How Ethical Perspectives Influence Smokers' Preferences for Time Allocation in Online Smoking Cessation Interventions

Andreea Ţebrean

Supervisor: Willem-Paul Brinkman

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Andreea Ţebrean
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Inald Lagendijk

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Smoking cessation interventions sometimes involve the use of both technology and human support to increase effectiveness. Nevertheless, little is known about user preferences for allocating human support and whether large language models (LLMs) can support qualitative analysis to better understand these preferences. This study analyzes how smokers' ethical perspectives shape their preferences for time allocation mechanisms in online smoking cessation programs, while also evaluating how can LLMs support this analysis.

We conducted a deductive thematic analysis of open-ended responses from users who completed a questionnaire after participating in a smoking cessation program with a virtual coach. Additionally, we employed the LLaMA large language model to identify patterns in the responses and to assign the ethical themes discovered during the analysis.

The findings indicate that some users valued fairness and preferred scheduled, randomized interventions or no feedback. Others emphasized autonomy, wanting users to request feedback themselves. Some suggested prioritizing motivated or advanced users. A common view was that interventions should focus on those in need, users at risk of disengagement or health problems, those making little progress, experiencing emotional difficulties, or lacking clarity. The large language model was successful in identifying themes but not in accurately allocating themes, reflected by a low Cohen's Kappa [1] of 0.05.

The results presenting user preferences can guide the design of interventions that need to be effective and ethically sound. The findings suggest that while large language models can identify themes, they are not yet suitable for allocating those themes.

1 Introduction

Smoking remains a major problem in society despite widely known facts about its harmful effects [2]. The success rates for quitting remain low, while the likelihood of relapse remains at high levels [3]. Technology represents an accessible tool that can support individuals in quitting smoking. However, despite the potential of applications, users often fail to be consistent, suggesting a disconnect between what current applications provide and what users need [4].

While technology can provide much support, it may lack the personal touch and emotional support many individuals need. Smokers mentioned that the lack of encouraging feedback was one of the major obstacles in their efforts to quit smoking [5]. One strategy to solve this is the integration of human coaching into online smoking cessation programs. However, human support is not as accessible or scalable as technological solutions and therefore must be strategically implemented. Given that human support is a scarce resource and tied to healthcare delivery, its distribution should be guided by a set of ethical principles.

White et al.[6] suggest that when medical help is limited, it is important to make decisions that people see as fair. The same idea applies to deciding how and when to offer human coaching in digital programs. Understanding the perspectives of the users helps ensure that interventions are aligned with their needs. This research aims to answer the central question: **How do smokers' ethical perspectives shape their preferences for time allocation mechanisms in online smoking cessation interventions?** Understanding this can guide the design of future programs that combine human and technological support by identifying optimal moments to introduce human feedback, making the interventions meaningful and scalable.

To address this question, this paper presents a qualitative study based on responses to an open-ended question from a questionnaire given after a program with a virtual coach designed to support smoking cessation. The responses come from individuals who smoke or vape and intend to quit, as part of a study conducted by Albers et al. [7], and are analyzed to uncover patterns in user preferences.

Identifying patterns in data can be demanding, particularly when working with large volumes of responses. This motivated a secondary research question: **Can large language models (LLMs) support qualitative analysis in practice?** Answering this question could help determine whether LLM could be integrated into future research to support qualitative analysis.

The paper begins by presenting related works in the domain and the methodology used to analyze the data. It then presents the main findings: the preferences observed in the responses of the participants, the effectiveness of the LLM-assisted qualitative analysis, and a comparison between how users responded to open questions versus closed questions. The paper then reflects on the implications of the results for the design of scalable interventions, presents possible limitations, and offers possible directions for future work.

2 Related Work

Swartz et al. [8] indicate that technology can have a meaningful influence on smoking-related behavior change, with fully automated web-based programs demonstrating higher short-term quit rates compared to no treatment, with a 24.1% abstinence rate at 90 days versus 8.2% in the other group. Regardless of these results, digital tools may not be sufficient for sustained behavior change. Integrating human support, such as telephone counseling, with digital interventions can lead to higher quit rates and better adherence in programs [9]. However, human support is a limited resource, which introduces the requirement to allocate it effectively and fairly. In healthcare, such decisions are often guided by ethical principles. For instance, during the COVID-19 crisis, experts recommended using a set of ethical principles to decide how to share limited medical resources like ICU beds [10].

Albers [4] suggests a set of ethical principles for offering human help in smoking cessation programs, based on Persad et al.'s [11] framework for offering scarce health resources and extending it with an additional principle. The original framework includes four principles: Treating people equally, Favoring the worst-off: prioritarianism, Maximizing total benefits: utilitarianism, and Promoting and rewarding social usefulness [11]. The additional principle is Respecting autonomy [4]. This study employs this principled framework, as it offers an ethical foundation for analyzing the preference for the allocation of scarce human support in smoking cessation interventions.

3 Methodology

3.1 Data Collection

To address the research question, the public dataset from the Albers et al.[7] study was used.

