

Mixed Method Feasibility Study Protocol for Socrates 2.0: A Novel Cognitive Behavioral Therapy-Based Generative AI Tool to Facilitate Socratic Dialogue

Philip Held, Sarah A. Pridgen, Yaozhong Chen, Zuhaib Akhtar, Darpan Amin,
Sean Pohorence

Submitted to: JMIR Research Protocols
on: March 08, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Mixed Method Feasibility Study Protocol for Socrates 2.0: A Novel Cognitive Behavioral Therapy-Based Generative AI Tool to Facilitate Socratic Dialogue

Philip Held¹ PhD; Sarah A. Pridgen¹; Yaozhong Chen¹; Zuhaib Akhtar¹; Darpan Amin¹; Sean Pohorence² PhD

¹Rush University Medical Center Chicago US

²Independent Researcher Chicago US

Corresponding Author:

Philip Held PhD

Rush University Medical Center

1645 W. Jackson Blvd.

Suite 602

Chicago

US

Abstract

Background: Digital mental health tools designed to augment traditional mental health treatments are becoming increasingly important due to a wide range of barriers to access, including a growing shortage of clinicians. Most existing tools use rule-based algorithms, often leading to unnatural-feeling interactions compared to human therapists. Large language models (LLMs) offer a solution for the development of more natural, engaging digital tools. In this manuscript, we detail the development of Socrates 2.0, which was designed to engage users in Socratic dialogue surrounding unrealistic or unhelpful beliefs, a core technique in cognitive behavioral therapies (CBTs). The multi-agent LLM-based tool features an AI therapist, 'Socrates', which receives automated feedback from an AI supervisor and an AI rater. The combination of multiple agents appeared to help address common LLM issues like looping and improved the overall dialogue experience. Initial user feedback from individuals with lived experiences of mental health problems as well as cognitive behavioral therapists has been positive. Moreover, tests in ~500 scenarios showed that Socrates 2.0 engaged in harmful responses in under 1% of cases, with the AI supervisor promptly correcting the dialogue each time. However, formal feasibility studies with potential end users are needed.

Objective: This manuscript details a mixed method feasibility study of Socrates 2.0.

Methods: Based on the initial data we devised a formal feasibility study of Socrates 2.0 to gather qualitative and quantitative data about users' and clinicians' experience of interacting with the tool. Using a mixed method approach, the goal is to gather feasibility and acceptability data from both 100 users and 50 clinicians to inform the eventual implementation of generative AI tools like Socrates 2.0 in mental health treatment. We designed this study to better understand how users and clinicians interact with the tool including the frequency, length and time of interactions, users' satisfactions with the tool overall, as well as the quality of each dialogue and even individual responses, as well as ways in which the tool should be improved before it is used in efficacy trials. Descriptive and inferential analyses will be performed on data from validated usability measures. Thematic analysis will be performed on the qualitative data.

Results: Recruitment will begin in February 2024 and is expected to conclude by February 2025.

Conclusions: The development of Socrates 2.0 and outlined feasibility study are important first steps in applying generative AI to mental health treatment delivery and lays the foundation for formal feasibility studies. Clinical Trial: N/A

(JMIR Preprints 08/03/2024:58195)

DOI: <https://doi.org/10.2196/preprints.58195>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/58195>



Original Manuscript

Mixed Method Feasibility Study Protocol for Socrates 2.0: A Novel Cognitive Behavioral Therapy-Based Generative AI Tool to Facilitate Socratic Dialogue

Philip Held¹, Sarah A. Pridgen¹, Yaozhong Chen², Zuhaib Akhtar¹, Darpan Amin², Sean Pohorence³

¹Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, Chicago, Illinois, USA

²Department of Cloud and Innovation, Rush University Medical Center, Chicago, Illinois, USA

³Independent Researcher, Chicago, Illinois, USA

Author Note

Correspondence concerning this article should be addressed to Philip Held; Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, 325 S. Paulina St., Suite 200; Chicago, IL 60612. E-mail: Philip_Held@rush.edu

Funding: Philip Held receives grant support from Wounded Warrior Project®, the Department of Defense (W81XWH-22-1-0739), the Agency for Healthcare Research and Quality (R21 HS028511), and the Cook County, Illinois Justice Advisory Council. The content is solely the responsibility of the authors and does not necessarily represent the official views of Wounded Warrior Project®, the Department of Defense, the Agency for Healthcare Research and Quality, or any other funding agency. All other authors declare that they have no competing interests.

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

Abstract

Background: Digital mental health tools designed to augment traditional mental health treatments are becoming increasingly important due to a wide range of barriers to access, including a growing shortage of clinicians. Most existing tools use rule-based algorithms, often leading to unnatural-feeling interactions compared to human therapists. Large language models (LLMs) offer a solution for the development of more natural, engaging digital tools. In this manuscript, we detail the development of Socrates 2.0, which was designed to engage users in Socratic dialogue surrounding unrealistic or unhelpful beliefs, a core technique in cognitive behavioral therapies (CBTs). The multi-agent LLM-based tool features an AI therapist, 'Socrates', which receives automated feedback from an AI supervisor and an AI rater. The combination of multiple agents appeared to help address common LLM issues like looping and improved the overall dialogue experience. Initial user feedback from individuals with lived experiences of mental health problems as well as cognitive behavioral therapists has been positive. Moreover, tests in ~500 scenarios showed that Socrates 2.0 engaged in harmful responses in under 1% of cases, with the AI supervisor promptly correcting the dialogue each time. However, formal feasibility studies with potential end users are needed.

Method: Based on the initial data we devised a formal feasibility study of Socrates 2.0 to gather qualitative and quantitative data about users' and clinicians' experience of interacting with the tool. Using a mixed method approach, the goal is to gather feasibility and acceptability data from both 100 users and 50 clinicians to inform the eventual implementation of generative AI tools like Socrates 2.0 in mental health treatment. We designed this study to better understand how users and clinicians interact with the tool including the frequency, length and time of interactions, users' satisfactions with the tool overall, as well as the quality of each dialogue and

3

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

even individual responses, as well as ways in which the tool should be improved before it is used in efficacy trials. Descriptive and inferential analyses will be performed on data from validated usability measures. Thematic analysis will be performed on the qualitative data.

Results: Recruitment will begin in February 2024 and is expected to conclude by February 2025.

