nature medicine

Article

https://doi.org/10.1038/s41591-024-03252-8

Digital AVATAR therapy for distressing voices in psychosis: the phase 2/3 AVATAR2 trial

Received: 10 April 2024

Accepted: 15 August 2024

Published online: 28 October 2024

Check for updates

A list of authors and their affiliations appears at the end of the paper

Distressing voices are a core symptom of psychosis, for which existing treatments are currently suboptimal; as such, new effective treatments for distressing voices are needed. AVATAR therapy involves voice-hearers engaging in a series of facilitated dialogues with a digital embodiment of the distressing voice. This randomized phase 2/3 trial assesses the efficacy of two forms of AVATAR therapy, AVATAR-Brief (AV-BRF) and AVATAR-Extended (AV-EXT), both combined with treatment as usual (TAU) compared to TAU alone, and conducted an intention-to-treat analysis. We recruited 345 participants with psychosis; data were available for 300 participants (86.9%) at 16 weeks and 298 (86.4%) at 28 weeks. The primary outcome was voice-related distress at both time points, while voice severity and voice frequency were key secondary outcomes. Voice-related distress improved, compared with TAU, in both forms at 16 weeks but not at 28 weeks. Distress at 16 weeks was as follows: AV-BRF, effect −1.05 points, 96.5% confidence interval (CI) = −2.110 to 0, $P = 0.035$, Cohen's $d = 0.38$ (CI = 0 to 0.767); AV-EXT −1.60 points, 96.5% CI = −3.133 to −0.058, $P = 0.029$, Cohen's $d = 0.58$ (CI = 0.021 to 1.139). Distress at 28 weeks was: AV-BRF, −0.62 points, 96.5% CI = −1.912 to 0.679, $P = 0.316$, Cohen's $d = 0.22$ (CI = −0.247 to 0.695); AV-EXT −1.06 points, 96.5% CI = −2.700 to 0.586, $P = 0.175$, Cohen's $d = 0.38$ (CI = −0.213 to 0.981). Voice severity improved in both forms, compared with TAU, at 16 weeks but not at 28 weeks whereas frequency was reduced in AV-EXT but not in AV-BRF at both time points. There were no related serious adverse events. These findings provide partial support for our primary hypotheses. AV-EXT met our threshold for a clinically significant change, suggesting that future work should be primarily guided by this protocol. ISRCTN registration: ISRCTN55682735.

Digital innovation carries the promise of transforming mental health treatment, addressing long-standing issues in access, engagement and effectiveness[1,2]. Auditory verbal hallucinations (henceforth voices), commonly associated with a diagnosis of schizophrenia, are often distressing and impair quality of life. However, the response to pharmacological and psychological treatments is suboptimal[3,4], highlighting the need for new interventions. AVATAR therapy is one such digital innovation, which targets voices[5]. It is part of a wave of relational approaches, informed by advances in theory, which position voice-hearing as an experience of social communication[6,7]. The defining aspect of AVATAR therapy is the digital embodiment of the distressing voice in the form of an avatar. Bespoke software enables the voice-hearer to customize how the avatar looks and sounds. Treatment is focused on a series of 'face-to-face' dialogues between the person and their avatar, supported by the therapist. The aim is to reduce voice-related distress and build empowerment in daily life.

A proof-of-concept study found that a six-session course of AVATAR therapy was safe, with positive effects on voice severity[5].

e-mail: Philippa.Garety@kcl.ac.uk; Thomas.Ward@kcl.ac.uk

A previous fully powered single-site randomized controlled trial (AVATAR1) compared AVATAR therapy with supportive counseling[8] and demonstrated a substantial reduction in the severity of voices in the AVATAR therapy group at 12 weeks. An independent pilot also reported feasibility and efficacy findings[9]. Examination of AVATAR1 therapy content identified a wide range of potential treatment targets, including developmental trauma[10], suggesting that the intervention could be optimized through personalization to diverse voice-hearer experiences[11,12]. Early evidence for AVATAR therapy is based on delivery by a small and experienced cohort of therapists within research settings. There is consequently a need to test effectiveness when treatment is delivered by a wider workforce, across geographically and demographically diverse locations, including frontline mental health services.