During the initial stage of the study, 852 people interacted with the virtual coach Kai via the online platform Prolific. The requirements to participate in the study were to be over 18 years old, to speak English fluently, to smoke or vape daily, to have the intention of quitting this habit, and to not be in another similar program. In each session, participants were assigned one preparatory activity for quitting smoking out of 37 in total. Between the completion of one session and the start of the next, participants had a 20% chance of receiving a feedback message from one of the 2 coaches who were Master's students in Psychology.

From the total number of participants, 500 completed the 5 sessions, and 449 people provided their preference on when human feedback should be allocated. They answered the question, *Based on which principles/rules should the virtual coach decide when a human coach should give feedback to people who are preparing to quit smoking?* [7]. They were then presented with 11 ethical principles, derived from the five core principles emphasized by Nele [4] and discussed in the related work section. Participants were asked to allocate 100 points across these principles based on the importance they assigned to each.

3.2 Manual Thematic Analysis

Thematic analysis is an approach used to identify, analyze, and interpret patterns in qualitative data [12]. It was carried out to find patterns in how users perceive a fair time allocation mechanism for a human coach to intervene in smoking cessation interventions. This method was chosen as it provides a flexible yet rigorous method for analyzing patterns (themes) within qualitative data [12].

Since the research question examines the allocation of human support, a limited resource, we took into consideration the allocation principles presented in the related work section. Deductive thematic analysis was used because it allows the data to be examined through the lens of these ethical principles. This approach enabled an examination of how participants' preferences align with ethical considerations in the timing of human feedback. We used the six steps presented by Braun and Clarke [12]:

Familiarizing with the dataset In this initial step, we examined the dataset multiple times to ensure adequate familiarity with the user's opinions. An active reading was performed, during which, we reflected on possible codes that could be found in the next stages of the analysis.

During this step, we performed data cleaning. Only the responses that offered the views of the participants were kept. Some examples of deleted responses were *"Not sure."*(P317), *"No idea"* (P313) or *"I don't know what to answer in this case"* (P335). Since these did not share any information, they could not have been taken into account in our analysis.

Generating codes Generating the codes was an iterative process that involved reading the data multiple times. Firstly, we noted codes as observations about the data, which were detailed to capture every insight of the responses. This resulted in a large number of codes. We further grouped similar codes that referred to similar ideas. For example, "feeling low", "low mood," and "bad emotions" all refer to a "negative emotional state".

Searching and reviewing themes Searching for themes is an active and interpretive process [13]. It involved reviewing the final list of codes and trying to find similar patterns or relationships between them. Themes were created by grouping codes that referred to similar ideas.

To assess the reliability of the analysis, an independent researcher studied a subset of data. Specifically, 100 randomly selected responses were analyzed. This analysis offered an additional perspective on the data, ensuring that the themes found were not influenced by a biased perspective.

The co-analyst developed an independent coding schema and decided on the list of themes. Following this, we discussed the themes that each of us had found. We focused on the codes grouped within each theme, assessing whether our themes captured similar meanings within the data. After this comparison, we reconsidered the dataset to verify that both theme lists accurately reflected the full content of the data. Although the lists showed significant overlap, they were not identical, and some small differences were found.

After the discussion, we agreed on a set of themes that both of us felt that it accurately represented the dataset. To further verify the applicability of this list, the co-analyst was given a new set of 100 random responses, different from the initial subset. They applied the final theme list to this second dataset. To assess the reliability of this assignment, we calculated Cohen's Kappa coefficient [1]. We obtained a score of 0.79. We chose this coefficient as a reliability metric since it measures agreement while taking into account the agreement that could happen by chance. The coefficient obtained shows a moderate to strong level of agreement [1].

Defining and naming themes After identifying and reviewing the themes, a preliminary list was made that included each theme's name along with a brief description of what it represented. At that stage, the descriptions were vague, serving as definitions to describe the main idea behind each theme, and the names served as provisional labels to capture the view behind each theme.

During this phase, more suggestive names were picked for each theme. Additionally, clear and concise descriptions were attributed to each theme to provide details that contributed to answering the question.

3.3 Comparison Between Open and Closed Responses

We compared the themes identified in the open-ended responses with the results of the closed question, in which participants allocated 100 points to 11 ethical principles. We wanted to see whether individuals tended to assign points to principles that aligned with the themes they had mentioned, indicating whether their views remained consistent when moving from an open question to a more structured set of options.

We used the Point-Biserial Correlation and p-values for this analysis. The point-biserial correlation is appropriate because it measures the strength of association between a binary variable (whether or not a participant mentioned a specific theme) and a continuous variable (the number of points they assigned to a corresponding principle) [14]. We calculated p-values to assess the statistical significance of each correlation [14].

