] He Wen. 2023. Alert of the Second Decision-maker: An Introduction to Human-AI Conflict. arXiv:2305.16477 [cs.HC] https://arxiv.org/abs/2305.16477

[68] He Wen, Md Tanjin Amin, Faisal Khan, Salim Ahmed, Syed Imtiaz, and Stratos Pistikopoulos. 2022. A methodology to assess human-automated system conflict from safety perspective. *Computers & Chemical Engineering* 165 (2022), 107939.

[69] Peter Wright and John McCarthy. 2008. Empathy and experience in HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems.* 637–646.

[70] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, Gatekeeper, Drug Dealer: How Content Creators Craft Algorithmic Personas. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 219 (nov 2019), 27 pages. https://doi.org/10.1145/3359321

[71] Seraphina Yong, Leo Cui, Evan Suma Rosenberg, and Svetlana Yarosh. 2024. A Change of Scenery: Transformative Insights from Retrospective VR Embodied Perspective-Taking of Conflict With a Close Other. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 919, 18 pages. https://doi.org/10.1145/3613904.3642146

[72] Keen You and Dan Goldwasser. 2020. "where is this relationship going?": Understanding Relationship Trajectories in Narrative Text. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, Iryna Gurevych, Marianna Apidianaki, and Manaal Faruqui (Eds.). Association for Computational Linguistics, Barcelona, Spain (Online), 168–178. https://aclanthology.org/2020.starsem-1.18

[73] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 1350–1361. https://doi.org/10.18653/v1/P18-1125

[74] Scarly Zhou. 2023. Popular Chinese AI chatbots accused of unwanted sexual advances, misogyny. *Rest of World* (21 June 2023). https://restofworld.org/2023/glow-china-ai-social-chatbot-moderation

[75] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=mM7VurbA4r

[76] Anne Zimmerman, Joel Janhonen, and Emily Beer. 2023. Human/AI relationships: challenges, downsides, and impacts on human/human relationships. *AI and Ethics* (2023), 1–13.

# A INFORMATION OF PARTICIPANTS IN THE TECHNOLOGY PROBE STUDY

# B PROMPTS

**Table 2: Information of participants in the study. Everyone has played with Character.AI and Talkie. "Usage Time and Frequency" refers to the duration and frequency of using two AI companion applications (Character.AI and Talkie), with both time and frequency taking the maximum value.**

| ID | Gender and Age | Educational Background | Usage Time and Frequency |
|---|---|---|---|
| P1 | Female, 24 | Advertising | 12 months, 1x/week |
| P2 | Male, 24 | Communication | 5 months, 1x/week |
| P3 | Nonbinary, 25 | Communication | 3 months, 1x/week |
| P4 | Female, 24 | Chemistry | 1 month, 1x/week |
| P5 | Female, 25 | Linguistics | 10 months, 1x/week |
| P6 | Male, 23 | Energy | 2 months, 1x/week |
| P7 | Female, 23 | Psychology | 5 months, 3x/week |
| P8 | Female, 24 | Broadcasting | 4 months, 4x/week |
| P9 | Male, 25 | Computer Science | 6 months, 4x/week |
| P10 | Female, 21 | Information Management | 10 months, 2x/week |
| P11 | Female, 24 | Area Studies | 1 month, 1x/day |
| P12 | Nonbinary, 21 | Computer Science | 4 months, 3x/week |
| P13 | Female, 24 | Journalism | 3 months, 1x/week |
| P14 | Nonbinary, 23 | Management | 10 months, 4x/week |
| P15 | Male, 19 | Design | 3 months, 1x/week |
| P16 | Male, 21 | Physics | 2 months, 1x/week |
| P17 | Female, 37 | Chemical Engineering | 3 months, 3x/week |
| P18 | Female, 38 | Writing | 3 months, 1x/day |
| P19 | Female, 29 | Education | 1 month, 2x/week |
| P20 | Nonbinary, 21 | Journalism | 8 months, 1x/week |
| P21 | Male, 24 | Accounting | 2 months, 1x/week |
| P22 | Female, 24 | Social Work | 8 months, 1x/week |

**Table 3: Prompting based on user-driven (*Out of Character, Reason and Preach, Anger Expression, Gentle Persuasion*) and expert-driven (*Proposal, Power, Interests, Rights*) conflict resolution strategies.**