The main objective of this late phase 2/3 multisite AVATAR2 trial is to test, compared with treatment as usual (TAU) alone, the efficacy of two forms of AVATAR therapy, that is, AVATAR-Brief (AV-BRF), with a standardized focus on exposure, assertiveness and self-esteem, and AVATAR-Extended (AV-EXT), with a phase 1 mirroring AV-BRF, augmented by a more personalized, developmentally focused phase 2 based on the voice-hearer's life history. We hypothesized that both AV-BRF and AV-EXT, when added to TAU, would be superior to TAU alone at 16 and 28 weeks in reducing voice-related distress (primary outcome), frequency and severity (key secondary outcomes).

Results

Patient disposition

Between 1 January 2021 and 30 November 2022, we assessed 642 people for eligibility, recruiting 345 participants and randomly allocating them to AV-BRF (n = 116), AV-EXT (n = 114) and TAU control (n = 115). Data were available for 300 participants at the 16-week follow-up (86.9%) and 298 (86.4%) at 28 weeks; at the 16-week follow-up, 12 participants were lost in TAU, 17 in AV-BRF and 16 in AV-EXT; the numbers lost were 11 (TAU), 15 (AV-BRF) and 21 (AV-EXT) at 28 weeks (see Fig. 1 for the participant Consolidated Standards of Reporting Trials (CONSORT) diagram).

The participants' baseline demographic and clinical characteristics showed no differences between trial arms at baseline (Table 1). As is typical in a sample of people with psychosis, overall there was a greater proportion who were male (61.4%); most were single, unemployed and the most common diagnosis was schizophrenia (43.8%). Participants had been in contact with mental health services for an average of approximately 13 years (mean = 13.33, s.d. = 11.15), and approximately 40% belonged to a minoritized ethnic group. Their voices were assessed at baseline on the Psychotic Symptoms Rating Scale-Auditory Hallucinations (PSYRATS-AH) scale as similar in severity to those in the AVATAR1 trial[8], with high mean scores for voice severity[13]. On average, 61.2% reported highly characterized voices.

Treatment completion

A total of 95 of 116 (81.90%) participants assigned to the AV-BRF and 66 of 114 (57.89%) assigned to the AV-EXT completed treatment (against prespecified criteria, that is, four of six active sessions for AV-BRF and ten of 12 sessions for AV-EXT). Four (3.45%) participants allocated to AV-BRF and 37 (32.46%) to AV-EXT had partial treatment but did not reach the completion criterion. Seventeen people (14.66%) allocated to AV-BRF and 11 to AV-EXT (9.65%) attended no treatment sessions. For AV-BRF, the overall mean number of sessions attended was 5.11 (s.d. = 2.42; range = 0–8). For those who completed treatment, the mean was 6.16 sessions (s.d. = 0.94; range = 4–8). The mean active treatment session duration was 65.65 min (s.d. = 13.97, minimum = 30; maximum = 148), including a mean avatar dialogue duration of 9.51 min (s.d. = 3.79, minimum = 4, maximum = 28). For AV-EXT, the overall mean number of sessions was 8.18 (s.d. = 4.43; minimum = 0; maximum = 13). For those who completed treatment, the mean was 11.53 sessions (s.d. = 0.92; range = 10–13). The active treatment session time was 65.93 min (s.d. = 13.34; minimum = 20; maximum = 122), including a mean active dialogue duration of 10.47 min (s.d. = 3.95; minimum = 1; maximum = 27) (see Supplementary Materials 1–4 for additional data on treatment completion and mode of delivery).