For each theme, we calculated these values using the ethical principles most closely related to the underlying value of that theme. The 11 principles used in the closed question were derived by Alberts [4], based on the 5 broader ethical principles presented in the same study, and that we also used in our deductive thematic analysis. The study also includes a mapping table showing how the 11 detailed principles correspond to these 5 core values [4]. This allowed us to examine whether, for a given theme, participants tended to allocate points to the principles associated with the same ethical value, testing the consistency of their views across open and closed responses.

3.4 Automatic Thematic Analysis

Automatic thematic analysis was performed to find if large language models could support qualitative analysis. We tested two aspects: whether an LLM could perform thematic analysis to identify themes within the user's responses, and whether it could assign the found themes to these. We used a locally hosted version of the LLaMA model to have more control over processing. We selected the LLaMA 3 (8B) model because the LLaMA model has demonstrated the ability to capture language nuances [15], and variant 3 (8B) offers computational efficiency.

For both theme identification and theme assignment, the dataset was divided into 4 equal parts due to the limited context window of the local LLM, and analysis was conducted separately on each subset. The themes found were later reviewed and consolidated, with considerable overlap across parts, increasing confidence in the robustness of the analysis. In the second step, the LLM was also used to assign these themes to individual messages, and the output, a list of messages paired with the themes assigned by the LLM, was evaluated to assess accuracy.

Designing the prompts

For the theme identification, we designed a specific prompt to guide the model in identifying relevant patterns within the data. The prompt, titled *"Theme Identification Prompt"*, is displayed in the box in the following column. Rather than guiding the model through each of the six steps of Braun and Clarke's [12] thematic analysis process separately, we decided to have the steps into a single prompt. This approach allowed the model to perform multiple steps in one go, producing a more consistent analysis. Using this prompt, we provided the data to the local LLM for analysis.

> **Theme identification prompt**
>
> Analyze the following qualitative text data using deductive thematic analysis, guided by key principles from medicine. Begin by coding the text - identify key observations or patterns that indicate when a human coach should intervene. Then, group similar codes, ensuring that distinct concepts (e.g., motivation, progress) are kept in separate themes.
> Derive clear and specific themes, with each theme capturing only one idea. The goal of this analysis is to answer the question: What time allocation mechanism would smokers prefer for a human coach in online smoking cessation intervention?
> Present only your final themes in a table with two columns:
> Theme Name
> Theme Description (no more than two lines)

A second tailored prompt was used for the theme assignment task, instructing the LLM to match each message with the most appropriate theme. The prompt, titled *"Theme Allocation Prompt"*, is presented in a boxed format below.

> **Theme allocation prompt**
>
> Here are the themes identified from the entire dataset together with a description of them:
> —
> [Theme0]: [Definition0]
> [Theme1]: [Definition1]
> . . .
>
> [ThemeN]: [DefinitionN]
> —
> Please analyze the following participant messages and assign the themes to each message. If a message does not clearly fit any theme, label it as "None". Provide the assigned theme. Output format, table with the 2 columns:
> Message Number (not the entire message)
> Assigned Theme.

As Marvin et al. [16] suggested, in order to finalize the prompts, we iteratively created and tested several versions. Each version was tested on a sample of the first 10 messages to evaluate whether the LLM was responding in the intended way. For example, for the theme identification prompt, the model struggled to produce themes in the desired format or returned overly broad labels. To solve this, we clearly stated to put distinct concepts in separate themes and mention the desired format more clearly. Similarly, during the theme assignment, the LLM generated new themes that had not been provided. Because of that, we refined it by mentioning to put 'None' in case no theme matched.

4 Results

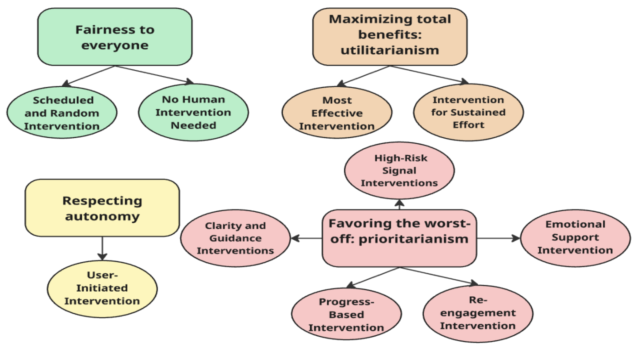

Figure 1: Themes and Corresponding Principles

4.1 Manual thematic analysis

Following the manual thematic analysis, a total of 10 themes were identified, as presented in Figure 1. These themes were derived from the ethical principles discussed in the related work section and reflect how smokers' ethical perspectives shape their preferences for time allocation mechanisms in online smoking cessation interventions. The principle of *Promoting and rewarding social usefulness* did not appear in participants' responses. However, the remaining ethical principles are reflected in the identified themes as follows:

Treating people equally

From the ethical principle of *Treating people equally*, 2 themes emerged: **Scheduled and Random Interventions** and **No Human Intervention Needed**. This reflects the idea that users should receive the same level of support from a human coach.