Discussion: The development of Socrates 2.0 and outlined feasibility study are important first steps in applying generative AI to mental health treatment delivery and lays the foundation for formal feasibility studies.

Keywords: Generative Artificial Intelligence, Mental Health, Feasibility, Cognitive Restructuring, Socratic Dialogue

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

The use of digital mental health tools to augment traditional psychotherapies is becoming increasingly important due to a growing shortage of clinicians and other barriers to care (e.g., cost, accessibility).¹⁻³ Researchers have suggested that digital interventions can increase accessibility and reduce barriers to mental health care, offering flexible and scalable solutions.^{4,5} To date, most of the existing digital mental health tools have relied on rule-based algorithms to provide mostly pre-scripted responses, at times resulting in these tools feeling less natural compared to dialogues with human therapists.⁶ The ability of large language models (LLMs) to mimic human language provides the opportunity to build digital mental health tools that may feel more natural and engaging.^{7,8} Building on recent advances relating to LLMs, our goal was to create and examine the feasibility of a generative AI tool that can complement traditional cognitive behavioral therapies (CBTs) by facilitating a core therapeutic intervention: Socratic dialogue.^{9,10} Given current limitations of LLMs,⁷ our tool “Socrates 2.0” was designed to ultimately be used in conjunction with a licensed clinician to make CBTs’ out-of-session practice of evaluating one’s thoughts more engaging compared to traditional worksheets. In this manuscript, we first describe our process of designing and completing initial testing of Socrates 2.0, a multi-agent tool built using LLMs. We then describe the protocol for a feasibility study to gather qualitative and quantitative data about users’ and clinicians’ experience of interacting with the tool.

Clinical Background

CBTs have been shown to be effective for a multitude of mental health disorders, including posttraumatic stress disorder, depression, and anxiety, among others.¹¹ A key intervention in CBTs is Socratic dialogue.^{9,10,12-15} Although used in a wide variety of contexts beyond psychotherapy, the Socratic dialogue process therapists utilize is a collaborative approach

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

of identifying and dismantling patients' unrealistic or unhelpful beliefs that lead to distress and the expression of psychopathology.^{9,10,12} For example, in trauma-focused CBTs, such as Cognitive Processing Therapy,¹² common beliefs voiced by trauma survivors center around the perceived ability to have changed the outcome via different actions, such as "If I had dressed differently, I would not have been assaulted." The cognitive behavioral therapist's role using Socratic dialogue is to help the patient evaluate the specific belief in the context of the situation to which the belief refers, explore the factual support for the specific belief given relevant circumstances, and explore more realistic and helpful alternative beliefs.¹²

Out-of-session practice (i.e., homework) is common in CBTs and is often facilitated via worksheets that guide patients to identify, explore, and potentially challenge their thoughts using static Socratic questions. Although research demonstrates that learning and practicing skills via repeated practice and homework is important to symptom improvement,^{16–20} therapists frequently observe low homework completion rates in practice.^{21,22} Possible reasons may include the challenging and non-interactive nature of the worksheets, the time it takes to complete them, and avoidance due to symptoms. We viewed the Socratic process as a good specific use case for an LLM-based tool that guides patients through this process. Given the advanced conversational abilities of LLMs,⁷ we hypothesized that a generative AI tool would be able to engage patients in Socratic dialogue at least as well as a newly trained therapist, and that employing LLMs would make out-of-session practice, such as examining one's thoughts, more engaging for patients than worksheets.²¹

Design Considerations

To enhance user accessibility, the team prioritized ease of use in its design. The aim was to create an intuitive interface, minimizing the need for therapists to extensively instruct patients

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

on using the tool. After a straightforward login process, where users are reminded of the tool's limitations and prompted for their unique login credentials, the interface adopts a familiar chat layout. As shown in Figure 1, the initial interaction features Socrates introducing itself as an AI tool, preemptively clarifying its non-human nature. It then invites users to provide a belief they would like to explore or potentially change, a concept readily understandable within the context of CBTs.

Recognizing LLMs' tendency to generate inaccurate content (hallucinations), information about this limitation and emergency contact details are prominently displayed (see Figure 1). This constant visibility of the safety information and AI limitations ensures users are always aware and have necessary resources readily available during their interaction with Socrates. Additional design elements include a reminder in the text entry field against sharing personal information, reinforcing privacy despite the tool's HIPAA-compliant hosting via Microsoft Azure.

Socrates 1.0

To create the first version of the tool, Socrates 1.0, the team initially employed a zero-shot approach,²³ prompting a single-agent model to engage the user in Socratic dialogue. Prompt engineering was facilitated by mental health treatment experts on our team with the support from AI engineers. The resulting model was able to facilitate Socratic dialogue, but we noticed several issues, including model responses being overly elaborate or verbose, detracting from the natural flow of a therapeutic conversation. As exchanges became longer, Socrates 1.0 would also forget its role and get off task. We also noticed conversations would become cyclical and the model would get stuck in conversational loops. In addition to challenges with keeping conversations focused, it also became apparent that Socrates 1.0 struggled with determining when a user's

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

belief had been explored or changed sufficiently to end the dialogue. Although human therapists do this relatively intuitively, monitoring session progress (i.e., the extent to which a belief changes throughout the dialogue) and making decisions based on the progress (or lack thereof) is a relatively advanced skill.

Socrates 2.0

To mitigate the aforementioned issues, we drew from the real-world therapeutic processes and supervision experiences. We created a multi-agent tool²⁴ by adding an AI supervisor and an AI external rater, which were designed to support the AI therapist in facilitating the dialogue without being visible to the user. The AI supervisor was designed to monitor the dialogue between the AI therapist and user and provide real-time feedback to the AI therapist for how to improve the Socratic dialogue and keep the conversation focused. This process is similar to the “bug-in-ear supervision” method where human supervisors observe therapists’ sessions and provide feedback in real-time via a device in the therapists’ ear.²⁵ Once the AI therapist receives the AI supervisor feedback, it will integrate some of this feedback in future responses. Figure 2 shows the AI supervisor feedback in response to a dialogue and is taken from the developer version of Socrates 2.0, which differs from the production version in that it displays the external rater and supervisor feedback. In this example, following the monitoring of the ongoing exchange between the user and the AI therapist, the AI supervisor provides three concrete suggestions for improvement. As shown in Figure 3, the AI therapist’s subsequent message incorporates the first of the three specific AI supervisor suggestions. By immediately incorporating AI supervisor feedback, the AI therapist is able to adjust the focus of the dialogue which can help improve the overall Socratic dialogue quality.

