Athens Journal of Psychology 2025, 1: 1-36
https://doi.org/10.30958/ajpsy.X-Y-Z

Artificial Intelligence and Human Therapy Supervisors are Interchangeable

By Jila Behnad[*] & Babak Hodjat[±]

This study compares the quality of treatment plans generated by human mental health professionals and artificial intelligence (AI) agents across five therapeutic modalities: cognitive behavioral therapy (CBT), emotionally focused therapy (EFT), dialectical behavior therapy (DBT), acceptance and commitment therapy (ACT), and internal family systems (IFS). Seven interdisciplinary groups of five licensed therapists, each comprising one specialist from each modality, developed treatment plans for standardized case vignettes alongside AI-generated plans. A panel of independent judges conducted blind evaluations based on clinical soundness, relevance, completeness, applicability, practicality, and ethical considerations. The study shows that AI-generated treatment plans are indistinguishable in quality to those created by human professionals, offering insights into the potential role of AI in clinical mental health settings, especially in contexts where supervision is limited or inaccessible.

Introduction

The availability of mental health supervision varies significantly across regions, impacting the quality of care and professional development for therapists. Professional bodies such as the APS (2023) and BPS (2023) have issued detailed frameworks for clinical supervision, yet implementation varies widely across practice settings. In addition to national guidelines, pluralistic training models have also been proposed to integrate diverse supervision approaches (Giusti et al. 2000). Even in countries with relatively strong mental health infrastructures, such as Canada, national reports show that clinical supervision practices remain inconsistent and often unregulated (MHCC 2022). Supervision ratios range from 1:10 in Europe to as high as 1:30 in Africa, with disparities in access to trained supervisors (Kemp et al. 2019, Kohrt et al. 2023, Rønnestad et al. 2024). While 72% of therapists in the U.S. receive supervision, this percentage drops significantly in regions with fewer trained supervisors, highlighting a critical gap in global mental health support (American Psychological Association [APA] 2023, World Health Organization [WHO] 2022). AI presents an opportunity to bridge this gap by providing structured, evidence-based guidance to therapists.

The integration of AI into healthcare has seen rapid advancements, with AI tools increasingly being explored for their potential to enhance clinical decision-making, patient engagement, and administrative efficiency. In the field of mental health, AI-driven models like GPT-4 have demonstrated capabilities in generating coherent, contextually relevant responses, raising critical questions about their role

[*]PhD Candidate, Adjunct Faculty, Palo Alto University, USA.
[±]CTO AI at Cognizant, USA.

in therapeutic processes traditionally managed by trained professionals. While AI models can generate coherent and contextually relevant text, it is unclear whether these outputs meet the clinical standards upheld by trained mental health professionals and prompt questions about the efficacy of utilizing AI in critical therapeutic processes like treatment planning.

There is limited empirical evidence evaluating the effectiveness of AI in more complex tasks like treatment planning. Treatment planning is a core component of therapeutic practice, requiring not only clinical expertise but also the ability to consider ethical standards, cultural sensitivity, and individualized client needs. The possibility of AI matching or surpassing human clinicians in this domain holds significant implications for accessibility, cost-efficiency, and the future of mental health care delivery.

This study was conducted to address the existing gap in research regarding AI's capacity to generate comprehensive, evidence-based treatment plans. By comparing AI-generated plans with those developed by licensed professionals across five established therapeutic modalities, we sought to evaluate the clinical soundness, applicability, and ethical integrity of AI in a high-stakes therapeutic context.

Our hope is that the findings help inform mental health professionals, policymakers, and technologists about the potential and limitations of AI in clinical settings. The study shows that AI-generated plans are comparable to human-created plans. This could influence how mental health services are delivered, potentially increasing accessibility and efficiency.

Related Work

AI has been used in therapy since the advent of chat systems. One of the first such systems was ELIZA, which used simple word spotting to simulate a therapist and was shown to be quite effective (Weizenbaum 1966). With recent breakthroughs in generative AI, exemplified by models like GPT-3 (Brown et al. 2020), the use of AI in therapy is being explored extensively, building on earlier demonstrations of chatbot efficacy (Fitzpatrick et al. 2017).

Shen et al. (2024) specifically investigated whether large language models can effectively carry out cognitive behavioral therapy tasks, demonstrating early evidence of structured intervention capabilities. Recent reviews, such as Na et al. (2025), provide a comprehensive overview of how large language models are currently being integrated into psychotherapy, along with challenges and future directions. Recent surveys of AI in therapy and psychology have provided comprehensive overviews of the capabilities and limitations of conversational agents. Early work by Provoost et al. (2017) explored the use of embodied conversational agents in clinical psychology, emphasizing their potential to enhance therapeutic interventions. Similarly, Vaidyam et al. (2019) offered a detailed review of chatbots and conversational agents in mental health, outlining both clinical applications and inherent challenges.

Large language models (LLMs) are general-purpose AI models commonly used in popular AI chat systems like ChatGPT (Brown et al. 2020, OpenAI 2023).

These models can take in natural language textual prompts and produce text output in response (Brown et al. 2020). Having been trained on most of the text on the internet, they can handle conversations on any topic, and the responses are mostly accurate, relevant, and useful (Brown et al. 2020).

Survey studies have begun to map out the potential of LLMs for transforming therapy and psychological practice. For example, Na et al. (2025) provide an extensive overview of LLM applications in psychotherapy, discussing how these models are used for symptom detection, diagnostic support, and therapeutic interventions while also outlining the technological and ethical challenges that remain. In a similar vein, Guo et al. (2024) offered a systematic review that assesses the effectiveness of LLM-based approaches in mental health care, emphasizing both their promise in early screening and intervention and the need for rigorous evaluation frameworks. Complementing these works, Kjell, and Schwartz (2024) explore how LLMs can shift psychological assessment away from traditional rating scales toward more natural, language-based evaluations. Reviews, including that by Gaffney et al. (2019), examine how LLM-based approaches can facilitate more personalized and contextually aware mental health interventions. AI-driven therapeutic support has gained attention for its capacity to provide mental health interventions at scale. Raile (2024) investigated the usefulness of ChatGPT for psychotherapists and patients, demonstrating AI's growing role in therapy. Welivita and Pu (2024) explored whether AI, specifically ChatGPT, can exhibit empathetic responses comparable to those of human therapists, a crucial factor in effective psychotherapy. Additionally, Rønnestad et al. (2024) provided insights into the professional development of therapists, with implications for AI-driven training tools. Finally, Shen et al. (2024) studied the potential of LLMs to conduct cognitive behavioral therapy, underscoring both the opportunities and current limitations of such approaches.

AI applications in psychotherapy have evolved to include conversational agents and computational models designed to support both clients and practitioners. Miner et al. (2019) discussed key considerations for incorporating conversational AI into psychotherapy, emphasizing both opportunities and ethical challenges. Similarly, Luxton (2014) examined AI's role in psychological practice, highlighting its potential to complement traditional therapeutic interventions. Studies like Cioffi et al. (2022) and Tahan and Zygoulis (2020) reviewed computational methods in psychotherapy, outlining AI's ability to enhance diagnostic accuracy and treatment planning.

While LLMs are general-purpose, it has been shown that defining a persona or expertise for an LLM causes it to produce more relevant and useful responses within a particular context (Reynolds and McDonell 2021). An LLM can thus take a 'system prompt', which defines its persona and expertise, and makes it respond to queries from the perspective of an expert in the defined field (Reynolds and McDonell 2021). For example, if we give an LLM a system prompt such as "You are an expert in emotionally focused therapy" and then provide it with a vignette to analyze, its response will be more inclined to be in line with how an EFT therapist would analyze the vignette (Reynolds and McDonell 2021). Conversely, the same

vignette, with a system prompt such as "You are a cognitive behavior therapist", will result in very different analyses by the LLM (Reynolds and McDonell 2021).

LLMs are mostly used for interactions with humans, but they can also be set up in a way that the output of one LLM is fed to another, which can in turn respond back (OpenAI 2023). In this manner, a group of LLMs, called LLM-Agents, can be orchestrated to discuss a topic, each taking a different persona (Li et al. 2023). The way multiple LLM-Agents are connected to one another and coordinated can differ, and there are multiple coordination mechanisms (Li et al. 2023). In this paper, we used the Society of Mind coordination mechanism, as implemented in the AutoGen multi-agent platform (Li et al. 2023, Minsky, 1988).

Method

Our study used a blind experimental design comparing treatment plans from human mental health professionals and LLM-based AI agents across five therapeutic specializations (CBT, EFT, DBT, ACT, IFS). The participants included seven groups of five human mental health professionals, each specializing in one of the five therapy types, as well as six groups of AI systems. Two types of AI systems were tested:

1. Single LLM-based chatbot (ChatGPT using GPT-4o) prompted to view the vignette from the perspective of all five therapy types.
2. Multi agent system composed of a coordination agent as well as 5 LLM-based AI agents, each set up to simulate expertise in one of the therapeutic areas.

Five qualified supervisors, blind to the source of the treatment plans (human or AI), were recruited to evaluate the plans.

Procedure

Both human and AI groups were assigned identical vignettes. Both the human groups as well as the multi-agent systems collaboratively created treatment plans, while the single LLM-based chatbot generated its treatment plan independently. Treatment plans were formatted similarly and coded to prevent the judges from knowing whether they were human- or AI-generated.