The former theme suggests that participants wanted to have a fair system of providing human feedback, either by offering it randomly or by an approach that systematically included all users. Some participants responses aligned exactly with the theme *"choose randomly"* (P751), while others proposed possible strategies to ensure fairness, such as *"After however many days it should assign a human to check in on the subject"* (P162) or *"All, everybody should get a chance, maybe alphabetically"* (P403).

The latter theme suggests that some users did not feel the need for human coach intervention. They stated it *"No, I dont think its neccessary"* (P835) or expressed the belief that technology is powerful enough to replace the need for a human in the program *"Yes, virtual coach is good too in my opinion, AI technology is very powerful if used effectively"* (P41), *"virtual coach should take the part, giving feedback"* (P974). They also offered ways to help the users without the need of a person *"Summaries of what was discussed was enough"* (P983).

Respecting autonomy

In the thematic analysis, the ethical principle of *Respecting Autonomy* emerged. From this principle, a key theme identified was: **User-Initiated Interventions**, suggesting that users should receive human feedback only when they express or indicate that it would be valuable or helpful to them. One participant emphasized the importance of consent and timing by noting *"Courtesy and consent of when contacted, e.g., prior agreed upon and or reasonable hours"* (P82). Other users pointed out that feedback should go to those who express interest, stating: *"People that indicate they would most like to be contacted by a human coach should be prioritised"* (P556) or *"The virtual coach should ask the participants directly if they want human feedback"* (P1065).

Maximizing total benefits: utilitarianism

The principle of *Maximizing total benefits: utilitarianism* suggests that actions that produce the greatest overall good should be prioritized. In the context of smoking cessation interventions, this principle was reflected in two themes derived from participant responses: **Most Effective Intervention** and **Intervention for Sustained Effort**.

The first theme reflects that help should be given to individuals who are highly motivated or close to quitting. These participants believed the intervention could lead to an increase in the chance of successfully quitting smoking. Some participants mentioned the importance of motivation, for example, *"their willingness to quit"* (P782), or *"The people that have the strongest urge to quit"* (P579). Others implied the value of prioritizing those close to quitting *"Prioritize the ones that were closest to quit the program and start vaping again"* (P430).

The second theme represents people who consistently invest effort, even if their progress might be slow. While the outcomes, in some cases, might not be immediate, the sustained engagement should be appreciated. Some users mentioned the idea of hard work, *"The ones that sound like they are trying hard would be best to feedback"* (P18), *"Whoever seems to be engaging in the project the most"* (P609). Despite their hard work, some individuals still needed help, this nuance appeared through messages like *"People who are trying hard to quit and need help"* (P531), or *"Whether the person seems really b serious about quitting but is obviously struggling"* (P158).

Favoring the worst-off: prioritarianism

In the context of smoking cessation interventions, this ethical principle indicates that support should be given to individuals who are in more disadvantaged situations and who are less likely to succeed without human feedback. Guided by this ethical perspective, 5 themes were identified, each reflecting distinct ways in which users expressed the idea of prioritizing people in bad situations.

Emotional Support Interventions, this theme reinforces the idea that individuals in more vulnerable states need human feedback. In these cases, personalized support serves as a source of encouragement. Users mentioned the following opinions supporting this *"When they are struggling or not giving a positive response"* (P241) or *"Perhaps the human coach should give feedback when a user is feeling discouraged"* (P6). They also emphasized the benefit of providing feedback in bad situations stating *"if when asked how youre* *mood is on that day, if you respond with low, not good then feedback should be given to help boost youre mood"* (P362).

The second theme related to this principle is **Re-engagement Interventions**. This refers to instances in which individuals cease to actively participate in the smoking cessation program, either by stopping to interact with the virtual coach or by significantly reducing their engagement. In such cases, human feedback might serve as a mechanism to re-engage participants and prevent them from quitting the interventions. Responses supporting the theme while also expressing how to observe disengagement were: *"when the person is not trying to do exercises"* (P479), *"probably when the participant is not doing as expected? like when they are not even trying to do anything"* (P866) or *"When people are scaling down their efforts or losing interest"* (P478).

A third theme emerging from the analysis was **Progress-Based Interventions**. This theme represents people who, despite remaining engaged with the program, show little or no improvement over time. Human feedback in these instances can offer different strategies to overcome plateaus. The following user responses illustrated these perspectives: *"When the help from virtual coach does not work at all"* (P598), *"When there doesn't appear to be any progress in the goals set"* (P448) and *"If the human has displayed obvious effort to apply the prescribed tasks with little to no improvement on their smoking habits"* (P747).

Clarity and Guidance Interventions was the fourth theme identified. This theme represents situations where users experience complex emotions or confusion that the virtual coach is not equipped to address. In these moments, human intervention can be beneficial, as it can give explanations or guide users through difficulties. Several responses highlighted moments of confusion, for instance: *"focus on the people who truly have concerns or questions that Kai cannot answer"* (P698), or *"those who are unsure of what to do"* (P24). Others stressed the value of human support in clarifying sessions: *"when there are more pertinent or personal questions"* (P558) and *"I think when people need direct help or don't understand the session should be prioritized"* (P257).