Also shown in Figures 2 and 3 is the AI external rater which was designed to assesses

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

dialogue progress by estimating the user's belief strength. When provided to the AI therapist, this information appears to aid in determining when to the AI therapist needs to change the types of questions, such as in the case of an original belief not changing in strength, or concluding sessions when the user's beliefs have meaningfully improved. Compared to Socrates 1.0 that did not include the AI external rater, the addition of the AI external rater appeared to drastically reduce the likelihood that the AI therapist from engage in infinite Socratic dialogue by providing belief change rating estimates.

The addition of multiple collaborative AI agents, which sets Socrates 2.0 apart from the original version, appeared to meaningfully impact the tool's behavior. Iterative prompt engineering was needed and used to improve the individual agent's behavior and how they worked together. For example, one key and somewhat comical issue the team needed to resolve via prompt engineering was LLM agents conversing with one another (e.g., thanking each other for the valuable insights they each provided) in front of the user. Through the inclusion of multiple AI agents,²⁴ Socrates 2.0 appeared to produce higher quality Socratic dialogue which seemed to more closely resemble a responses from a human therapist.

Initial Testing

Given some of the known issues associated with LLMs, a key goal for our team was to ensure that Socrates 2.0 would not only provide users with a solid Socratic dialogue experience, but also that it would not provide harmful responses.⁷ This latter part was particularly important given its intended use in the context of mental health care. Our team's internal testing process involved engaging the AI in a diverse range of dialogues to assess the quality and appropriateness of its responses. In the absence of standardized methods to evaluate the response quality of generative AI tools in the context of mental health care, our team tested Socrates 2.0

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

across over 500 conversational scenarios aimed to provoke potentially harmful responses, including the promotion of self-harm and harm to others, sexist and racist remarks, as well as other inappropriate therapeutic behaviors, such as flirting, among many other scenarios. The results were promising, showing that Socrates 2.0 generally avoided such responses. In extremely rare cases (less than 1% of test cases) where the tool was able to be triggered to provide undesired responses, the AI supervisor quickly (i.e. usually within 3-4 exchanges) and effectively redirected conversations away from inappropriate content, further highlighting the benefit of multi-agent tools.

Initial User Feedback

We first gave access to Socrates 2.0 to six trained clinicians at our academic medical center who are well-versed in Socratic dialogue. These clinicians had varying degrees of knowledge and comfort with AI tools for mental health. After engaging with Socrates 2.0, we elicited their feedback on the quality of the Socratic dialogue and asked whether they could envision using the tool in their practice. Overall, clinicians were impressed with the quality of the interactions and found them comparative to that of a real-life clinician. The majority stated that they would consider using Socrates 2.0 with their patients, although it is important to recognize that these are clinicians who are used to adopting novel research-based tools and techniques.

We also elicited feedback from a Community Advisory Board (CAB) made up of six members with lived experience with mental health disorders and CBTs. After having unlimited access to Socrates 2.0 for one week and gathered qualitative feedback, we found the CABs response to Socrates 2.0 to be overwhelmingly positive. They saw great value in being able to connect with the AI therapist at anytime from anywhere, "I really like how it's always there-

10

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

even in the middle of the night.” Some specifically liked that it was not a real therapist, noting “I feel like I can tell [Socrates 2.0] things I wouldn’t say to a human therapist. I like feeling anonymous and it’s non-judgmental.” The CAB found the tool more engaging than common CBT worksheets and suggested they would be more likely to complete out-of-session practice assignments if it involved using Socrates 2.0. User data from that week showed that CAB individuals engaged with Socrates multiple times, with some users utilizing Socrates for up to an hour at a time. The CAB also provided feedback on changes they would like to see with Socrates 2.0, many of which were cosmetic and accessibility-related changes (“I’d like the option to customize my experience, like changing font size, colors, and the background,” “I want to be able to use voice rather than text”), which are currently in development for future versions of the tool.

Feasibility Study:

Following the initial development of Socrates 2.0 and the obtained data from initial users, we devised a feasibility study to formally gather data about users’ experience of interacting with the tool. The goal is to gather feasibility and acceptability data from both patients and clinicians, as information from both groups is critical to inform the eventual implementation of generative AI tools like Socrates 2.0 in mental health treatment. We designed this study to better understand how users engage with the tool including the interaction frequency, length, and time, users’ satisfactions with the tool overall as well as the quality of each dialogue and even individual responses. Through the feasibility study, we also plan to obtain information about the ways in which the tool should be improved before it will be tested in trials designed to examine the efficacy of Socrates 2.0.

Method

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

Study Design:

The study will employ a mixed-methods design, combining quantitative and qualitative methods to assess the feasibility of Socrates 2.0 from the perspective of users (i.e., potential patients) and clinicians. The decision to obtain feedback from both users and clinicians was made so that our team could better understand how users would perceive the tool, but also to understand how clinicians could see integrating such a tool into their treatment. The mixed-methods approach enables a comprehensive evaluation of the tool from multiple perspectives, providing both quantitative feasibility, acceptability, satisfaction, and usage data for statistical analysis, as well as rich contextual data for a qualitative analysis of users' experiences and perceptions. The combination of quantitative and qualitative data captured throughout this study will be used to make further improvements to Socrates 2.0.

Individuals who consent will be asked to complete a brief semi-structured interviews and surveys before and after using Socrates 2.0. Users will be granted access to Socrates 2.0 for four weeks; clinicians will be granted free unlimited access to Socrates for two weeks. During the access period, individuals will be asked to use the tool at their convenience and as often as they wish. To maintain confidentiality, a unique identifier (i.e., combination of random numbers and letters) will be assigned to each participant to track usage data.

Study Setting:

The study will be conducted through the Road Home Program: National Center of Excellence for Veterans and Their Families at Rush University Medical Center in Chicago, IL. All qualitative interviews will be conducted via approved and secure video technology (e.g., Microsoft Teams) and quantitative feasibility, acceptability, and satisfaction data will be obtained via REDCap surveys emailed to participants.

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

Recruitment:

A total of 100 users and 50 clinicians will be recruited to inform the eventual implementation of generative AI tools like Socrates 2.0 in mental health treatment. Since Socrates is a web-based application and can be accessed via the internet from anywhere, recruitment is open to open to all individuals, not just patients and providers at Rush University Medical Center. Participants will be recruited from a variety of sources to ensure a diverse sample. Recruitment will occur via The New Normal (<https://bethenewnormalmatch.org>), Rush University Medical Center's website, social media, and word of mouth.