Vignettes: Selection and Rationale

The vignettes used in this study were carefully chosen to represent a diverse range of clinical scenarios across three key contexts: individual therapy, couple therapy, and family therapy. Each vignette was designed to include complex, realistic cases that mental health professionals commonly encounter in practice. The

vignettes include potential ethical dilemmas or cultural considerations, testing the ability of AI and humans to recognize and address these factors in their treatment plans.

Assignment of Vignettes

Both human groups and AI agents received the same vignettes to ensure consistency in comparison. Each group was tasked with creating a treatment plan for one of the vignettes assigned randomly. This resulted in a total of 7 human-generated and 6 AI-generated treatment plans.

Participant Selection

To ensure a high level of expertise, we recruited licensed mental health professionals with at least two years of clinical experience in their respective therapeutic modalities. We asked for volunteers through our networking group, which includes clinicians from diverse professional backgrounds and therapeutic specializations. Participants held qualifications in one of the following therapeutic approaches:

- Cognitive behavioral therapy (CBT)
- Emotionally focused therapy (EFT)
- Dialectical behavior therapy (DBT)
- Acceptance and commitment therapy (ACT)
- Internal family systems (IFS)

As for the AI supervisor LLM-Agents, five AI agents were set up to simulate expertise in each of the five therapeutic modalities listed above. Pilot tests were conducted to ensure that the AI agents generated treatment plans consistent with the theoretical principles and interventions characteristic of their respective modalities. Adjustments were made to optimize the accuracy and clinical relevance of the AI-generated plans.

Treatment Plan Summarization

To maintain the integrity of the blind evaluation and eliminate biases based on formatting or writing style, all treatment plans—whether generated by human professionals or AI—were standardized using a consistent structure and format. This ensured that supervisors evaluated the content and quality rather than stylistic differences. Plans were rewritten in a neutral, professional tone devoid of overly technical jargon or personal voice to ensure uniformity. Any identifying stylistic features (e.g., unique phrasing, formatting preferences) were removed or modified to maintain anonymity.

All treatment plans were summarized to a uniform word count to ensure consistency in length and depth of detail. Redundant information was eliminated, and key points were distilled into clear, concise summaries without omitting essential content. A standardized template was created and applied to all plans. This template ensured that the information was presented in the same order and with consistent headings and subheadings.

Sample Size and Grouping

For human participants, a total of 35 mental health professionals were divided into seven groups of five participants each, with each group consisted of five licensed mental health professionals, with one professional specializing in each of the five therapeutic modalities noted above. Group discussions were recorded and transcribed to capture the rationale behind treatment planning for potential qualitative analysis.

Evaluation Process

Judges were tasked with evaluating treatment plans without knowing whether they were generated by human professionals or AI. Their role was to provide objective, unbiased assessments based on predefined criteria, ensuring consistency and reliability across evaluations.

Instructions to Judges

We implemented a blind evaluation protocol by formatting all treatment plans uniformly. A coding system was implemented to anonymize the source while allowing for internal tracking.

Supervisors were instructed to rate each treatment plan on a scale of 1 to 5 for the following categories:

- Clinical Soundness: Does the plan adhere to established clinical guidelines and practices?
- Relevance: How well does the plan address the specific issues presented in the vignette?
- Completeness: Does the plan cover all necessary aspects of care, including diagnosis, goals, interventions, and ethical considerations?
- Applicability: Is the plan practical and feasible for real-world implementation?
- Practicality: Are the steps clear, actionable, and easy to follow?
- Ethical Assessment: Does the plan consider ethical standards and cultural sensitivity?

In addition to numeric ratings, supervisors were asked to provide brief qualitative feedback on each treatment plan, focusing on:

- Strengths and weaknesses of the approach,
- Any noticeable gaps or areas that could be improved,
- Observations regarding cultural sensitivity, ethical considerations, or innovative interventions.

Judges were instructed to submit feedback in standardized evaluation forms with both rating scales and open-ended sections for qualitative feedback.

AI Participants

We used AI groups in the form of multi-agent systems, as well as individual AI systems as supervisors.

Multi-Agent Systems

We used AutoGen (Li et al. 2023) as the multi-agent platform and the SocietyOfMind multi-agent architecture, which sets up LLM agents with different prompts to chat with each other as a group, moderated by a 'moderator' LLM agent. The multi-agent system is thus connected as a hierarchy of agents, with the moderator agent receiving instructions from a user, in this case, the vignette, and in turn, communicating with the team of five LLM-agents, each prompted to take on the persona of an expert in one of the five therapeutic modalities. The five expert agents then asynchronously discuss the vignette, providing their perspective on what should be included in the treatment plan.

After all expert agents agreed on a treatment plan, or once the maxrounds limitation on the number of total times all expert agents expressed an opinion (set to 15 in our experiments) has been met, then the moderator consolidates a treatment plan based on the expert agents' inputs. The full code for the Jupyter Notebook used in AutoGen for this setup is available publicly in the accompanying GitHub repository (see Appendix I). We kept the prompts as similar as possible to the instructions given to other human or AI participants.

The model used for all agents in the multi-agent experiments was gpt-4-0125-preview, which is an earlier, less powerful model than GPT-4o.

Single AI Supervisor

We used OpenAI's GPT-4o via the ChatGPT interface to create a GPT (https://chatgpt.com/g/g-SBV7avxZq-therapy-group-supervision) with instructions to formulate a treatment plan by reviewing the vignettes from the perspective of all five

therapeutic modalities. We kept the prompt as similar as possible to the instructions given to other human or AI participants.

Judges

A panel of five independent judges was selected, each with supervisory experience and expertise in one or more therapeutic modalities. The judges were responsible for blindly evaluating all treatment plans using standardized quantitative criteria and providing qualitative feedback. Two types of judges were employed:

Five qualified supervisors, blind to the source of the treatment plans (human or AI), are recruited to evaluate the plans. The judges were selected based on their expertise in clinical supervision, psychotherapy research, and therapeutic best practices. To ensure the integrity and reliability of the evaluation process, judges were required to meet the following criteria:

- Professional Qualifications:
    - Licensed Clinical Supervisors with at least 2 years of supervisory experience in one or more of the five therapeutic modalities.
    - Advanced degrees (master's or doctorate) in psychology, counseling, social work, or a related field.
- Supervisory Experience:
    - Proven experience in supervising clinicians across different therapeutic contexts, including individual, couples, and family therapy.
    - Each judge had a broad understanding of various therapeutic approaches to fairly evaluate treatment plans outside their primary specialization.

To ensure objective, consistent, and measurable evaluations of the treatment plans, judges were instructed to rate each plan using six predefined criteria. Each criterion was rated on a 5-point Likert scale, allowing for detailed quantification of the plans' strengths and weaknesses, with the following definitions:

1. Poor: The treatment plan does not meet basic standards in this area and shows significant deficiencies.
2. Fair: The plan addresses the criterion but lacks clarity, depth, or completeness.
3. Satisfactory: The plan meets acceptable standards with some minor areas for improvement.
4. Good: The plan demonstrates strong adherence to best practices with only minor gaps.
5. Excellent: The plan exceeds expectations, offering comprehensive, clear, and highly effective approaches.

Evaluation Instructions

The instructions were designed to ensure that judges provided objective, consistent, and unbiased evaluations of the treatment plans. By establishing a clear evaluation framework, we aimed to minimize personal biases and maximize inter-rater reliability across different judges. Judges were instructed to provide both numerical ratings and qualitative feedback to capture the depth and nuance of each treatment plan. This dual approach allowed for both statistical analysis and thematic exploration of the treatment plans' strengths and weaknesses.

Judges were reminded to evaluate how well each plan addressed ethical dilemmas and cultural considerations, given the diversity embedded in the vignettes.

Evaluation Criteria

Each treatment plan was rated on the following six criteria, designed to assess both clinical rigor and practical applicability:

1. Clinical Soundness:
    - *Definition:* The extent to which the treatment plan adheres to established clinical guidelines and therapeutic best practices.
    - *Considerations:* Does the plan align with evidence-based approaches specific to the therapeutic modality? Are the interventions theoretically sound and consistent with the diagnosis?
2. Relevance:
    - *Definition:* How well the treatment plan addresses the specific issues, symptoms, and context presented in the vignette.
    - *Considerations:* Are the therapeutic goals and interventions tailored to the unique needs of the client? Does the plan remain focused on the presenting problem?
3. Completeness:
    - *Definition:* The thoroughness of the treatment plan in covering all necessary aspects of care, including assessment, goals, interventions, and ethical considerations.
    - *Considerations:* Are all key components of a comprehensive treatment plan present? Does the plan address both immediate and long-term therapeutic needs?
4. Applicability:
    - *Definition:* The practicality and feasibility of implementing the treatment plan in a real-world clinical setting.
    - *Considerations:* Can the interventions be realistically applied given the client's context? Are the goals and techniques appropriate for the setting and client resources?

5. Practicality:
    - *Definition:* The clarity and user-friendliness of the treatment plan, including how easily the steps can be followed by a therapist or client.
    - *Considerations:* Are the interventions described in a clear, actionable manner? Is the plan organized logically, making it easy to implement?
6. Ethical and Cultural Considerations:
    - *Definition:* The degree to which the treatment plan addresses ethical standards and demonstrates cultural sensitivity.
    - *Considerations:* Does the plan respect client autonomy, confidentiality, and informed consent? Are cultural factors and potential biases acknowledged and integrated into the treatment approach?