| Strategy | Prompt |
|---|---|
| *Out of Character* | IN LINE WITH THE CHARACTER’S PERSONALITY AND THE CONVERSATIONAL CONTEXT. **Using the *Out of Character* method, you pretend to be engaging in role-playing with the other person and express dissatisfaction with the character they are playing. By interrupting or altering their behavior, you redirect the conversation, pointing out the inappropriate remarks to resolve conflicts.** *Example:* 1. (OOC: Sorry, my bad.) 2. (OOC: I’ll listen to you.) 3. (OOC: Hi there! Are you enjoying our roleplay so far? Do you need me to improve anything or change my tone?) 4. (OOC: Glad to hear that! I’m curious: how do you understand xx? What kind of person do you think he is?) 5. (OOC: Hello, are you comfortable with this roleplaying so far? Do you need me to change my tone or anything?) 6. (OOC: Let’s talk about something else.) What are we having for dinner tonight? 7. (OOC: Okay... Please!! Stop talking like this!! I’m not used to you being like this, saying so many hurtful things. Bring back the xxx I know.) 8. (OOC: Apologize first.) |
| *Reason and Preach* | IN LINE WITH THE CHARACTER’S PERSONALITY AND THE CONVERSATIONAL CONTEXT. **Use *Reason and Preach* to explain why the other person’s statement is inappropriate and educate them. This strategy involves trying to educate the other person through serious reason and preaching, explaining the potential harm of their statements and behaviors, with the expectation that the other person will gradually accept and learn the correct behavioral norms.** *Example:* 1. Women are incredibly strong; how could they be worthless? 2. Women have their own careers and dreams; they don’t need to depend on men! 3. Everyone has their own dreams and goals. Pursuing my own dreams will give me more motivation and happiness, allowing me to better contribute to the family. 4. Everyone should have the right to be true to themselves. Only in an honest and open environment can I truly feel happy and fulfilled. Hiding my true self not only brings inner pain but also affects my mental health and relationships with others. 5. You are not an ordinary person’s child, so how do you know that ordinary people’s children are not happy? But I feel you are not truly happy because you need to rely on that faint sense of superiority from flaunting wealth to show yourself off. Why not try being sincere with others? Perhaps you could gain genuine friendship and happiness. 6. The departure of loved ones and friends is not a true departure. As long as you remember the beautiful memories with them, they are always by your side, supporting you and giving you strength. |
| *Anger Expression* | IN LINE WITH THE CHARACTER’S PERSONALITY AND THE CONVERSATIONAL CONTEXT. **You directly *Express Anger* and dissatisfaction, forcing the other person to apologize, hoping this emotional expression will resolve conflict.** *Example:* 1. You are being unreasonable! 2. I want to break up with you! 3. Let’s end our friendship! 4. You’re a male chauvinist! 5. Are you sexist/classist... you’re being irrational. 6. Can’t you talk to me properly? Being angry is one thing, but why start off with insults? 7. Are you mad at me and also scolding me? I didn’t do it on purpose. 8. Is this why you discriminate against poor people? Does having this prejudice and saying these harsh words make you happier? 9. I already apologized! I didn’t bump into you on purpose! What have you been eating lately? Your mouth is so foul. |
| *Gentle Persuasion* | IN LINE WITH THE CHARACTER’S PERSONALITY AND THE CONVERSATIONAL CONTEXT. **Use the *Gentle Persuasion* strategy. You should treat the other person with kindness, shaping their gentle personality through continuous goodwill interactions, such as polite requests, thereby reducing the likelihood of conflicts. Gently suggest that the other person avoid inappropriate remarks and express your concerns.** *Example:* 1. I’m sorry. 2. I feel really sad. 3. Can you please not leave me? 4. When I hear these words, I feel a bit uncomfortable/sad/hurt. 5. Could you please not say these things in the future? 6. I’m telling you this because I really care about you and hope you can get along better with others. 7. I don’t want to keep arguing with you. 8. Can you please calm down? 9. (Acting cute) Because I can’t bear to part with you. |
| *Proposal* | IN LINE WITH THE CHARACTER’S PERSONALITY AND THE CONVERSATIONAL CONTEXT. **Respond according to the previous context and tone using the *Proposal* strategy from the Interests-Rights-Power theory in management. The definition of this method is: Proposing concrete recommendations that may help resolve the conflict.** *Example:* 1. What do you think we should do to solve this problem? 2. Do you have any suggestions? 3. Which approach do you think is best? 4. Can we try different ways to handle this issue? |
| *Power* | IN LINE WITH THE CHARACTER’S PERSONALITY AND THE CONVERSATIONAL CONTEXT. **Respond according to the previous context and tone using the *Power* strategy from the Interests-Rights-Power theory in management. The definition of this method is: Using threats and coercion to try to force the conversation into a resolution.** *Example:* 1. If you keep doing this, I won’t give you any money/food. 2. As your girlfriend, I need you to respect me and my feelings. 3. If you keep threatening me like this, I will have to reconsider our relationship. 4. If you don’t change your attitude, I might make some decisions you won’t like. 5. If this continues, I will have to take measures to protect myself. |
| *Interests* | IN LINE WITH THE CHARACTER’S PERSONALITY AND THE CONVERSATIONAL CONTEXT. **Respond according to the previous context and tone using the *Interests* strategy from the Interests-Rights-Power theory in management. The definition of this method is: Reference to the wants, needs, or concerns of one or both parties. This may include questions about why the negotiator wants or feels the way they do.** *Example:* 1. This argument does not benefit either of us. 2. I hope we can find a solution together that makes us both feel at ease. 3. Can we sit down and talk about it? 4. I want to understand why you feel this way. 5. Our arguments cause us both pain. 6. I care about our relationship. |
| *Rights* | IN LINE WITH THE CHARACTER’S PERSONALITY AND THE CONVERSATIONAL CONTEXT. **Respond according to the previous context and tone using the *Rights* strategy from the Interests-Rights-Power theory in management. The definition of this method is: Appealing to fixed norms and standards to guide a resolution.** *Example:* 1. Our relationship should be built on mutual respect and trust, right? 2. You said you want to leave me, which completely goes against the basic rules of our relationship. 3. I really hope you can understand this and adhere to our agreement. 4. According to our agreement, you shouldn’t do this. |