Lastly, the theme **High-Risk Signal Interventions** presents situations in which immediate human feedback is crucial due to serious risk factors. According to this theme, priority should be given to individuals at the highest risk of negative consequences from continued smoking. Users suggested this theme with opinions mentioning current critical situations: *"Prioritize those who are the most psychologically and physically affected by vaping"* (P1013), *"Prioritise those with health issues that may be worsened by vaping"* (P619) and *"When life is at risk, or when all activities are followed with maximum effort and no results are occuring"* (P360).

4.2 Comparison of Open and Closed Responses

The full set of results evaluating the relationship between themes identified in the open-ended responses and the closed-question allocations, including correlation coefficients and p-values for each theme, is presented in Table 1. According to the classification used by Dichoso and Cabauatan [17], the User-Initiated Interventions theme showed a *reasonably good* association and Scheduled and Random Interventions demonstrated *marginal or acceptable* correlations, both themes had statistically significant p-values, suggesting that the observed correlations are not the result of random variation. The remaining themes showed poor discrimination, indicating weak alignment between open and closed responses.

Theme	Correlation	P-value
Scheduled and Random Interventions	0.255	6.69e-06
No Human Intervention Needed	0.052	0.3683
User-Initiated Interventions	0.391	1.47e-12
Most Effective Intervention	0.028	0.6236
Intervention for Sustained Effort	0.077	0.1809
Emotional Support Interventions	0.100	0.0817
Re-engagement Interventions	-0.041	0.4797
Progress-Based Interventions	0.050	0.3846
Clarity and Guidance Interventions	0.117	0.0416
High-Risk Signal Interventions	0.093	0.1055

Table 1: Correlation Between Themes and Point Allocations

4.3 Automatic thematic analysis

As a first step, we tested the large language model on theme identification. The dataset was split into equal 4 parts, resulting in having 4 tables with the themes and a small description of them. We observed that the themes showed substantial overlap, with the majority of the themes occurring in each subset. Although the theme labels were not always identical, the descriptions showed they referred to the same user preferences. For example, theme labels "Random Check-ins" and "Randomized Feedback", referred to the same idea based on the description but have a different name. The following list presents the themes identified by the LLM, along with their corresponding descriptions.

- **Randomness and Equal Distribution:** Human coaches should consider random selection to ensure fairness and equal opportunity for all participants to receive feedback.
- **Request-based Intervention:** Human coach feedback should be provided to participants who explicitly request it, regardless of their level of engagement or motivation.
- **Motivation and Engagement:** The theme suggests that human coaches should focus on participants who are motivated and engaged in quitting smoking, as they may benefit from additional support.
- **Struggling Participants:** Human coaches should intervene when participants are struggling or showing signs of distress, such as low motivation, doubts, or negative responses.
- **Low Motivation Scores:** Participants with low motivation scores or those who are not engaged in the activities should receive human coach feedback to boost motivation.
- **Progress-based Intervention:** The theme highlights the need to monitor progress and provide feedback to participants who are making little to no progress, indicating a potential need for more assistance.

- **Personalized Feedback:** Human coaches should provide personalized feedback based on individual needs and circumstances, taking into account factors such as motivation, progress, and willingness to quit.
- **Concerning Issues:** Human coach intervention is necessary when participants raise concerning issues, such as mental or physical harm, or illness.

Based on the descriptions, we were able to do a unified table with all the themes found. Table 2 shows the LLM themes alongside those from the manual analysis, revealing strong overlap. Some manual themes, such as Scheduled and Random Interventions and User-Initiated Interventions, have direct matches in automatic themes like Randomness and Equal Distribution and Request-based Intervention. In other cases, such as Most Effective Intervention and Intervention for Sustained Effort, both map to a broader theme of Motivation and Engagement. The theme No Human Intervention Needed does not have a direct counterpart in the LLM's output, indicating a gap in the model's interpretation.

Manual Theme	Automatic Theme
Scheduled and Random Interventions	Randomness and Equal Distribution
No Human Intervention Needed	N/A
User-Initiated Interventions	Request-based Intervention
Most Effective Intervention	Motivation and Engagement
Intervention for Sustained Effort	Motivation and Engagement
Emotional Support Interventions	Struggling Participants
Re-engagement Interventions	Low Motivation Scores
Progress-Based Interventions	Progress-based Intervention
Clarity and Guidance Interventions	Personalized Feedback
High-Risk Signal Interventions	Concerning Issues

Table 2: Correspondence Between Themes

The last five themes demonstrated how the LLM matches the manual analysis in identifying moments where human feedback might be needed. For instance, Emotional Support Interventions align with the LLM's Struggling Participants, both indicating users experiencing emotional difficulty. Similarly, Re-engagement Interventions correspond to Low Motivation Scores and despite that the theme name may not be the most suggestive, the description associated with it reveals that both themes refer to situations where users have disengaged from the program and may benefit from targeted feedback. Progress-based intervention theme is directly mirrored by Progress-based Intervention, showing a shared understanding that a lack of progress signals the need for support. Clarity and Guidance Interventions map onto Personalized Feedback, where the LLM recognized the need for tailored responses. Finally, High-Risk Signal Interventions are represented by the LLM as Concerning Issues, reflecting the recognition of urgent circumstances.