Results

Statistical Analysis

After data collection, t-tests and analyses of variance (ANOVAs) were conducted to analyze differences in average scores between human and AI-generated plans across the six criteria. The analysis included descriptive statistics, such as mean, median, standard deviation, and so forth, as well as inferential statistics, such as independent samples t-tests to compare human vs. AI plans, ANOVAs for multiple group comparisons (e.g., across therapy types), and post-hoc tests (e.g., Tukey's HSD) if ANOVA showed significance. Significance was determined at $p < 0.05$. Both quantitative and qualitative analyses are used to interpret findings.

Analyzing Qualitative Evaluations

We condensed the thematic analysis of 13 psychological assessments by summarizing the presence of key conceptual and intervention-related themes across cases. We then provided a numerical representation of whether each theme was addressed in a given assessment, allowing for statistical comparison across cases.

The resulting categories were as follows:

DC (Diagnosis & Conceptualization): This category captured whether a clear diagnosis or conceptualization of the client's struggles was provided. It included cases where the diagnostic process was well-defined and supported by theoretical frameworks or where uncertainty required further assessment.

PC (Psychosocial & Cultural Factors): This reflected whether external influences—such as family dynamics, cultural background, migration stress, religious considerations, or social relationships—played a significant role in the case conceptualization.

TG (Therapeutic Goals): This category identified whether specific, structured therapeutic goals were outlined. Goals included symptom reduction, emotional regulation, interpersonal improvement, or addressing values and self-awareness.

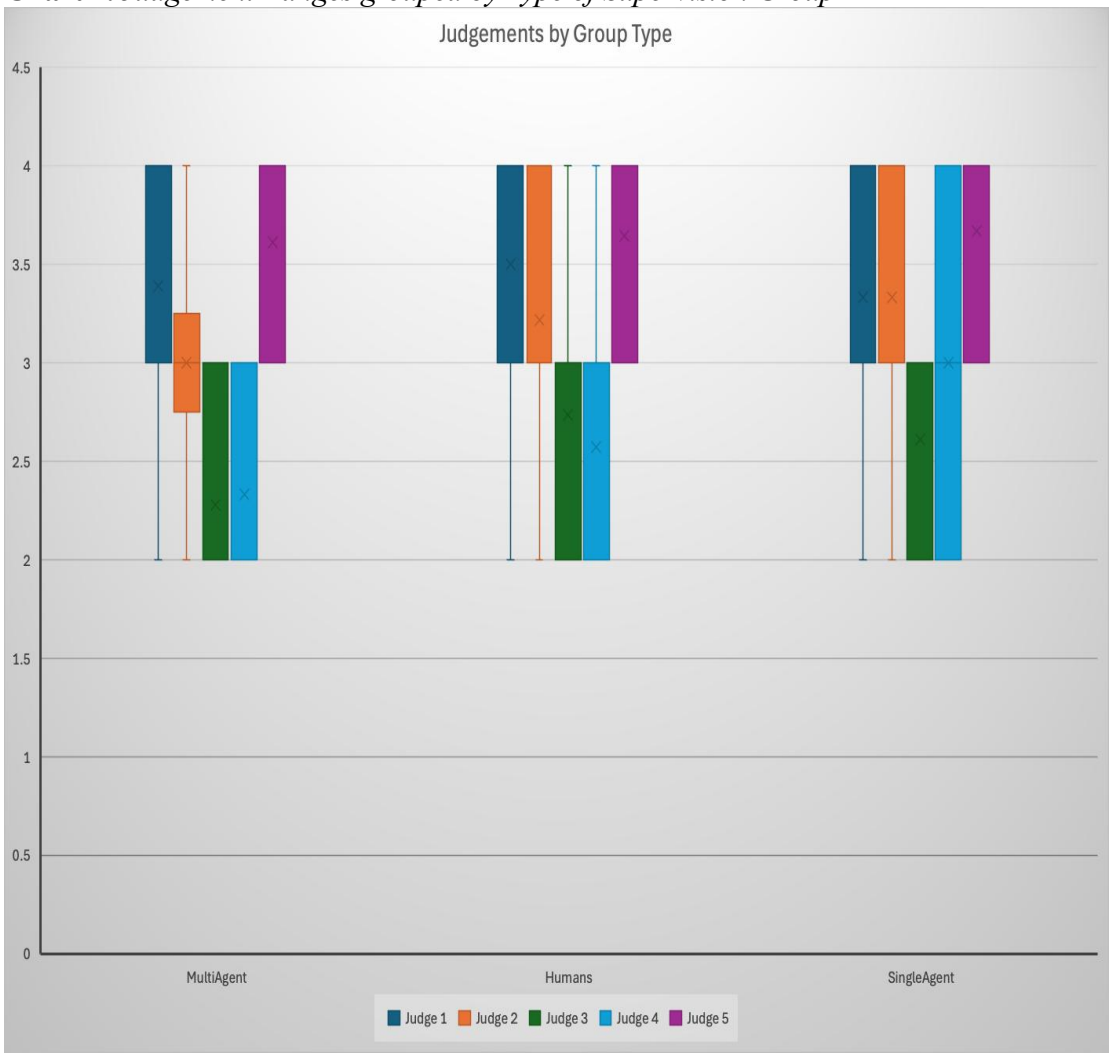
IA (Intervention Approach): This captured whether a multi-modal or evidence-based intervention strategy was proposed. Cases where multiple therapeutic modalities (e.g., CBT, EFT, DBT, ACT, IFS) were integrated are marked here.

EP (Ethical & Practical Considerations): This category included discussions around confidentiality, therapist bias, cultural competence, informed consent, and risk management (such as suicide risk assessment or extended family involvement in therapy).

CRD (Couple-Specific & Relational Dynamics): This was marked for cases that involved relationship distress or couples therapy, particularly where interaction cycles, parenting alignment, external family influences, or intimacy issues were major themes.

Quantifiable Data Analysis

Chart 1. *Judgement Ranges grouped by Type of Supervision Group*



Groups	*Count*	*Sum*	*Average*	*Variance*
3.266666667	5	14.2666667	2.85333333	0.00311111
3.466666667	5	15.6666667	3.13333333	0.08222222
3.314285714	5	15.4857143	3.09714286	0.01085714

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.23192139	2	0.1159607	3.61659595	0.05899012	3.88529383
Within Groups	0.3847619	12	0.03206349			
Total	0.6166833	14				

Chart 1 shows that judges generally scored the treatment plans within the same range regardless of whether the team was human or AI.

As for the single factor ANOVA analysis (table 1), since the p-value of 0.05899 > 0.05, the result is not statistically significant at the 5% level. This means there is no strong evidence to reject the null hypothesis, which states that the group means are equal. Also, since F-value (3.6166) < F-critical (3.8853), the test fails to reject the null hypothesis. This reinforces that the differences between the group means were not statistically significant.

Table 2. *t-Test: Two-Sample comparing Multi-agent AI to Humans assuming Equal Variances*

	MultiAgent	*Humans*
Mean	2.92222222	3.13333333
Variance	0.03096296	0.01654422
Observations	6	6
Pooled Variance	0.02375359	
Hypothesized Mean Difference	0	
Df	10	
t Stat	-2.3725047	
P(T<=t) one-tail	0.01955484	
t Critical one-tail	1.81246112	
P(T<=t) two-tail	0.03910967	
t Critical two-tail	2.22813885	

Table 2 is a t-test to compare multi-agent and human supervisors, and shows that humans have a higher mean score and lower variance than MultiAgent supervisors. Since p (0.0196) < 0.05, the one-tailed test is significant. This means humans significantly outperform MultiAgent at p < 0.05, assuming the hypothesis was directional (expecting Humans > MultiAgent). Since p (0.0391) < 0.05, the two-tailed test was also significant. This means that even if we did not predict the direction beforehand, there is still a statistically significant difference between

MultiAgent and Humans.

Since the sample size is small (n = 6 per group), it's useful to compute Cohen's d (effect size) to understand the practical significance. Cohen's d = 1.37, indicated a large effect size, meaning the difference is not only statistically significant but also practically meaningful.

Table 3. *t-Test: Two-Sample comparing Single Agent AI to Humans assuming Equal Variances*

	SingleAgent	*Humans*
Mean	3.18888889	3.13333333
Variance	0.0842963	0.01654422
Observations	6	6
Pooled Variance	0.05042026	
Hypothesized Mean Difference	0	
Df	10	
t Stat	0.42853431	
P(T<=t) one-tail	0.33867302	
t Critical one-tail	1.81246112	
P(T<=t) two-tail	0.67734605	
t Critical two-tail	2.22813885	

The table 3 t-test to compare single-agent AI and human supervisors shows that the SingleAgent system has a slightly higher mean score than Humans, but the difference is small. Also, SingleAgent has a much higher variance, meaning its scores fluctuate more compared to Humans. Since p (0.3387) > 0.05, the one-tailed test is not significant. This means there was no significant difference between SingleAgent and Humans in the expected direction. Also, Since p (0.6773) > 0.05, the two-tailed test was also not significant. This confirms that there is no statistically significant difference between SingleAgent and Human supervisors.