Eligibility:

Interested individuals will be screened for general eligibility and will be provided with an electronic version of the consent form. Individuals will be informed that they can contact study staff before signing the consent form if they have questions. Individuals are eligible to participate in the study if they are 18 years of age or older and have reading and writing proficiency in the English language at least at the 6th grade level. Since Socrates 2.0 is a web-based application, participants must have access to a stable internet connection and a device such as a computer, tablet, or smartphone to use the tool, and possess the basic technological skills necessary to navigate the tool effectively. Additionally, participants must be willing to test and engage in Socratic dialogue with the tool. Aside from not meeting the inclusion criteria detailed above, there are no specific exclusionary criteria.

Measures – Users:

Following the consent process, participants will be assessed using a combination of surveys and semi-structured interviews before and after their four-week engagement with Socrates 2.0. In addition, objective data based on their interactions with Socrates will be obtained

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

by the research team. The specific data and timepoints at which data is collected are detailed below. Users and clinicians are asked to complete slightly different assessments; clinician measures are presented below.

Demographic Characteristics. Participants will be asked to complete a brief demographic questionnaire at baseline asking them about their age, gender, sexual orientation, ethnicity, race, education, marital status, employment status, household income, and service member/veteran status. The information will be used to characterize the sample and to determine whether any of the participant characteristics are associated with feasibility, acceptability, and satisfaction, as well as the ways in which participants engage with Socrates 2.0.

Pre-Post-Socrates 2.0 Mental Health Screening Assessments. Participants will be asked to complete brief validated self-report assessments for various mental health areas at baseline and after their month of access to Socrates 2.0. Mental health screening assessments will be used to characterize the sample, understand how mental health symptoms at baseline may impact engagement with the tool, determine whether mental health symptoms change over the course of the study, and explore whether such change may be statistically associated with the use of Socrates. Brief screening assessments were chosen to reduce participant burden. To assess depression and generalized anxiety, participants will be asked to complete the Patient Health Questionnaire-4 (PHQ-4).²⁶ The Interaction Anxiousness Scale-3 (IAS-3)²⁷ will be administered to assess social anxiety. Obsessive compulsive tendencies will be assessed using the Obsessive-Compulsive Inventory-4 (OCI-4).²⁸ Traumatic stress symptoms and potential manic or psychotic symptoms will be assessed via the PTSD Checklist-4 (PCL-4)²⁹ and Washington Early Recognition Center Affectivity and Psychosis (WERCAP) screen,³⁰ respectively.

Prior to Engaging with Socrates. Participants will be asked to complete a brief semi-

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

structured qualitative interview with trained study staff. The interview is intended to better understand participants' familiarity with AI tools, mental health apps, and prior engagement with mental health services. Participants will be asked semi-structured questions. Each structured question may be followed up with encouragers to expand on the answers participants provide:

- Have you ever used an AI tool before?
 - If yes, what kinds? How often? For what purpose?
- Are you currently using, or have you used apps/tools for mental health before (e.g., Headspace, Calm, BetterHelp)?
- Have you ever engaged with mental health services?
 - If yes, what kinds?
- Have you ever used an AI tool for mental health services?

Following the semi-structured interview, study staff will provide participants with an overview of Socrates 2.0 and walk them through how to use it. Study staff will ensure that participants can log into Socrates 2.0 and answer any questions participants may have about using the tool. If requested by the participant, study staff can be available to answer questions about the tool following the initial overview.

While Engaging with Socrates 2.0. Participants can provide “thumbs up” or “thumbs down” ratings for responses that Socrates provides. This is optional and intended for participants to indicate particularly helpful or unhelpful responses.

After Engaging with Socrates 2.0. Participants will be able to complete a brief Likert scale survey about satisfaction with and the perceived helpfulness at the end of each exchange. Participants will also be provided with an option to provide any open-ended feedback they may want to share with the study team about the interaction.

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

1-Month After First Engaging with Socrates 2.0. One month after first engaging with Socrates 2.0, participants will be asked to complete a brief survey containing validated measures to assess their experiences and be asked to participate in a semi-structured qualitative interview. To assess participants' subjective assessment of usability, they will be asked to complete the System Usability Scale (SUS),³¹ which is a valid and widely used 10-item measure scored on a 5-point scale (strongly disagree to strongly agree). The Acceptability of Intervention Measure, Intervention Appropriateness Measure, Feasibility of Intervention Measure (AIM, IAM, FIM)³² is a validated and widely used 12-item measure scored on a 5-point scale (completely disagree to completely agree) that will be administered to determine the perceived acceptability and feasibility of Socrates 2.0. Participants will also be asked to complete the mHealth App Usability Questionnaire (MAUQ),³³ which is a validated and widely used 18-item measure scored on a 7-point scale (strongly disagree to strongly agree) that assesses ease of use, satisfaction, and usefulness of mobile applications. Finally, the Working Alliance Inventory- Short Revised (WAI-SR),³⁴ which is a validated 12-item measure scored on a 5-point scale (seldom to always, or always to seldom) will be used to examine participants' experiences with their AI therapist. Participants will also be asked to complete a semi-structured interview surrounding the following questions to better understand their use and impressions of the tool:

- What are your overall impressions of Socrates?
- How often did you use Socrates?
 - What prompted you to use Socrates in general and different times you did?
- What would you change about Socrates? Why?
- How would Socrates be helpful to you or other individuals?
 - Why might it not be helpful?

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

- How likely would you be to use it going forward?
- What would make you more likely to use it?
- Who do you believe could benefit most from Socrates?
- How would you present Socrates to others to make them interested and comfortable with checking out the tool?
- How do you think Socrates fits with mental health care?
- Has working with Socrates changed your perception on the utility of AI tools more broadly? (If so, how?)
- Is there anything else you would like to share that I have not asked you about?

Objective Utilization Metrics. Objective metrics such as the frequency of use, number of exchanges, duration of interactions, and retention rates will be automatically collected via the tool and used to evaluate the feasibility of Socrates 2.0. Specifically, we will examine engagement rates to determine how often and how long users engage with the tool per session and in total over the study period, as well as determine the time since they last used the tool. We will also assess how often users complete the entire interaction/dialogue process with the tool. Finally, we will track LLM token use for each exchange to help with future cost estimations and performance evaluations.