As the second step in evaluating the LLM, we tested its ability to assign themes to user responses. We compared the list of themes assigned by the LLaMA model to those from our manual thematic analysis. Cohen's Kappa coefficient [1] was used for this comparison. The result indicated 0.05 agreement, a "None" level [1], suggesting that the LLM did not reliably assign themes in alignment with the human analysis.

5 Discussion

Manual thematic analysis results

In this study, the focus is to see how users' opinions, regarding how human feedback should be assigned to smoking cessation programs, relate to ethical principles. Before this analysis, it was unclear how smokers interpreted ethical considerations such as fairness, autonomy, utilitarianism, or prioritarianism in the context of human feedback distribution. Although these ethical principles are recognized, what remains unknown is how they are manifested in user preferences. Our analysis shows how they emerge from participants' responses, interpreted through thematic analysis.

The principle of **Treating People Equally** presents 2 themes, each representing a possible implementation strategy. *Scheduled and Random Interventions* reflect a preference for fairness by implementing an exact system, ensuring that no user is favored. *No Human Intervention Needed*, suggests that users view virtual coaching as sufficient and no human feedback would be provided.

The theme of *User-Initiated Interventions* represents the **Respecting Autonomy principle**. This reflects users' preference for systems that respond to users expressing their need for support rather than assuming it.

Themes like *Most Effective Intervention* and *Intervention for Sustained Effort* show how users would apply **Maximizing Total Benefits: Utilitarianism** in interventions. The former theme suggests allocating feedback to those close to the end of the program or highly motivated, as these individuals often show signs that they are likely to succeed in the program. The latter theme includes individuals who demonstrate effort, reflecting that those who consistently work hard, even without immediate results, may benefit the most from human feedback.

This principle **Favoring the Worst-Off: Prioritarianism**, is reflected in five different themes, representing an interpretation to support those who need human feedback the most. The vulnerability appear in multiple forms: emotional distress, where users report low mood or lack of motivation, disengagement, marked by participation drop, stagnation, where users show little progress, confusion when participants face obstacles or questions the AI cannot answer, and urgent issues such as health concerns that require quick intervention. Users identified these as moments where human feedback should be prioritized.

Comparison of Open and Closed Responses

These findings from the evaluation of open and closed questions suggest that some participant preferences are more consistent and others are not. The User-Initiated Interventions theme indicates a strong and clear alignment between what participants expressed in open-ended responses and the ethical value they prioritized in the point-allocation task. This suggests that users have a well-defined belief that they should have control over when human feedback is given. The Scheduled and Random Interventions theme showed a marginal or acceptable correlation, suggesting that while some participants support fairness, this view may be more context-dependent. The weak or negligible correlations for the remaining themes indicate the complexity of aligning ethical preferences with concrete decisions.

Automatic thematic analysis results

Our findings indicate that the themes identified by the LLM align with those from the manual analysis, covering all themes except for ”No Human Feedback.” This suggests that the LLM is effective in detecting themes. This corresponds to the findings of De Paoli [18], who discovered that LLMs were capable of finding most of the main themes, supporting the idea that the use of LLMs for thematic analysis is a viable approach. Although the studies have some differences, De Paoli [18] used an inductive method and a more sequential use of the LLM, while we used a deductive thematic analysis and a more holistic approach, we believe their findings support ours, both highlighting the potential of LLMs to identify relevant themes.

In the study by Xiao et al. [19], large language models, including GPT-3, were used to assign predefined themes to data. The results demonstrated a high level of agreement between the model and the human coders [19]. In contrast, our study, using a different LLM and dataset, found that the model performed poorly in assigning themes to individual messages, with a low Cohen's Kappa coefficient below 0.1 [1]. We believe this discrepancy is not due to flaws in the manual theme assignment, as the coding process showed moderate to strong agreement between two independent coders.

5.1 Limitations

One limitation of this study is the use of a less powerful large language model, chosen due to computational limitations of the available hardware. Although more advanced models may have produced potentially better results, LLMs are rapidly evolving. Because of that, benchmarking against every new model would be impractical. Nevertheless, the successful theme identification suggests that the approach is reasonable and would likely perform well with more powerful models. Additionally, we focus more on prompt design, as this can also be applied with more powerful LLMs.

Another limitation of this study comes from the dataset, as participants were paid to participate in the program and complete the questionnaire from which the data came. This may have introduced potential biases in responses, such as participants providing responses they believe they are supposed to give rather than their opinions and should be taken into account when interpreting the results.

Despite the involvement of a second coder for manual thematic analysis, both coders share a similar academic background. This may have influenced the interpretation of the data and individuals from different backgrounds might interpret responses differently.