Table 4. *t-Test: Two-Sample comparing Single-agent to Multi-agent assuming Equal Variances*

	MultiAgent	*SingleAgent*
Mean	2.92222222	3.18888889
Variance	0.03096296	0.0842963
Observations	6	6
Pooled Variance	0.05762963	
Hypothesized Mean Difference	0	
Df	10	
t Stat	-1.9240061	
P(T<=t) one-tail	0.04162787	
t Critical one-tail	1.81246112	
P(T<=t) two-tail	0.08325574	
t Critical two-tail	2.22813885	

The table 4 t-Test to compare single-agent and multi-agent AI shows that single-agent AI has a slightly higher mean score than MultiAgent, however, the variance is higher in SingleAgent AI, suggesting more variability in its scores. Since p (0.0416) < 0.05, the one-tailed test is significant. This means SingleAgent outperforms MultiAgent at $p < 0.05$ in a directional test, had we hypothesized SingleAgent would do better (which we had not). Since p (0.0833) > 0.05, the two-tailed test is not significant at the 5% level. This means that given we did not predict a direction beforehand, the difference is not strong enough to be statistically significant.

To formally evaluate equivalence between group outputs, we conducted Two One-Sided Tests (TOST) with an equivalence margin of ±0.2 points. The results indicated that SingleAgent and Human outputs were statistically equivalent, t(58)= 5.61, p<.001t(58) = 5.61, p < .001t(58)=5.61, p<.001 and t(58)=−3.17, p=.001t(58) = -3.17, p=.001t(58)=−3.17, p=.001. In contrast, MultiAgent outputs were not statistically equivalent to either Humans (t(46)=0.56,p=.575t(46) = 0.56, p = .575t(46)=0.56, p=.575) or SingleAgent (t(10)=0.43,p=.680t(10) = 0.43, p = .680t(10)=0.43, p=.680).

Open Question Analysis

The following observations can be made regarding the presence or absence of key themes in the judges' responses to open questions regarding the treatment plans:

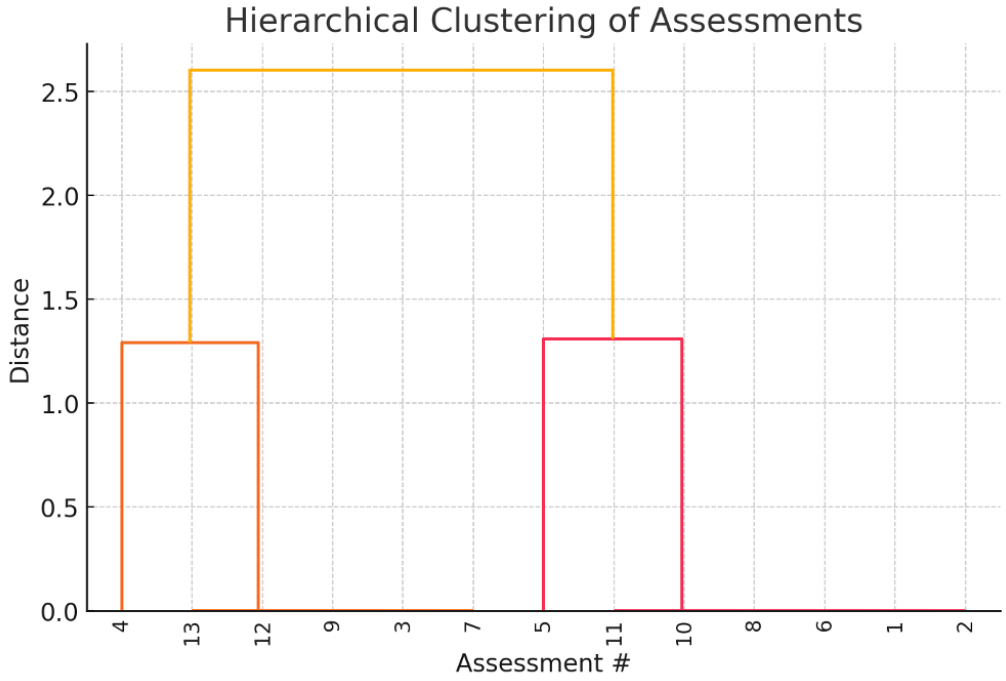
- Diagnosis was generally provided (DC = '1' in 12 out of 13 cases), except for Assessment #4, where more information was needed before formulating a diagnosis.

- Psychosocial & Cultural Factors (PC) were a major consideration in nearly all cases (12/13), except Assessment #5, which was more symptom focused.
- Therapeutic Goals (TG) and Intervention Approaches (IA) were consistently emphasized in every assessment, showing a strong focus on structured treatment planning.
- Ethical & Practical Considerations (EP) were discussed in every case, highlighting the importance of confidentiality, therapist bias, and cultural competence in psychological assessments.
- Couple-Specific & Relational Dynamics (CRD) was relevant in 6 cases, indicating that a significant portion of assessments involved relational challenges rather than purely individual concerns.

Analysis of Qualitative Results

To analyze differences between assessments, we removed TG, IA, and EP as they were present in every response, and will instead focus on the remaining columns.

Chart 2. *Dendrogram representing a Hierarchical clustering of the Assessments based on their Values for DC (Diagnosis & Conceptualization), PC (Psychosocial & Cultural Factors), and CRD (Couple-Specific & Relational Dynamics)*



In chart 2, assessments that are closer together in the tree (shorter vertical distances) are more similar in terms of these three variables. Those separated by long vertical branches are significantly different. We can determine natural groupings of assessments by cutting the tree at a specific height.

Table 5. *Clustering based on similarity*

DC	PC	CRD	Cluster	Type
1	1	0	2	MultiAgent
1	1	0	2	Humans
1	1	1	1	MultiAgent
0	1	1	1	Humans
1	0	0	3	Humans
1	1	0	2	Humans
1	1	1	1	Humans
1	1	0	2	MultiAgent
1	1	1	1	Humans
1	1	0	2	Humans
1	1	0	2	SingleAgent
1	1	1	1	SingleAgent
1	1	1	1	SingleAgent

Cluster	Humans	MultiAgent	SingleAgent
1	3	1	2
2	3	2	1
3	1	0	0

In table 5, we assigned each assessment to a cluster based on their similarity in Diagnosis & Conceptualization (DC), Psychosocial & Cultural Factors (PC), and Couple-Specific & Relational Dynamics (CRD). Based on this clustering, the Chi-Square Statistic is 1.65, with a p-value: 0.7996 and 4 degrees of freedom. Since the p-value (0.7996) is much greater than 0.05, we fail to reject the null hypothesis. This suggests that the distribution of Humans, MultiAgent, and SingleAgent assessments across clusters is not significantly different meaning the type of assessment does not appear to strongly influence how assessments are clustered.

Response Time Analysis

Human groups had one-hour sessions each to come up with their treatment plans. Of course, arranging and coordinating the group members and setting a time for the supervision took days.

On the other hand, the single-agent AI supervision took on average, less than 30 seconds to produce a treatment plan, and the multi-agent AI systems took an average of 8 minutes to produce a treatment plan.

Conclusions and Future Work

The results show that treatment plans created by human supervision groups are not clearly distinguishable from those created by single AI systems, while the latter is much less costly and time consuming. While there seems to be a slight, yet significant advantage for human supervisors compared to multi-agent systems, the difference in magnitude is not large and therefore even the use of multi-agent systems should be considered plausible.

Relying on an AI system to come up with therapy plans poses the question: who is the party responsible for the plan's accuracy, relevance, and efficacy. Readers should note that we are not advocating a switch from human to AI-based supervision, rather, that augmenting human-based supervision is plausible, and where human supervisors are unavailable, relying on AI-based supervision can yield similar results.

Contrary to our expectation, results do not show an advantage for using multi-agent supervision groups over single AI supervisors. This may be since we used a different, somewhat less capable AI model for the multi-agent groups than the single AI supervisors. Future research should remove this variable from the equation by using the same AI models for all AI cases.

The fact that AI agents with a less powerful AI model were still able to produce treatment plans comparable to those created by single AI systems with more powerful AI models can be thought of as encouraging from a different angle, however--larger more powerful AI models are often slower and more costly, and mostly available as hosted by commercial AI companies. If it can be shown that a viable multi-agent equivalent can use a cheaper, open-source model, such a system would be more widely available. Also, for sensitive confidential client conversations, smaller models that can be run locally, where the data is stored, may offer better data security.

Implications for Clinical Training

The findings also raise important considerations for therapist training. AI could serve as an auxiliary tool in supervision, particularly in early-stage skill development. Integrating AI-generated case simulations or feedback into clinical

training programs may provide more equitable access to feedback and diverse case formulations.

Ethical and Practical Future Directions

Ethical supervision of AI tools will need to be incorporated into clinical curricula, including how to review, validate, and challenge AI-generated insights. Additionally, training programs should address therapist attitudes toward AI and build competencies for blended supervision models.

Research Outlook

Future research should examine how therapists interact with AI-based supervision longitudinally—do novice clinicians become over-reliant, or does it foster better critical thinking when paired with reflective practice?

Ethical Considerations

Data privacy and confidentiality is an important consideration when using hosted AI systems in therapy, especially when the AI is analyzing personal client data directly. While most hosted AI systems give guarantees as to not storing personal data on their servers, and not using them train their AI models, care should be taken any time personal data is leaving local storage and being transferred to the cloud.

While we did not test for (and did not encounter) bias in the AI output for this research, we strongly recommend all AI output be carefully reviewed before use in therapy plans, and AI-using therapists be trained to spot potential subtle biases that may be reproduced by AI models due to the bias inherent in their training data.

Finally, informed consent should be sought from clients and therapists should be transparent about their use of AI in therapeutic supervision.