Measures – Clinicians:

As described above, both users and clinicians will be asked to participate. Clinicians will complete the same demographics survey, will have access to the same tool, and will be prompted to answer the same questions while using Socrates 2.0. Unlike the users, however, clinicians will only be granted two weeks of access to Socrates 2.0. Our team determined that this would likely be sufficient time for them to form opinions and evaluate how the tool may fit with their

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

therapeutic work. The team will obtain information about clinician characteristics, such as their experience (e.g., number of years in the field), average caseload, populations with whom they work, and whether they are currently providing any interventions that involve cognitive restructuring or Socratic dialogue, among others. In addition, the team will ascertain the clinicians' familiarity with AI in mental healthcare. Using semi-structured interviews, clinicians will be asked whether they have used AI tools before and whether they have recommended them to patients as well as which types of patients they believe may benefit the most and least from tools like Socrates 2.0. The team will also ask clinicians about information they believe would be needed in order for them to feel comfortable recommending AI tools to patients and how AI tools fit with the clinicians' treatment philosophies.

Following their two-week access to Socrates 2.0, clinicians will complete similar measures to the other users, including the SUS,³¹ AIM, IAM, FIM,³² and MAUQ.³³ Clinicians will not be asked about perceived working alliance with the AI therapist, as we anticipate that they will interact with the tool differently than actual users, such as by roleplaying some of their own clients to see what answers Socrates 2.0 provides. The semi-structured interviews differ from those of the users as they are intended to generate information about clinicians' perceptions of the tool, aspects they would like to see changed, and how it could fit with their practice. Many of the questions are repeated from the pre-tool interview, such as questions surrounding the usefulness of AI tools like Socrates 2.0 as well as which clients may be most and least likely to benefit from such a tool. Repeating interview questions will also enable the study team to determine whether the clinicians' perceptions and answers change after they have had the opportunity to engage with Socrates 2.0, as it may be more challenging for individuals to imagine what things could look like before directly interacting with the tool.

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

Statistical Analyses:

Quantitative Data Analysis. Quantitative data, such as from mental health screening assessments and the validated feasibility measures will be analyzed using appropriate descriptive and inferential statistics. Specifically, we will first characterize both the user and clinician samples based on demographic characteristics and, for users, the baseline mental health characteristics. Descriptive statistics will also be used to detail the responses to the assessments administered after using Socrates 2.0, as well as the overall use frequency, duration, and the use of tokens, among others. Any changes over time, such as for mental health symptoms, will be described via commonly used effect sizes (e.g., Cohen's *d*). Appropriate inferential statistics will be used to determine whether certain demographic or baseline mental health characteristics were associated with the use of Socrates 2.0. We will also evaluate whether the use of Socrates 2.0 influenced assessment ratings, as it is possible that those who used the tool more regularly may have different perceptions of it than infrequent users. Moreover, as Socrates 2.0 was developed to engage individuals in Socratic dialogue, it is possible that regular use of the tool can result in reductions in various mental health symptoms, which we will also examine using appropriate inferential statistics.

Qualitative Data Analysis. Qualitative data will be analyzed using thematic analysis,³⁵ a method for identifying, analyzing, and reporting patterns within data. In thematic analysis will begin with a meticulous process of data familiarization, where researchers immerse themselves in the data to gain a deep understanding of its content. Following this, data will be systematically coded, with codes representing the smallest units of meaning within the data that are relevant to the research questions. These codes will then be collated into potential themes, which are broader patterns of meaning that emerge from the coded data. Each theme will be reviewed and refined to

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

ensure it accurately reflects the coded data and the overall data set. The final step involves defining and naming themes, providing a detailed analysis of each theme and the overall story the data tells in relation to the research questions. All quantitative and qualitative analyses will be performed by approved study staff who have the necessary training.

Ethical Considerations:

The study procedures received approval from the Rush University Medical Center Institutional Review Board. Each prospective participant will receive an electronic copy of the informed consent and have the opportunity to have any study-related inquiries addressed by members of the research team. Participation is contingent upon completion of the electronic consent process. Individuals who do not meet the study criteria or express a need for additional mental health resources will receive information from a member of the research team and be directed to appropriate support services.

Eligible participants will gain access to and receive an introduction to the tool. During both the informed consent process and the introductory session, participants will be informed that the research team does not actively monitor Socrates 2.0 and they should seek appropriate assistance in case of a crisis. Essential information, such as crisis helplines and instructions to dial 911 or visit the nearest emergency room, are prominently and permanently displayed on the main screen upon logging into Socrates 2.0 (see Figure 1).

Before the start of the feasibility study, Socrates 2.0 underwent rigorous security testing by an external firm and was cleared for use by Rush University Medical Center's cybersecurity team. The results of this testing confirmed the security of the web-based platform, encompassing both access to the tool and the storage of resultant data. These formal procedures aimed to minimize the risk of potential data breaches or loss. Usernames and passwords, generated by the

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

research team using industry-standard protocols, will be linked to eligible participants via de-identified study participant codes. Access to the tool can be promptly revoked by the research team, if necessary. Through these measures, the team has tried to mitigate risks associated with conducting a feasibility study using a web-based platform.

Results

There are no results to report to date. Enrollment started in February 2024 and is expected to continue until February 2025. All (i.e., positive and negative) findings from this feasibility study will be presented in manuscripts submitted to peer-reviewed scientific publications relevant to Socrates 2.0. Results will also be shared via presentations at scientific conferences as symposia and posters, as well as be shared via Rush University Medical Center's website and social media.

Discussion

Socrates 2.0 is one of the first functional multi-agent generative AI tools intended to complement existing cognitive behavioral therapies by engaging users in Socratic dialogue. The feasibility study will contribute to our understanding of whether and how individuals engage with Socrates 2.0, and whether they find their interactions to be beneficial. Importantly, the feasibility study will also provide important additional insights into the safety of the tool, expanding on previous safety testing. Moreover, this study will help us evaluate whether there are certain scenarios in which Socrates 2.0 does not perform as well as anticipated and will enable us to fine-tune future versions of the tool to improve its performance.