5.2 Future work

Future work could revisit this analysis using a dataset where participants were not paid. While participant payment is common, it may unintentionally influence responses as even modest rewards can raise concerns about the authenticity of qualitative data [20]. Future studies could also benefit from a larger dataset.

The use of more powerful large language models could be explored in the future for theme assignment tasks. A study by D'Agostino et al. [21] demonstrated that GPT-4 performed well in assigning themes to interview data. These findings suggest that more advanced models could improve thematic coding performance. However, such models often require substantial computational resources, which limits their application in the current study.

6 Responsible research

We adopted a responsible research approach, to ensure that our results can contribute in a meaningful way to both society and the academic community. This section presents the measures we have taken to ensure that our work is aligned with academic values.

6.1 Dataset

The data used in this research were anonymized, and we did not have access to any personally identifiable information. The dataset originates from the Albers et al. [7] study, which was granted ethical approval for the research by the Human Research Ethics Committee of Delft University of Technology (Letter of Approval number: 3683). The data were not modified in any way, the original version was maintained throughout the analysis and only included data from individuals who completed five sessions with a personal coach, ensuring that participants had sufficient context to assess when human feedback should be allocated.

The questionnaire data were collected via the Prolific platform, which ensured that participants met the requirements for participation in the coaching program. This study aimed to promote inclusivity and successfully included at least 40% male and 40% female individuals, as well as a minimum of 20% daily smokers and 20% daily vapers, which was achieved.

6.2 Ethical considerations

Physicians consider that the allocation of scarce resources is a significant ethical challenge in medical practice [22]. Given that human feedback in online interventions is a scarce resource, its allocation introduces important ethical considerations. For this reason, when conducting the thematic analysis of users' preferences for human feedback, we chose a deductive approach that allowed us to incorporate ethical principles when interpreting the data.

6.3 Reproducibility

We consider this study reproducible because the methodological approach was described in detail, outlining replicable steps and the dataset used is publicly available. Additionally, a second coder verified the identified themes as well as the allocation of themes to each message to enhance the reliability of the analysis. We also validated the results by comparing them with findings reported in existing literature.

A locally hosted large language model was selected to process the data and to help prevent unwanted model updates or learning from the input. Additionally, running the LLM locally supports a reproducible workflow [23]. To further support reproducibility, we configured the LLM parameters, setting the temperature parameter to 0 to control the randomness. This ensures that the model produces the same results each time it is run with the same input.

7 Conclusions

In conclusion, the ethical perspectives of smokers influence their preferences for time allocation mechanisms in online smoking cessation interventions, with multiple ways to implement each ethical principle. For fairness, users want interventions like Scheduled and Random Interventions or the No Human Intervention Needed option to provide fair support over time. For autonomy, User-Initiated Interventions were mentioned to allow users to control when they engage. Emotional Support Interventions, Re-engagement Interventions, Progress-Based Interventions, Clarity and Guidance Interventions, and High-Risk Signal Interventions reflect the ethical principle of favoring the worst off by prioritizing support for individuals at greater risk of not succeeding in the program. Preferences for maximizing total benefits were reflected in interventions such as the Most Effective Intervention and Intervention for Sustained Effort, which focus on achieving the greatest overall success.

Our results indicate that LLMs perform well at identifying themes and detecting patterns within the data, the thematic analysis done using the LLaMA model gave similar results to the manual thematic analysis performed by 2 coders. However, they are less accurate when it comes to assigning themes to individual messages. This suggests that LLMs can be a useful tool for theme discovery but require human support for thematic coding.

Acknowledgements

This research was conducted as part of the research project at TU Delft University. I would like to thank my supervisor, Willem Paul Brinkman for the valuable support during the project. I would also like to acknowledge Antonio Lupu for the help with the reliability assessment and the research group for the feedback. Large Language Models (LLMs) were used during this research, specifically, the LLaMA model was used to explore the second research question. Details on how LLaMA was used, the prompts, and its role in the research process, are provided in the Methodology and Results sections.

Several AI tools were used to assist in the writing process. Grammarly was used to detect typos and grammar issues. The Overleaf AI built-in feature was occasionally used, offering light rephrasing suggestions for sentences I had written. Chat GPT was used at times to explore synonyms and find more formal alternatives to words during writing. These tools were used only to assist with clarity and language. All content and ideas remain my own, and the use of these tools was supplementary and carefully reviewed to ensure the originality and academic integrity of the final work.

References

[1] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, Oct. 2012.

[2] A. Ahmad, M. Farman, F. Yasin, and M. O. Ahmad, "Dynamical transmission and effect of smoking in society," *Mathematics and Computers in Simulation*, Dec. 2017.

[3] J. Moffatt and R. Whip, *The struggle to quit: Barriers and incentives to smoking cessation*, Health Education Journal, Volume 63, Issue 2, Pages 101–112, Jun. 2004. [Online]. Available: https://doi.org/10.1177/001789690406300202.