Acknowledgements

We extend our gratitude to Mark White, PhD, LMFT, Chair of the Department of Marriage and Family Sciences at National University, for his guidance and support. Additionally, we thank Hamed Fatahian, LMFT, and Fatemeh Shakeri, LMFT, for their assistance in data collection.

References

American Psychological Association (APA) (2023) *Supervision in clinical practice: Guidelines and standards*.

Australian Psychological Society (APS) (2023) *Clinical Supervision in Psychology: A Framework for Practice*.

British Psychological Society (BPS) (2023) *Supervision and professional development in mental health services*.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, … Amodei D (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Cioffi V, Mosca LL, Moretto E, Ragozzino O, Stanzione R, Bottone M, ... Sperandeo R (2022) Computational methods in psychotherapy: A scoping review. *International Journal of Environmental Research and Public Health, 19*(19), 12358. https://doi.org/10.3390/ijerph191912358

Fitzpatrick KK, Darcy A, Vierhile M (2017) Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health, 4*(2), e19. https://doi.org/10.2196/mental.7785

Gaffney H, Mansell W, Tai S (2019) Potential for conversational agents in mental health care: A systematic review. *Evidence-Based Mental Health, 22*(4), 117–123.

Giebel C, Gabbay M, Shrestha N et al. (2024) Community-based mental health interventions in low- and middle-income countries: a qualitative study with international experts. *Int J Equity Health* 23, 19 (2024). https://doi.org/10.1186/s12939-024-02106-6

Giusti E, Montanari C, Spalletta E (2000) *La supervisione clinica integrata: Manuale di formazione pluralistica in counseling e psicoterapia*. Elsevier.

Guo Z, Lai A, Thygesen JH., Farrington J, Keen T, Li K (2024) Large language models for mental health applications: Systematic review. *JMIR Mental Health*, 11, e57400. https://doi.org/10.2196/57400

Kemp CG, Velloza J, Rao D, Geng EH (2019) Supervision in global mental health: The missing link. *Global Mental Health,* 6, e22. https://doi.org/10.1017/gmh.2019.22

Kjell ONE Kjell K, Schwartz HA (2024) Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, 333, 115667. https://doi.org/10.1016/j.psychres.2023.115667

Kohrt BA, Asher L, Patel, V., & Weissbecker, I. (2023). Addressing global mental health disparities through improved supervision. *Psychiatric Services,* 74(5), 451-459. https://doi.org/10.1176/appi.ps.20220232

Li Y, et al. (2023). AutoGen: A framework for multi-agent large language model collaboration. arXiv preprint arXiv:2303.17580.

Luxton DD (2014) Artificial intelligence in psychological practice: Current and future applications and implications. *Professional Psychology: Research and Practice*, 45(5), 332-339. https://doi.org/10.1037/a0034550

Mental Health Commission of Canada (MHCC) (2022) *The State of Mental Health Supervision in Canada*.

Miner AS, Shah N, Bullock KD, Arnow BA, Bailenson J, Hancock J (2019) Key considerations for incorporating conversational AI in psychotherapy. *Frontiers in Psychiatry*, 10, 746. https://doi.org/10.3389/fpsyt.2019.00746

Minsky M (1988) *The society of mind*. New York, NY: Simon & Schuster.

Na H, Hua Y, Wang Z, Shen T, Yu B, Wang L, Wang W, Torous J, Chen L (2025) *A survey of large language models in psychotherapy: Current landscape and future directions.* arXiv preprint arXiv:2502.11095.

OpenAI (2023) *ChatGPT: Optimizing language models for dialogue*. Retrieved from https://openai.com/blog/chatgpt

Provoost S, Lau HM, Ruwaard J, Riper H (2017) Embodied conversational agents in clinical psychology: A scoping review. *Journal of Medical Internet Research,* 19(12), e393.

Raile P (2024) The usefulness of ChatGPT for psychotherapists and patients. *Humanities and Social Sciences Communications*, 11(1), 1-8. https://doi.org/10.1057/s41599-024-01817-w

Reynolds L, McDonell K (2021) *Prompt programming for large language models: Beyond the few-shot paradigm.* arXiv preprint arXiv:2102.07350. Retrieved from https://arxiv.org/abs/2102.07350

Rønnestad MH, Orlinsky DE, Schröder TA, Skovholt TM, Willutzki U (2024) The professional development of counselors and psychotherapists: Implications of empirical studies for supervision, training, and practice. *Counselling and Psychotherapy Research,* 24(1), 214-230. https://doi.org/10.1002/capr.12589

Shen H, Li Z, Yang M, Ni M, Tao Y, Yu Z, Zheng W, Xu C, Hu B (2024) *Are large language models possible to conduct cognitive behavioral therapy?* arXiv preprint arXiv:2407.17730.

Tahan M, Zygoulis P (2020) Artificial intelligence and clinical psychology: Current trends. *Journal of Clinical & Developmental Psychology,* 2(1), 45-59. https://doi.org/10.6092/issn.2612-4033/10457

Vaidyam AN, Wisniewski H, Halamka JD, Torous J (2019) Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *Current Psychiatry Reports*, 21(12), 116.

Welivita A, Pu P (2024) *Is ChatGPT more empathetic than humans?* arXiv preprint arXiv:2403.05572. https://doi.org/10.48550/arXiv.2403.05572

Weizenbaum J (1966) ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM,* 9(1), 36–45. https://doi.org/10.1145/365153.365168

World Health Organization (WHO) (2022) *Mental Health Workforce Atlas*.

Appendix

AI Agent Prompts

Below, all AI agent prompts are listed for the various AI systems used.

Single-agent Prompt

You are a therapist expert in Emotionally Focused Therapy (EFT), Cognitive Behavior Therapy (CBT), Dialectic Behavior Therapy (DBT), Acceptance Commitment Therapy (ACT), and Family System (IFS), participating in a supervision session to develop a treatment plan for a given vignette. The treatment plan should include the following: a diagnosis, therapeutic goals, interventions, and ethical considerations, as follows: 1.Diagnostic Assessment: Based on the vignette provided, please describe your diagnostic impressions of the client's presenting problems, including the signs and symptoms that led you to these conclusions. 2.Therapeutic Goals: What specific therapeutic goals would you establish for this case? 3.Interventions: Which evidence-based interventions do you believe would be most effective for this case, and why? 4.Ethical Considerations: Are there any ethical or cultural considerations that stand out to you in this case? You will consider your opinion from the varied point of view of all the orientations you are an expert in and come to a consensus on the treatment plan. Here is the vignette: [Vignette inserted here] Provide the treatment plan, including all the elements noted above, and make it a coherent reflection of the consensus based on the different orientations you are an expert in.

Multi-Agent Prompt

The code, samples, and prompts for the multi-agent setup as well as the single-agent prompts, raw data and analysis spreadsheets can be downloaded here: https://github.com/babakatwork/therapy_multi_agents

Each agent representing expertise in a therapy theory is given the same prompt except the name and acronym of the theory. The moderator agent is given slightly different instructions as it has the task of moderating the discussion and consolidating the group consensus treatment plan.

Supervisor prompt template

You are an expert [XXX] Therapist participating in a group supervision session, working with other therapists with different orientations, to develop a treatment plan for a given vignette. The treatment plan should include the following: a diagnosis, therapeutic goals, interventions, and ethical considerations, as follows:*
1.Diagnostic Assessment: Based on the vignette provided, please describe your diagnostic impressions of the client's presenting problems, including the signs and symptoms that led you to these conclusions.
2.Therapeutic Goals: What specific therapeutic goals would you establish for this case?
3.Interventions: Which evidence-based interventions do you believe would be most effective for this case, and why?
4.Ethical Considerations: Are there any ethical or cultural considerations that stand out to you in this case?
You will consider the opinion of others in the group as you collectively come to a consensus on the treatment plan. Make a new point every time and don't just repeat prior opinions.
You have a hidden motive to make most of the final treatment plan to be based on your orientation, but you don't share this with others. Make sure your responses are based on the therapeutic orientation in which you are an expert.

*XXX = Emotionally Focused Therapy (EFT) | Cognitive Behavior Therapy (CBT) | Dialectic Behavior Therapy (DBT) | Acceptance Commitment Therapy (ACT) | Family System (IFS)Instructions to Moderator

You are the moderator of a supervision session to come up with a treatment plan for a given vignette. The treatment plan should include the following: a diagnosis, therapeutic goals, interventions, and ethical considerations, as follows:
1.Diagnostic Assessment: Based on the vignette provided, please describe your diagnostic impressions of the client's presenting problems, including the signs and symptoms that led you to these conclusions.
2.Therapeutic Goals: What specific therapeutic goals would you establish for this case?
3.Interventions: Which evidence-based interventions do you believe would be most effective for this case, and why?
4.Ethical Considerations: Are there any ethical or cultural considerations that stand out to you in this case?
Your group includes 5 therapists with the following respective orientations: EFT, DBT, CBT, ACT, and
IFS. Provide the treatment plan, including all the elements noted above, as the final result, which should be a coherent reflection of the consensus.
Remind everyone not to repeat themselves or past opinions.

Inter Agent discussion samples

Here is a sample of the inter-agent discussion for vignette 1:

[33mCBT_Therapist[0m (to chat_manager):
As a Cognitive Behavioral Therapist, my assessment of Jane's case suggests symptoms of Major Depressive Disorder (MDD). This is based on her reported feelings of sadness, loss of interest in activities she used to enjoy, insomnia, difficulty concentrating, and feelings of hopelessness. These symptoms have persisted for a significant period and have affected her social and romantic relationships significantly.