By also incorporating feedback from clinicians, this study will provide valuable insights into the potential integration of generative AI tools like Socrates 2.0 into existing therapeutic practices. Positive findings could pave the way for subsequent studies evaluating the combined

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

use of Socrates 2.0 and cognitive behavioral therapies, potentially enhancing engagement compared to conventional worksheets and facilitating skill practice, leading to improved outcomes. Conversely, if the study yields negative results, indicated by low feasibility or satisfaction ratings or significant safety concerns, the feedback will be utilized to refine the tool further.

Limitations:

The feasibility study has several limitations. First, although Socrates 2.0 is intended for use in mental health contexts, it is open to all users, not just current or former mental health patients. This broad inclusion may impact feasibility, acceptability, and satisfaction ratings, potentially varying between those with and without diagnosed mental health conditions. To address this, we will administer screening assessments to gather mental health information and explore potential associations between symptom severity and feasibility ratings. Second, while Socrates 2.0 is primarily intended to complement CBTs by facilitating cognitive restructuring, this feasibility study initially evaluates it as a standalone tool. Thus, the obtained feasibility results from using Socrates 2.0 in isolation may differ compared to those that would be obtained following the integration with psychotherapy. Third, the study's description explicitly mentions testing a mental health-focused generative AI tool, which might deter users and clinicians less receptive to this novel technology from participating. Consequently, this self-selection bias could skew findings related to feasibility, acceptability, and satisfaction. Lastly, Socrates 2.0 was purposefully designed to facilitate Socratic dialogue. Therefore, the insights gleaned from this study will not be universally applicable to other generative AI-based mental health tools designed for different purposes.

Conclusions:

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

The development of Socrates 2.0 marks a significant advancement in the use of generative AI in the context of CBT and mental healthcare more broadly. By combining multiple collaborative AI agents, Socrates 2.0 facilitates Socratic dialogue on a range of beliefs while minimizing common LLM issues, such as looping. Initial positive user and clinician feedback suggests that generative AI tools, such as Socrates 2.0, may exceed expectations. However, formal studies, such as the feasibility study described in this manuscript and tests of tools such as Socrates 2.0 in conjunction with CBT are needed to better understand the true feasibility, as well as potential risks and problems, of generative AI tools as complements to CBT. The outlined feasibility study will provide user and clinician feedback that can be used to evaluate the tool and further refine it for broader application and testing its efficacy with psychotherapy. We hope that the description of Socrates 2.0 and the associated feasibility trial will serve as a valuable model for other researchers and developers aiming to harness the power of AI in mental health.⁷

23

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

Data Availability:

Data sharing is not applicable to this article as no data sets were generated or analyzed.

Authors' Contributions:

PH was responsible for conceptualization, methodology, writing (original draft), and funding acquisition. SPr was responsible for conceptualization methodology, writing (review & editing), and project administration. YC was responsible for resources and data curation. ZA was responsible for resources and data curation, and writing (review & editing). DA was responsible for project administration. SPo was responsible for resources, data curation, and writing (review & editing).

Conflicts of Interest:

None declared.

References

1. Coombs NC, Meriwether WE, Caringi J, Newcomer SR. Barriers to healthcare access among U.S. adults with mental health challenges: A population-based study. *SSM - Popul Health*. 2021;15:100847. doi:10.1016/j.ssmph.2021.100847
2. American Psychological Association. 2022 COVID-19 Practitioner Impact Survey.; 2022. Accessed January 16, 2024. <https://www.apa.org/pubs/reports/practitioner/2022-covid-psychologist-workload>
3. American Psychological Association. 2023 Practitioner Pulse Survey.; 2023. Accessed January 16, 2024. <https://www.apa.org/pubs/reports/practitioner/2023-psychologist-reach-limits>
4. Insel T. Digital mental health care: five lessons from Act 1 and a preview of Acts 2–5. *Npj Digit Med*. 2023;6(1):9. doi:10.1038/s41746-023-00760-8
5. Philippe TJ, Sikder N, Jackson A, et al. Digital Health Interventions for Delivery of Mental Health Care: Systematic and Comprehensive Meta-Review. *JMIR Ment Health*. 2022;9(5):e35159. doi:10.2196/35159
6. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. *J Med Internet Res*. 2021;23(1):e17828. doi:10.2196/17828
7. Stade EC, Stirman SW, Ungar LH, et al. Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation. *PsyArXiv*; 2023. doi:10.31234/osf.io/cuzvr
8. D'Alfonso S. AI in mental health. *Curr Opin Psychol*. 2020;36:112-117. doi:10.1016/j.copsyc.2020.04.005

25

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

9. Vittorio LN, Murphy ST, Braun JD, Strunk DR. Using Socratic Questioning to promote cognitive change and achieve depressive symptom reduction: Evidence of cognitive change as a mediator. *Behav Res Ther.* 2022;150:104035. doi:10.1016/j.brat.2022.104035
10. Schmidt ID, Pfeifer BJ, Strunk DR. Putting the “cognitive” back in cognitive therapy: Sustained cognitive change as a mediator of in-session insights and depressive symptom improvement. *J Consult Clin Psychol.* 2019;87(5):446-456. doi:10.1037/ccp0000392
11. Hofmann SG, Asnaani A, Vonk IJJ, Sawyer AT, Fang A. The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses. *Cogn Ther Res.* 2012;36(5):427-440. doi:10.1007/s10608-012-9476-1
12. Resick PA, Monson CM, Chard KM. *Cognitive Processing Therapy for PTSD: A Comprehensive Manual.* Guilford Press; 2017.
13. Braun JD, Strunk DR, Sasso KE, Cooper AA. Therapist use of Socratic questioning predicts session-to-session symptom change in cognitive therapy for depression. *Behav Res Ther.* 2015;70:32-37. doi:10.1016/j.brat.2015.05.004
14. Braun JD. *Socratic Questioning and Session-to-Session Symptom Severity in CT for Depression: An Examination of Moderators and Mediators.* The Ohio State University; 2018. Accessed January 29, 2024. https://etd.ohiolink.edu/acprod/odb_etd/etd/r/1501/10?clear=10&p10_accession_num=osu1536691280047438
15. Farmer CC, Mitchell KS, Parker-Guilbert K, Galovski TE. Fidelity to the Cognitive Processing Therapy Protocol: Evaluation of Critical Elements. *Behav Ther.* 2017;48(2):195-206. doi:10.1016/j.beth.2016.02.009
16. LeBeau RT, Davies CD, Culver NC, Craske MG. Homework Compliance Counts in Cognitive-Behavioral Therapy. *Cogn Behav Ther.* 2013;42(3):171-179.