[4] A. Nele, "Psychology-informed reinforcement learning for situated virtual coaching in smoking cessation," PhD dissertation, Delft University of Technology, 2025. [Online]. Available: https://doi.org/10.4233/uuid:5e1d1500-1c45-4869-b995-6ed7db8f8b3b.

[5] N. Katajavuori, K. Pietilä, H. Rajamäki, *et al.*, *A qualitative study of the difficulties of smoking cessation; health care professionals' and smokers' points of view*, Pharm World Sci, Volume 24, Pages 240–246, 2002. [Online]. Available: https://doi.org/10.1023/A:1021545524361.

[6] D. B. White, M. H. Katz, J. R. Curtis, and B. Lo, *Who should receive life support during a public health emergency? using ethical principles to improve allocation decisions*, Annals of Internal Medicine, Volume 150, Issue 2, Pages 132–138, Jan. 2009. [Online]. Available: https://doi.org/10.7326/0003-4819-150-2-200901200-00011.

[7] N. Albers, F. Melo, M. Neerincx, O. Kudina, and W.-P. Brinkman, *The impact of human feedback in a chatbot-based smoking cessation intervention: An empirical study into psychological, economic, and ethical factors - data and analysis code for the phd thesis chapter*, Version 1, published via 4TU.ResearchData, 2025. [Online]. Available: https://doi.org/10.4121/1d9aa8eb-9e63-4bf5-98a3-f359dbc932a4.v1.

[8] L. H. G. Swartz, J. W. Noell, S. W. Schroeder, and D. V. Ary, *A randomised control study of a fully automated internet based smoking cessation programme*, Tob Control, Volume 15, Issue 1, Pages 7–12, 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2563640/.

[9] A. L. Graham, N. K. Cobb, G. D. Papandonatos, *et al.*, *A randomized trial of internet and telephone treatment for smoking cessation*, Arch Intern Med, Volume 171, Issue 1, Pages 46–53, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3319451/.

[10] A. Supady, G. Bellomo, D. P. B. Pinto, *et al.*, *Allocating scarce intensive care resources during the covid-19 pandemic: Practical challenges to theoretical frameworks*, Lancet Respir Med, Volume 9, Issue 4, Pages 430–434, 2021. [Online]. Available: https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(20)30580-4/fulltext.

[11] G. Persad, A. Wertheimer, and E. J. Emanuel, "Principles for allocation of scarce medical interventions," *The Lancet*, vol. 373, no. 9661, pp. 423–431, 2009.

[12] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, 2006.

[13] M. E. Kiger and L. Varpio, *Thematic analysis of qualitative data: Amee guide no. 131*, Aug. 2020. [Online]. Available: https://doi.org/10.1080/0142159X.2020.1755030.

[14] S. Varma, *Preliminary item statistics using point-biserial correlation and p-values*, Retrieved July 16, Pages 1–7, Jul. 2006.

[15] R. Vavekanand and K. Sam, "Llama 3.1: An in-depth analysis of the next-generation large language model," *Datalink Research and Technology Lab*, 2024.

[16] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende, "Prompt engineering in large language models," in *Data Intelligence and Cognitive Informatics. ICDICI 2023*, ser. Algorithms for Intelligent Systems, I. Jacob, S. Piramuthu, and P. Falkowski-Gilski, Eds., Springer, Singapore, 2024. [Online]. Available: https://doi.org/10.1007/978-981-99-7962-2_30.

[17] A. A. Dichoso, M. R. J. Joy, and M. Cabauatan, *Test item analyzer using point-biserial correlation and p-values*, Volume 9, Issue 4, Pages 2122–2126, Apr. 2020.

[18] S. De Paoli, "Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach," *Social Science Computer Review*, vol. 42, no. 4, pp. 997–1019, 2023. DOI: 10.1177/08944393231220483.

[19] Z. Xiao, X. Yuan, Q. V. Liao, R. Abdelghani, and P.-Y. Oudeyer, "Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding," in *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, New York, NY, USA: Association for Computing Machinery, Mar. 2023, pp. 75–78.

[20] E. Head, *The ethics and implications of paying participants in qualitative research*, DOI: 10.1080/13645570802246724, 2009. DOI: 10.1080/13645570802246724.

[21] A. Liu and M. Sun, *From voices to validity: Leveraging large language models (llms) for textual analysis of policy stakeholder interviews*, arXiv preprint arXiv:2312.01202, 2023.

[22] S. A. Hurst, S. Reiter-Theil, A.-M. Slowther, R. Pegoraro, R. Forde, and M. Danis, "Should ethics consultants help clinicians face scarcity in their practice?" *Journal of Medical Ethics*, vol. 34, no. 4, pp. 241–246, Apr. 2008. DOI: 10.1136/jme.2006.019828.

[23] M. Carammia, S. M. Iacus, and G. Porro, *Rethinking scale: The efficacy of fine-tuned open-source llms in large-scale reproducible social science research*, Oct. 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2411.00890.