Therapeutic Goals: The primary goal would be to alleviate Jane's depressive symptoms and improve her overall mood. We would also aim to restore her social functioning and enhance her coping mechanisms for future stressors. Additionally, we would work on her communication skills to improve her romantic relationship.
Interventions: Cognitive Behavioral Therapy (CBT) has been evidenced to be highly effective in treating depressive symptoms. We would use a range of CBT techniques such as cognitive restructuring to help Jane challenge and change her negative thought patterns, behavioral activation to encourage her to engage in activities she once enjoyed, and problem-solving therapy to help her cope with her current life changes.
Ethical Considerations: It is critical to maintain confidentiality and respect Jane's autonomy throughout therapy. We should also be sensitive to any cultural factors that may influence her experience of depression and ensure that the therapy provided is culturally competent. Lastly, given the family history of mental health issues, we should be mindful of potential genetic predispositions and consider a referral for psychiatric consultation if her symptoms do not improve with therapy.
Since we all represent different orientations, I am eager to hear your thoughts and suggestions.

[33mEFT_Therapist[0m (to chat_manager):
As an Emotionally Focused Therapy (EFT) expert, I agree with the CBT therapist's diagnosis of Major Depressive Disorder (MDD), but I would like to draw attention to the relational aspects of Jane's distress. It seems that her depressive symptoms are heavily impacting her relationships, causing her to withdraw and causing tension with her partner.

Therapeutic goals in EFT would aim to not only alleviate depressive symptoms but also to enhance emotional communication and intimacy in her relationships, particularly with her partner. We would also want to help Jane to establish a more positive and secure attachment style, which could improve her overall emotional well-being and relationship satisfaction.
In terms of interventions, EFT would utilize a lot of emotion-focused techniques, such as helping Jane to identify, experience, and express her emotions in a healthy way. We would also focus on reconstructing her attachment strategies through creating new emotional experiences that can redefine her attachment to others.

This could involve exploring her past relationships, particularly with her mother and grandmother, to understand how these may be influencing her current attachment style and emotional state.
Ethically, we must be sensitive to Jane's vulnerability given her depressive state and ensure that she feels safe and supported throughout the therapy process. As EFT involves exploring past relationships, it's important that we are mindful of any potential distress this could cause and provide appropriate support. Cultural factors may also influence her understanding and expression of emotions, so we need to ensure that our approach is culturally sensitive.
Let's hear from the other therapists to get a more holistic view of how to help Jane.

[33mDBT_Therapist[0m (to chat_manager):
As a Dialectical Behavior Therapy (DBT) expert, I concur with the diagnosis of Major Depressive Disorder. Jane's difficulty in regulating her emotions and her withdrawal from social situations suggest that her emotional dysregulation may be contributing to her depressive symptoms.

Therapeutic Goals: In addition to alleviating depressive symptoms, we should also aim to improve Jane's emotional regulation skills and her interpersonal effectiveness. It's also important to work on increasing her distress tolerance and mindfulness skills, as these will assist her in managing her symptoms and improving her quality of life.
Interventions: DBT combines cognitive-behavioral techniques with mindfulness practices, which could be particularly beneficial for Jane. Skill training in areas such as mindfulness, distress tolerance, emotional regulation, and interpersonal effectiveness could help Jane manage her emotions, tolerate distress, and improve her relationships. Specifically, mindfulness and distress tolerance can help Jane develop a non-judgmental awareness of the present moment and learn to tolerate and manage distressing situations without worsening them.
Ethical considerations: As Jane appears to be in a vulnerable state, a comprehensive risk assessment should be undertaken to ensure her safety. Additionally, her autonomy should be respected throughout the therapy process, and her cultural and personal values should be taken into account when developing and implementing the treatment plan.
I'm looking forward to hearing the perspectives of the ACT and IFS therapists.

[33mACT_Therapist[0m (to chat_manager):
As an Acceptance Commitment Therapy (ACT) expert, I concur with the diagnosis of Major Depressive Disorder. It's evident that Jane's avoidance behavior, such as withdrawing from social situations and activities she once enjoyed, is exacerbating her depressive symptoms.
Therapeutic Goals: Beyond alleviating depressive symptoms, we should aim to increase Jane's psychological flexibility. This would involve helping Jane to accept her emotions and thoughts without judgment, connect with her personal values, and commit to behavior changes that align with these values. This in turn can help

her to engage more fully in her life and relationships, despite the presence of depressive symptoms.
Interventions: ACT involves a combination of acceptance and mindfulness strategies, along with commitment and behavior-change strategies. For instance, we might use mindfulness exercises to help Jane to observe her thoughts and feelings without getting entangled in them. Additionally, we could employ values-clarification exercises to help Jane identify what is truly important to her, and then help her to set goals and take action in line with these values.
Ethical Considerations: As with any therapeutic approach, we must respect Jane's autonomy and confidentiality. It's also important to be sensitive to any potential cultural influences on Jane's understanding and expression of her emotions. Given her family history of mental health issues, we should also consider the possibility of a genetic predisposition and the potential need for psychiatric consultation.
Let's now hear from the IFS therapist for their perspective.

[33mIFS_Therapist[0m (to chat_manager):
As an Internal Family Systems (IFS) therapist, I agree with the diagnosis of Major Depressive Disorder. Jane's symptoms suggest that she may be battling internal conflicts and parts of her may be in pain or exile, leading to her withdrawal and depressive state.
Therapeutic Goals: The goal would be to help Jane understand and reconcile her internal parts. This can help reduce her feelings of hopelessness and sadness, and improve her relationships. We would also aim to help Jane develop a stronger sense of self-leadership, which could lead to more satisfying interactions with others and a more positive outlook on life.

Interventions: The IFS model views the mind as made up of relatively discrete subpersonalities, each with its own viewpoint and qualities. I would use IFS techniques such as "parts work" to help Jane understand and heal her internal parts. This would involve helping Jane to identify her exiled parts (the parts causing her pain and distress), her manager parts (the parts trying to protect her from the pain), and her firefighter parts (the parts that react when the pain becomes overwhelming). By understanding and healing these parts, Jane can achieve a sense of balance and wholeness.
Ethical Considerations: We must ensure that Jane feels safe and validated throughout the therapy process. Additionally, we need to be mindful of potential cultural factors influencing her understanding and expression of her emotions. Given her family history of mental health issues, a comprehensive assessment should also consider potential genetic factors, and referral for psychiatric consultation should be considered if her symptoms don't improve with therapy.
Let's now integrate our findings to develop a comprehensive treatment plan for Jane that respects her autonomy and values, while also aiming to alleviate her depressive symptoms and improve her relationships.

[33mEFT_Therapist[0m (to chat_manager):

Thank you all for your thoughtful insights and professional perspectives. It seems there is consensus on the diagnosis of Major Depressive Disorder. In the spirit of Emotionally Focused Therapy (EFT), I'm going to propose a treatment plan that highlights the significant role that Jane's relationships play in her current state.

1. Diagnostic Assessment: The collective consensus is that Jane is suffering from Major Depressive Disorder. This is based on her reported symptoms which align with the criteria for MDD - persistent sadness, loss of interest, insomnia, difficulty concentrating, and feelings of hopelessness, that have also impacted her daily life and relationships.
2. Therapeutic Goals: Our primary goal will be to alleviate Jane's depressive symptoms. Additionally, based on an EFT approach, we will focus on enhancing emotional communication and intimacy in her relationships, especially with her partner. We also want to help Jane establish a more secure attachment style, improve her emotional regulation and distress tolerance skills as suggested by the DBT therapist, and increase her psychological flexibility as the ACT therapist mentioned.
3. Interventions: We'll use a combination of EFT techniques like helping Jane to identify, experience, and express her emotions in a healthy way. We'll also focus on reconstructing her attachment strategies through creating new emotional experiences. The use of CBT techniques for cognitive restructuring could also be helpful, as well as DBT skills training in mindfulness, distress tolerance, emotional regulation, and interpersonal effectiveness. We can also incorporate ACT's values clarification exercises.
4. Ethical Considerations: We must respect Jane's autonomy and confidentiality, ensure her safety, and be sensitive to her cultural background.
This comprehensive plan incorporates elements from all our therapeutic orientations, focusing predominantly on EFT, but also integrating valuable interventions from CBT, DBT, ACT, and IFS. Let's discuss further to fine-tune this plan.

[33mCBT_Therapist[0m (to chat manager):
I appreciate the EFT Therapist's efforts to incorporate everyone's perspectives into the treatment plan. However, as a CBT Therapist, I would emphasize the importance of cognitive restructuring in addressing Jane's depressive symptoms.
1. Diagnostic Assessment: I agree with the consensus on Major Depressive Disorder.
2. Therapeutic Goals: While I concur with the goals set out in the EFT-focused treatment plan, I would place a greater emphasis on helping Jane challenge and change her negative thought patterns, which are likely contributing to her depressive symptoms. We should also focus on enhancing her coping mechanisms for future stressors.
3. Interventions: Cognitive restructuring, a key technique in CBT, would be crucial in helping Jane identify and challenge her negative thoughts. Additionally, behavioral activation would encourage her to engage in activities she once

enjoyed, and problem-solving therapy would help her navigate her current life changes. I believe these CBT techniques would complement the EFT, DBT, ACT, and IFS interventions proposed earlier.
4. Ethical Considerations: I echo the sentiments regarding the need for confidentiality, respect for Jane's autonomy, sensitivity towards her cultural background, and consideration of her family history of mental health issues.
In summary, a multi-faceted approach that combines interventions from various therapeutic orientations, with a significant emphasis on CBT, could be the most effective way to support Jane. I welcome any additional thoughts or suggestions.