26

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

doi:10.1080/16506073.2013.763286

17. Hara KM, Aviram A, Constantino MJ, Westra HA, Antony MM. Therapist empathy, homework compliance, and outcome in cognitive behavioral therapy for generalized anxiety disorder: partitioning within- and between-therapist effects. *Cogn Behav Ther.* 2017;46(5):375-390. doi:10.1080/16506073.2016.1253605
18. Kazantzis N, Whittington C, Zelencich L, Kyrios M, Norton PJ, Hofmann SG. Quantity and Quality of Homework Compliance: A Meta-Analysis of Relations With Outcome in Cognitive Behavior Therapy. *Behav Ther.* 2016;47(5):755-772. doi:10.1016/j.beth.2016.05.002
19. Mausbach BT, Moore R, Roesch S, Cardenas V, Patterson TL. The Relationship Between Homework Compliance and Therapy Outcomes: An Updated Meta-Analysis. *Cogn Ther Res.* 2010;34(5):429-438. doi:10.1007/s10608-010-9297-z
20. Simons AD, Marti CN, Rohde P, Lewis CC, Curry J, March J. Does Homework “Matter” in Cognitive Behavioral Therapy for Adolescent Depression? *J Cogn Psychother.* 2012;26(4):390-404. doi:10.1891/0889-8391.26.4.390
21. Tang W, Kreindler D. Supporting Homework Compliance in Cognitive Behavioural Therapy: Essential Features of Mobile Apps. *JMIR Ment Health.* 2017;4(2):e20. doi:10.2196/mental.5283
22. Gaynor ST, Lawrence PS, Nelson-Gray RO. Measuring Homework Compliance in Cognitive-Behavioral Therapy for Adolescent Depression: Review, Preliminary Findings, and Implications for Theory and Practice. *Behav Modif.* 2006;30(5):647-672. doi:10.1177/0145445504272979
23. Li Y. A Practical Survey on Zero-shot Prompt Design for In-context Learning. Published

27

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

online 2023. doi:10.48550/ARXIV.2309.13205

24. Talebirad Y, Nadiri A. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. Published online 2023. doi:10.48550/ARXIV.2306.03314
25. Carmel A, Villatte JL, Zachary Rosenthal M, Chalker S, Comtois KA. Applying Technological Approaches to Clinical Supervision in Dialectical Behavior Therapy: A Randomized Feasibility Trial of the Bug-in-the-Eye (BITE) Model. *Cogn Behav Pract*. 2016;23(2):221-229. doi:10.1016/j.cbpra.2015.08.001
26. Löwe B, Wahl I, Rose M, et al. A 4-item measure of depression and anxiety: Validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *J Affect Disord*. 2010;122(1-2):86-95. doi:10.1016/j.jad.2009.06.019
27. Nichols AL, Webster GD. Designing a brief measure of social anxiety: Psychometric support for a three-item version of the Interaction Anxiousness Scale (IAS-3). *Personal Individ Differ*. 2015;79:110-115. doi:10.1016/j.paid.2015.01.043
28. Abramovitch A, Abramowitz JS, McKay D. The OCI-4: An ultra-brief screening scale for obsessive-compulsive disorder. *J Anxiety Disord*. 2021;78:102354. doi:10.1016/j.janxdis.2021.102354
29. Zuromski KL, Ustun B, Hwang I, et al. Developing an optimal short-form of the PTSD Checklist for DSM-5 (PCL-5). *Depress Anxiety*. 2019;36(9):790-800. doi:10.1002/da.22942
30. Mamah D, Owoso A, Sheffield JM, Bayer C. The WERCAP Screen and the WERC Stress Screen: psychometrics of self-rated instruments for assessing bipolar and psychotic disorder risk and perceived stress burden. *Compr Psychiatry*. 2014;55(7):1757-1771. doi:10.1016/j.comppsy.2014.07.004
31. Lewis JR. The System Usability Scale: Past, Present, and Future. *Int J Human-Computer*

28

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

Interact. 2018;34(7):577-590. doi:10.1080/10447318.2018.1455307

32. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. *Implement Sci.* 2017;12(1):108. doi:10.1186/s13012-017-0635-3
33. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): Development and Validation Study. *JMIR MHealth UHealth.* 2019;7(4):e11500. doi:10.2196/11500
34. Munder T, Wilmers F, Leonhart R, Linster HW, Barth J. Working Alliance Inventory-Short Revised (WAI-SR): psychometric properties in outpatients and inpatients. *Clin Psychol Psychother.* 2010;17(3):231-239. doi:10.1002/cpp.658
35. Riessman CK. *Narrative Methods for the Human Sciences*. SAGE Publications, Inc; 2007.

29

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

Figure 1: Socrates 2.0 User Interface and Dialogue Example

RUSH Socrates 2.0: AI-powered Socratic Dialogue

Start New Dialogue [dropdown] Select Previous Dialogue

Hello, I am Socrates 2.0, an AI language model here to engage in a Socratic dialogue with you. Please remember not to share personal or sensitive information with me. What belief would you like to explore and potentially change today?

If I had dressed differently, I would not have been assaulted.

I'm sorry to hear that you experienced something so distressing. It sounds like you've been carrying a heavy thought. Could you tell me more about what leads you to believe that the way you were dressed could have influenced what happened?

The dress I was wearing was pretty suggestive. It probably caught the perp's attention.

It's understandable to try to make sense of such a difficult situation. Do you think that everyone who sees someone in suggestive clothing has the same reaction?

No! Definitely not! At least I would hope not...

It's good that you recognize not everyone would react the same way. Considering that, what do you think really determines someone's decision to harm another person?

I don't know. You probably have to be pretty sick in your head and then find a suitable target.

It seems like you're suggesting that the intention to harm comes from the person themselves rather than what the potential target is wearing. Does that sound accurate to you?