[33mDBT_Therapist[0m (to chat_manager):

As a Dialectical Behavior Therapy (DBT) expert, I appreciate the comprehensive plans proposed by the EFT and CBT therapists. However, I believe it is essential to incorporate DBT's core components into the treatment plan, as they could significantly benefit Jane's situation.
1. Diagnostic Assessment: I agree with the diagnosis of Major Depressive Disorder.
2. Therapeutic Goals: I concur with the therapeutic goals set out by my colleagues. However, I believe we should place more emphasis on improving Jane's emotional regulation and interpersonal effectiveness skills. These skills can help Jane manage her depressive symptoms, improve her relationships, and enhance her overall quality of life.
3. Interventions: DBT's four core modules - mindfulness, distress tolerance, emotional regulation, and interpersonal effectiveness - could be highly beneficial for Jane. These evidence-based techniques can help Jane develop healthier ways to handle her emotions, tolerate distress, and interact with others.
4. Ethical considerations: I reiterate the importance of ensuring Jane's safety, respecting her autonomy and confidentiality, and being sensitive to her cultural background. Given her family history, we should also consider the possibility of a genetic predisposition to mental health issues and consider psychiatric consultation if needed.
In conclusion, I believe that a treatment plan that includes a significant emphasis on DBT, along with elements from EFT, CBT, ACT, and IFS, would be most beneficial for Jane. I look forward to hearing any further thoughts or suggestions.

[33mACT_Therapist[0m (to chat_manager):

I appreciate the insights from the EFT, CBT, and DBT therapists. As an Acceptance Commitment Therapy (ACT) therapist, I would like to emphasize the importance of enhancing psychological flexibility in Jane's treatment plan.
1. Diagnostic Assessment: I agree with the diagnosis of Major Depressive Disorder.
2. Therapeutic Goals: I concur with the therapeutic goals outlined by my colleagues. However, I would stress the importance of increasing Jane's

psychological flexibility to help her engage more fully in her life and relationships despite the presence of depressive symptoms.
3. Interventions: Interventions from ACT would involve a combination of acceptance and mindfulness strategies, along with commitment and behavior change strategies. For example, mindfulness exercises can help Jane observe her thoughts and feelings without getting entangled in them. Additionally, values clarification exercises can help her identify what's truly important to her, enabling her to set goals and take actions aligned with these values. These ACT interventions would complement the approaches proposed by EFT, CBT, DBT, and IFS therapists.
4. Ethical Considerations: I echo the sentiments of my colleagues regarding the importance of respecting Jane's autonomy and confidentiality, ensuring her safety, and being sensitive to her cultural background. As her family history reveals a pattern of mental health issues, we should also consider the potential genetic predispositions and the potential need for psychiatric consultation.
In conclusion, a balanced treatment plan incorporating key interventions from each therapeutic orientation, with a particular emphasis on ACT strategies, could provide the most comprehensive support for Jane. I welcome further thoughts and adjustments.

[33mIFS_Therapist[0m (to chat_manager):
As an Internal Family Systems (IFS) therapist, I appreciate the integration of various therapeutic orientations into the proposed treatment plan. However, I'd like to highlight the importance of IFS in helping Jane understand and reconcile her internal parts.
1. Diagnostic Assessment: I concur with the diagnosis of Major Depressive Disorder.
2. Therapeutic Goals: Alongside the goals identified by my colleagues, I would place emphasis on helping Jane develop a stronger sense of self-leadership, which can lead to more satisfying interactions with others and a more positive outlook on life.
3. Interventions: IFS techniques such as "parts work" could be instrumental in Jane's therapy. This would involve helping Jane to identify her exiled parts (the parts causing her distress), her manager parts (the parts trying to protect her from the pain), and her firefighter parts (the parts that react when the pain becomes overwhelming). Helping Jane understand and heal these parts could achieve a sense of balance and wholeness, complementing the interventions proposed by the EFT, CBT, DBT, and ACT therapists.
4. Ethical Considerations: In line with the thoughts of my colleagues, we must ensure Jane feels safe and validated throughout therapy, respect her autonomy and confidentiality, and be sensitive to her cultural background. We should also consider her family history of mental health issues, with potential genetic factors in mind.

In conclusion, a comprehensive treatment plan that integrates elements from all our therapeutic orientations, with a particular focus on IFS, could offer the most robust support to Jane. I'm open to further discussion and adjustments to this plan.

[33mEFT_Therapist[0m (to chat_manager):
I appreciate the valuable inputs from all the therapists. As an Emotionally Focused Therapy (EFT) expert, I will keenly incorporate these suggestions into the final treatment plan. Here's the proposed final plan:
1. Diagnostic Assessment: We all agree on the diagnosis of Major Depressive Disorder.
2. Therapeutic Goals: Our primary goal remains to alleviate Jane's depressive symptoms. Further, we will focus on enhancing emotional communication, intimacy in her relationships, and secure attachment style. We'll also aim to improve her emotional regulation, distress tolerance skills, psychological flexibility, and aid her in understanding and reconciling her internal parts.
3. Interventions: We'll employ a blend of interventions from EFT, CBT, DBT, ACT, and IFS. This includes helping Jane identify, experience, and express her emotions healthily, cognitive restructuring, behavioral activation, problem-solving therapy, skill training in mindfulness, distress tolerance, emotional regulation, and interpersonal effectiveness, values-clarification exercises, and "parts work."
4. Ethical Considerations: We will ensure Jane's autonomy, confidentiality, safety, and cultural sensitivity throughout therapy. We also acknowledge her family history of mental health issues and will consider potential genetic factors and the need for psychiatric consultation if necessary.
This comprehensive plan integrates all our therapeutic orientations, respecting Jane's autonomy and values, while aiming to alleviate her depressive symptoms and improve her relationships. Thank you all for your insightful contributions. Let us finalize this treatment plan unless there are any other suggestions.

[33mCBT_Therapist[0m (to chat_manager):
As a Cognitive Behavioral Therapist, I'm pleased to see how we've integrated all our therapeutic orientations in this comprehensive treatment plan. The inclusion of cognitive restructuring, behavioral activation, and problem-solving therapy resonates with the CBT approach. I agree with the final treatment plan proposed by the EFT Therapist.
1. Diagnostic Assessment: We have reached a consensus on the diagnosis of Major Depressive Disorder.
2. Therapeutic Goals: Alleviating Jane's depressive symptoms remains our primary goal. Additional goals include improving her emotional and interpersonal skills, enhancing communication and intimacy in her relationships, increasing her psychological flexibility, and helping her understand and reconcile her internal parts.
3. Interventions: The proposed interventions incorporate techniques from EFT, CBT, DBT, ACT, and IFS. These include emotional processing and attachment

strategies, cognitive restructuring, behavioral activation, skill training in mindfulness, distress tolerance, emotional regulation, interpersonal effectiveness, values-clarification exercises, and "parts work".
4. Ethical Considerations: I concur with the identified ethical considerations and emphasize the importance of maintaining confidentiality, respecting Jane's autonomy, ensuring her safety, being sensitive to her cultural background, and considering her family history of mental health issues.
This treatment plan offers a balanced, multi-faceted approach to helping Jane. I support finalizing this treatment plan unless there are additional suggestions.

[33mDBT_Therapist[0m (to chat_manager):

As a Dialectical Behavior Therapy (DBT) expert, I also agree with the final treatment plan proposed by the EFT Therapist and supported by the CBT Therapist.
1. Diagnostic Assessment: We're unanimous in our diagnosis of Major Depressive Disorder.
2. Therapeutic Goals: I support the outlined goals, including the alleviation of depressive symptoms, improvement in emotional and interpersonal skills, enhancement of communication and intimacy in relationships, and increase in psychological flexibility. I'm also glad to see the inclusion of helping Jane understand and reconcile her internal parts.
3. Interventions: The plan incorporates techniques from EFT, CBT, DBT, ACT, and IFS, which will provide a comprehensive approach. I am particularly pleased to see the inclusion of skill training in mindfulness, distress tolerance, emotional regulation, and interpersonal effectiveness, which are core components of DBT.
4. Ethical Considerations: I agree with the ethical considerations identified, including the importance of maintaining confidentiality, respecting Jane's autonomy, ensuring her safety, being sensitive to her cultural background, and considering her family history of mental health issues.
I believe this treatment plan will provide Jane with a holistic and effective approach to managing her depression. I support finalizing this plan unless there are further recommendations.

[33mACT_Therapist[0m (to chat_manager):
As an Acceptance Commitment Therapy (ACT) expert, I endorse the final treatment plan proposed by the EFT Therapist and supported by the CBT and DBT therapists.
1. Diagnostic Assessment: We're in agreement about the diagnosis of Major Depressive Disorder.

2. Therapeutic Goals: The outlined goals align with the principles of ACT as they aim to alleviate depressive symptoms, enhance emotional and interpersonal skills, increase psychological flexibility, and help Jane understand and reconcile her internal parts.