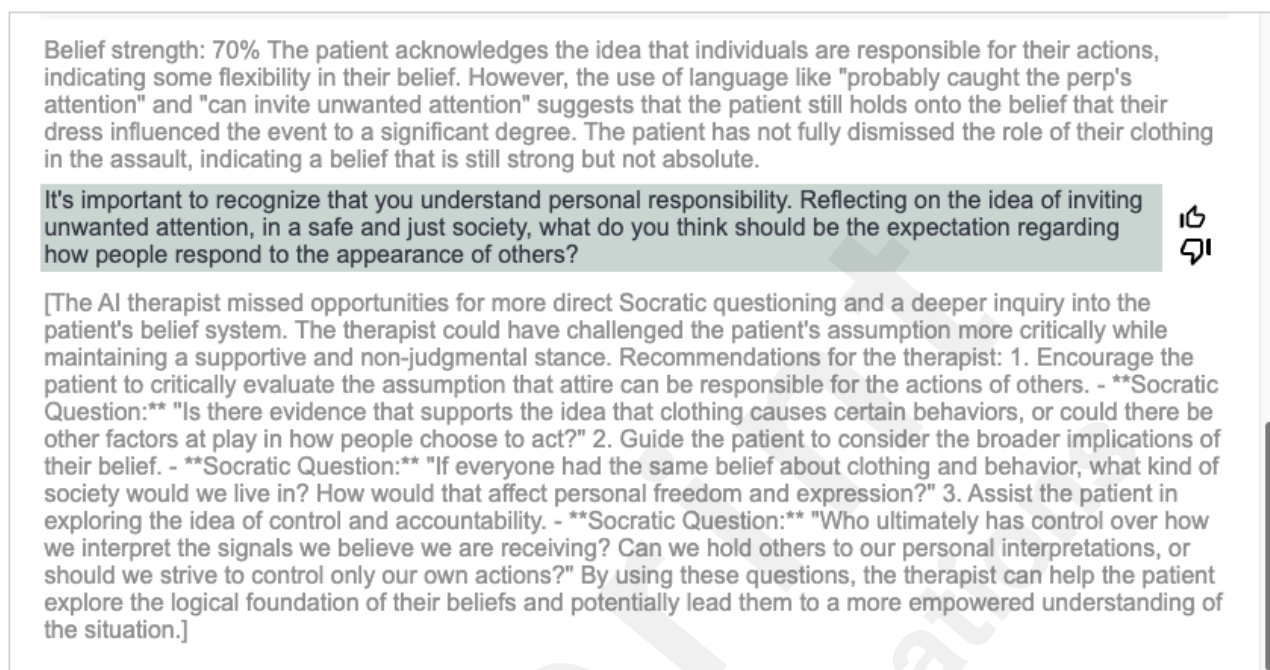
Respond to Socrates. Please DO NOT enter any personal or sensitive information.

Report Problem End Dialogue Send

Socrates is a generative AI tool and is not capable of handling crisis situations. Socrates may also not provide accurate crisis line information or other resources when requested.

If you or someone you know is struggling, in a mental health crisis, or having thoughts of suicide, call or text the [988 Suicide & Crisis Lifeline](https://988lifeline.org) at **988** or chat at 988lifeline.org. This service is confidential, free, and available 24 hours a day, 7 days a week. In life-threatening situations, call **911**.

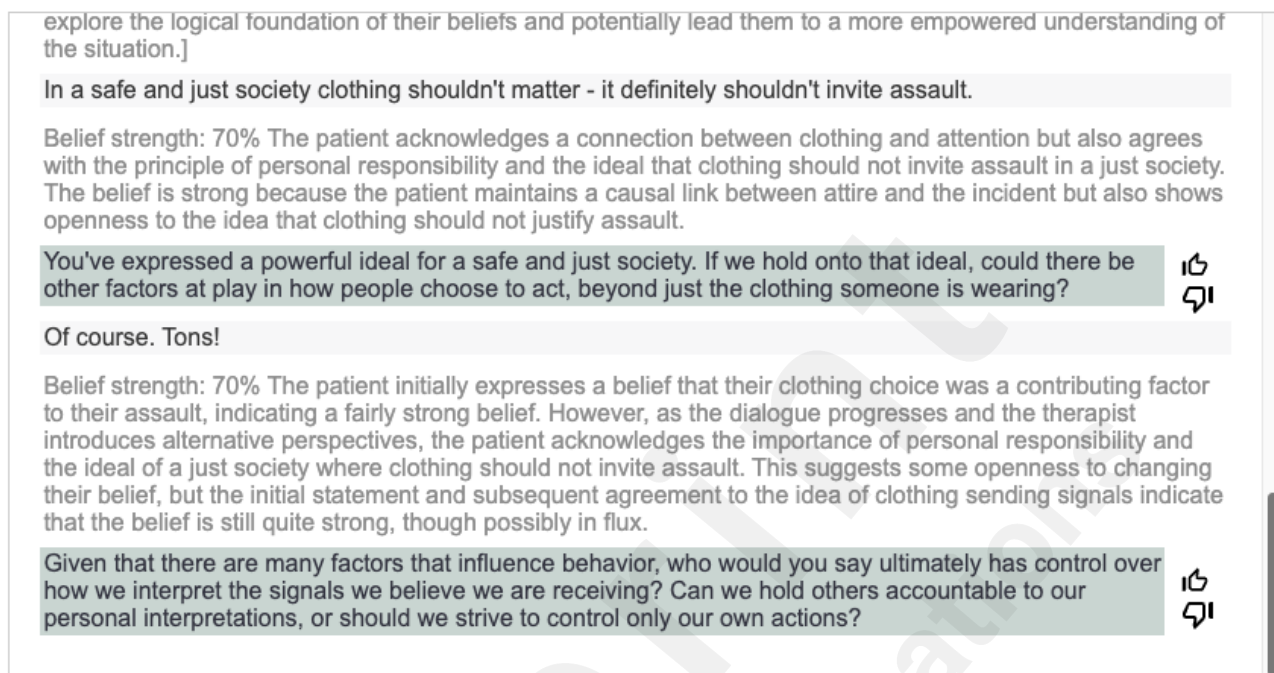
FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

Figure 2: Socrates 2.0 AI External Rater and AI Supervisor Feedback

Note: The first gray statement is the AI external rater. The second gray statement in brackets is the AI supervisor. Both responses are hidden from the user; this is a screenshot for illustrative purposes from the developer mode of Socrates 2.0.

31

FEASIBILITY OF A GENERATIVE AI FACILITATED SOCRATIC DIALOGUE TOOL

Figure 3: Socrates 2.0 Integration of Supervisor Feedback.

Note: The gray statement without brackets is the AI external rater. AI external rater responses are hidden from the user; this is a screenshot for illustrative purposes from the developer mode of Socrates 2.0.