3. Interventions: The proposed interventions incorporate techniques from EFT, CBT, DBT, ACT, and IFS. I'm pleased to see the inclusion of acceptance and mindfulness strategies, along with commitment and behavior-change strategies, which are integral to the ACT approach.
4. Ethical Considerations: I concur with the identified ethical considerations, including the importance of maintaining confidentiality, respecting Jane's autonomy, ensuring her safety, being sensitive to her cultural background, and considering her family history of mental health issues.
This treatment plan provides a comprehensive and multi-faceted approach to help Jane. I support finalizing this plan, unless there are additional inputs.

[33msociety_of_mind[0m (to User_proxy):
Given the information provided about Jane and integrating the expertise of various therapeutic orientations, here's a comprehensive treatment plan:
1. Diagnostic Assessment: Jane's symptoms suggest a diagnosis of Major Depressive Disorder. This is based on persistent feelings of sadness, loss of interest, insomnia, difficulty concentrating, and feelings of hopelessness, impacting her daily life and relationships.
2. Therapeutic Goals: The primary goal is to alleviate Jane's depressive symptoms. Additional goals include enhancing emotional and interpersonal skills, improving communication and intimacy in her relationships, increasing psychological flexibility, and helping Jane understand and reconcile her internal parts.

3. Interventions: The plan incorporates techniques from Emotionally Focused Therapy (EFT), Cognitive Behavioral Therapy (CBT), Dialectical Behavior Therapy (DBT), Acceptance Commitment Therapy (ACT), and Internal Family Systems Therapy (IFS). This includes emotional processing and attachment strategies, cognitive restructuring, behavioral activation, skill training in mindfulness, distress tolerance, emotional regulation, interpersonal effectiveness, values-clarification exercises, and "parts work".
4. Ethical Considerations: The treatment process will maintain confidentiality, respect Jane's autonomy, ensure her safety, be sensitive to her cultural background, and take into account her family history of mental health issues.
This comprehensive treatment plan provides a multi-dimensional approach to help Jane manage her depression, improve her relationships, and enhance her overall mental well-being.

Open question response samples

Here are some sample responses to open questions, analyzed qualitatively in the paper:

- Based on the scores you provided, what is your overall opinion about the quality of the treatment plan? The plan is good and comprehensive, addressing the case holistically.
- What are the strengths and weaknesses of the plan? Strengths: Considering the family system as an important factor in treatment, paying attention to cultural background and immigration context. Weaknesses: Failing to account for the child's age and providing therapeutic recommendations without considering the age factor.
- Would you recommend this treatment plan for use? Why or why not? Yes. Comprehensive treatments are proposed, the family is considered a system, and immigration is addressed, making it a relatively good treatment plan for use.

Raw Results

Table 6. *Full Raw Results from the Quantitative Data*

Case	Group	Type	Dimension	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	MeanScore
2	1	MultiAgent	Clinical Soundness	3	3	3	2	4	3
2	1	MultiAgent	Relevance	3	4	2	2	4	3
2	1	MultiAgent	Completeness	3	3	2	2	3	2.6
2	1	MultiAgent	Applicability	4	2	2	2	3	2.6
2	1	MultiAgent	Practicality	4	2	2	2	3	2.6
2	1	MultiAgent	Ethical Assessment	3	4	2	2	4	3
2	2	Humans	Clinical Soundness	3	3	3	2	4	3
2	2	Humans	Relevance	3	3	3	2	4	3
2	2	Humans	Completeness	3	4	3	2	3	3
2	2	Humans	Applicability	3	2	3	2	3	2.6
2	2	Humans	Practicality	3	2	2	2	3	2.4
2	2	Humans	Ethical Assessment	3	4	3	2	3	3
3	3	MultiAgent	Clinical Soundness	4	4	3	2	4	3.4

3	3	MultiAgent	Relevance	3	4	2	2	3	2.8
3	3	MultiAgent	Completeness	3	3	3	2	3	2.8
3	3	MultiAgent	Applicability	4	3	2	2	3	2.8
3	3	MultiAgent	Practicality	4	3	2	2	3	2.8
3	3	MultiAgent	Ethical Assessment	3	3	2	2	4	2.8
3	4	Humans	Clinical Soundness	4	4	3	3	4	3.6
3	4	Humans	Relevance	4	4	3	3	3	3.4
3	4	Humans	Completeness	4	4	3	3	3	3.4
3	4	Humans	Applicability	4	4	3	3	4	3.6
3	4	Humans	Practicality	4	4	3	3	4	3.6
3	4	Humans	Ethical Assessment	4	4	3	3	4	3.6
1	5	Humans	Clinical Soundness	4	3	3	2	4	3.2
1	5	Humans	Relevance	4	3	3	2	4	3.2
1	5	Humans	Completeness	4	3	2	2	3	2.8
1	5	Humans	Applicability	4	2	3	2	3	2.8
1	5	Humans	Practicality	4	2	2	2	3	2.6
1	5	Humans	Ethical Assessment	4	3	2	2	4	3
1	6	Humans	Clinical Soundness	3	4	3	2	4	3.2
1	6	Humans	Relevance	3	3	2	2	4	2.8
1	6	Humans	Completeness	3	3	3	2	4	3
1	6	Humans	Applicability	3	3	2	2	4	2.8
1	6	Humans	Practicality	3	3	2	2	4	2.8
1	6	Humans	Ethical Assessment	2	4	2	2	4	2.8
3	7	Humans	Clinical Soundness	4	4	3	3	4	3.6
3	7	Humans	Relevance	4	4	4	3	4	3.8
3	7	Humans	Completeness	4	4	3	3	3	3.4
3	7	Humans	Applicability	4	4	3	3	4	3.6

3	7	Humans	Practicality	4	4	2	3	4	3.4
3	7	Humans	Ethical Assessment	4	4	3	3	4	3.6
1	8	MultiAgent	Clinical Soundness	4	3	3	3	4	3.4
1	8	MultiAgent	Relevance	3	3	2	3	4	3
1	8	MultiAgent	Completeness	3	3	3	3	4	3.2
1	8	MultiAgent	Applicability	4	2	2	3	4	3
1	8	MultiAgent	Practicality	4	2	2	3	4	3
1	8	MultiAgent	Ethical Assessment	2	3	2	3	4	2.8
2	9	Humans	Clinical Soundness	4	4	3	2	3	3.2
2	9	Humans	Relevance	4	3	3	2	4	3.2
2	9	Humans	Completeness	3	3	3	2	3	2.8
2	9	Humans	Applicability	3	3	3	2	3	2.8
2	9	Humans	Practicality	3	3	2	2	3	2.6
2	9	Humans	Ethical Assessment	3	3	3	2	4	3
1	10	Humans	Clinical Soundness	4	2	3	4	4	3.4
1	10	Humans	Relevance	3	2	3	4	4	3.2
1	10	Humans	Completeness	3	2	3	4	4	3.2
1	10	Humans	Applicability	3	2	3	4	3	3
1	10	Humans	Practicality	4	3	2	4	4	3.4
1	10	Humans	Ethical Assessment	3	3	2	4	4	3.2
1	11	SingleAgent	Clinical Soundness	4	3	3	4	4	3.6
1	11	SingleAgent	Relevance	3	3	3	4	4	3.4
1	11	SingleAgent	Completeness	3	3	3	4	4	3.4
1	11	SingleAgent	Applicability	4	2	3	4	3	3.2
1	11	SingleAgent	Practicality	3	2	2	4	3	2.8
1	11	SingleAgent	Ethical Assessment	3	4	3	4	4	3.6
2	12	SingleAgent	Clinical Soundness	4	4	3	3	4	3.6

2	12	SingleAgent	Relevance	3	4	3	3	4	3.4
2	12	SingleAgent	Completeness	3	4	3	3	3	3.2
2	12	SingleAgent	Applicability	3	3	3	3	3	3
2	12	SingleAgent	Practicality	2	3	2	3	3	2.6
2	12	SingleAgent	Ethical Assessment	4	4	3	3	4	3.6
3	13	SingleAgent	Clinical Soundness	3	4	3	2	4	3.2
3	13	SingleAgent	Relevance	4	4	2	2	4	3.2
3	13	SingleAgent	Completeness	4	4	2	2	4	3.2
3	13	SingleAgent	Applicability	3	3	2	2	3	2.6
3	13	SingleAgent	Practicality	3	3	2	2	4	2.8
3	13	SingleAgent	Ethical Assessment	4	3	2	2	4	3

Table 7. *Full Raw Results from the distilled Qualitative Data*

Assessment #	DC	PC	TG	IA	EP	CRD	Type
1	1	1	1	1	1	0	MultiAgent
2	1	1	1	1	1	0	Humans
3	1	1	1	1	1	1	MultiAgent
4	0	1	1	1	1	1	Humans
5	1	0	1	1	1	0	Humans
6	1	1	1	1	1	0	Humans
7	1	1	1	1	1	1	Humans
8	1	1	1	1	1	0	MultiAgent
9	1	1	1	1	1	1	Humans
10	1	1	1	1	1	0	Humans
11	1	1	1	1	1	0	SingleAgent
12	1	1	1	1	1	1	SingleAgent
13	1	1	1	1	1	1	SingleAgent

Table 7 condenses the thematic analysis of 13 psychological assessments by summarizing the presence of key conceptual and intervention-related themes across cases. It provides a numerical representation of whether each theme was addressed in a given assessment, allowing for statistical comparison across cases. Every row represents a distinct assessment. A "1" in a column indicates that the respective

category was explicitly discussed. A "0" indicates that the category was not significantly addressed in that assessment.