Primary outcomes

As shown in Table 2 and Fig. 2, there was an improvement in voice-related distress on the distress subscale of the PSYRATS-AH-Distress (range = 0–20) in both forms at 16 weeks but not at 28 weeks (distress at 16 weeks: AV-BRF, effect −1.05 points, 96.5% confidence interval (CI) = −2.110 to 0, P = 0.035, Cohen's d = 0.38 (CI = 0 to 0.767); AV-EXT, −1.60 points, 96.5% CI = −3.133 to −0.058, P = 0.029, Cohen's d = 0.58 (CI = 0.021 to 1.139)). Distress at 28 weeks was as follows: AV-BRF, −0.62 points, 96.5% CI = −1.912 to 0.679, P = 0.316, Cohen's d = 0.22 (CI = −0.247 to 0.695); AV-EXT −1.06 points, 96.5% CI = −2.700 to 0.586, P = 0.175, Cohen's d = 0.38 (CI = −0.213 to 0.981). In the mixed-effects analysis of the primary outcome, the intraclass correlation coefficient (ICC) for the therapist clustering effect in both AV-EXT and AV-BRF was 0.054, indicating that approximately 5.4% of the residual variance in PSYRATS-AH-Distress was at the therapist level.

Key secondary outcomes

There was an improvement in PSYRATS-AH-Total voice severity (range = 0–44) in both forms at 16 weeks (AV-BRF, −2.04 points, 96.5% CI = −3.836 to −0.239, P = 0.017, Cohen's d = 0.45 (CI = 0.053 to 0.853); AV-EXT −2.32 points, 96.5% CI = −4.208 to −0.438, P = 0.009, Cohen's d = 0.52 (CI = 0.097 to 0.936)) but not at 28 weeks (AV-BRF, −1.61 points, 96.5% CI = −4.260 to 1.036, P = 0.199, Cohen's d = 0.36 (CI = −0.230 to 0.947); AV-EXT −1.87 points, 96.5% CI = −4.274 to 0.526, P = 0.100, Cohen's d = 0.42 (CI = −0.117 to 0.950)).

Voice frequency as measured by the PSYRATS-AH-Frequency subscale (range = 0–12) was significantly reduced in AV-EXT at both time points (16 weeks: −0.62 points, 96.5% CI = −1.140 to −0.104, P = 0.011, Cohen's d = 0.30 (CI = 0.051 to 0.556); 28 weeks: −0.89 points, 96.5% CI = −1.525 to −0.258, P = 0.003, Cohen's d = 0.43 (CI = 0.126 to 0.744)). Frequency was not reduced by AV-BRF at either time point (16 weeks: −0.50 points, 96.5% CI =−1.012 to 0.018, P = 0.042, Cohen's d = 0.24 (CI = −0.009 to 0.494); 28 weeks: −0.65 points, 96.5% CI = −1.331 to 0.030, P = 0.044, Cohen's d = 0.32 (CI = −0.015 to 0.649)).

Other secondary outcomes

Table 2 also shows the treatment effect estimates across all other secondary outcomes. For other voice-specific measures, there were improvements in voice acceptance and action for both AV-BRF and AV-EXT at both time points; for the Beliefs About Voices Questionnaire (BAVQ) malevolence or benevolence, there were no effects for AV-BRF or AV-EXT at either time point nor for omnipotence at 16 weeks, although there was an effect on omnipotence for AV-EXT only at 28 weeks; the Voice Power Differential Scale (VPDS) score improved in both arms at 16 weeks and in AV-EXT at 28 weeks; finally, there were no effects on the Hallucinations Remission Score in either arm at either time point. There were reductions in PSYRATS-Delusions and improvements in well-being (Warwick-Edinburgh Mental Well-being Scale (WEMWBS)) for AV-EXT at both time points and for AV-BRF at week 16. There were improvements in personal recovery (CHoice of Outcome In Cbt for PsychosEs (CHOICE)) for AV-EXT and AV-BRF at both time points, and reductions in anxiety, depression and stress (measured by the Depression, Anxiety and Stress Scale (DASS)) for AV-BRF at both time points but only at week 16 for AV-EXT. Depression measured by the Beck Depression Inventory (BDI) and anxiety in daily life measured with the experience sampling method (ESM) showed improvements at 16 but not 28 weeks in AV-EXT, but not AV-BRF. There were no effects on the International Trauma Questionnaire (ITQ) for either arm at either time point. Figure 3 summarizes the standardized mean differences on all outcomes at both time points (Extended Data Tables 1 and 2 give the descriptive statistics for all measures at each time point).

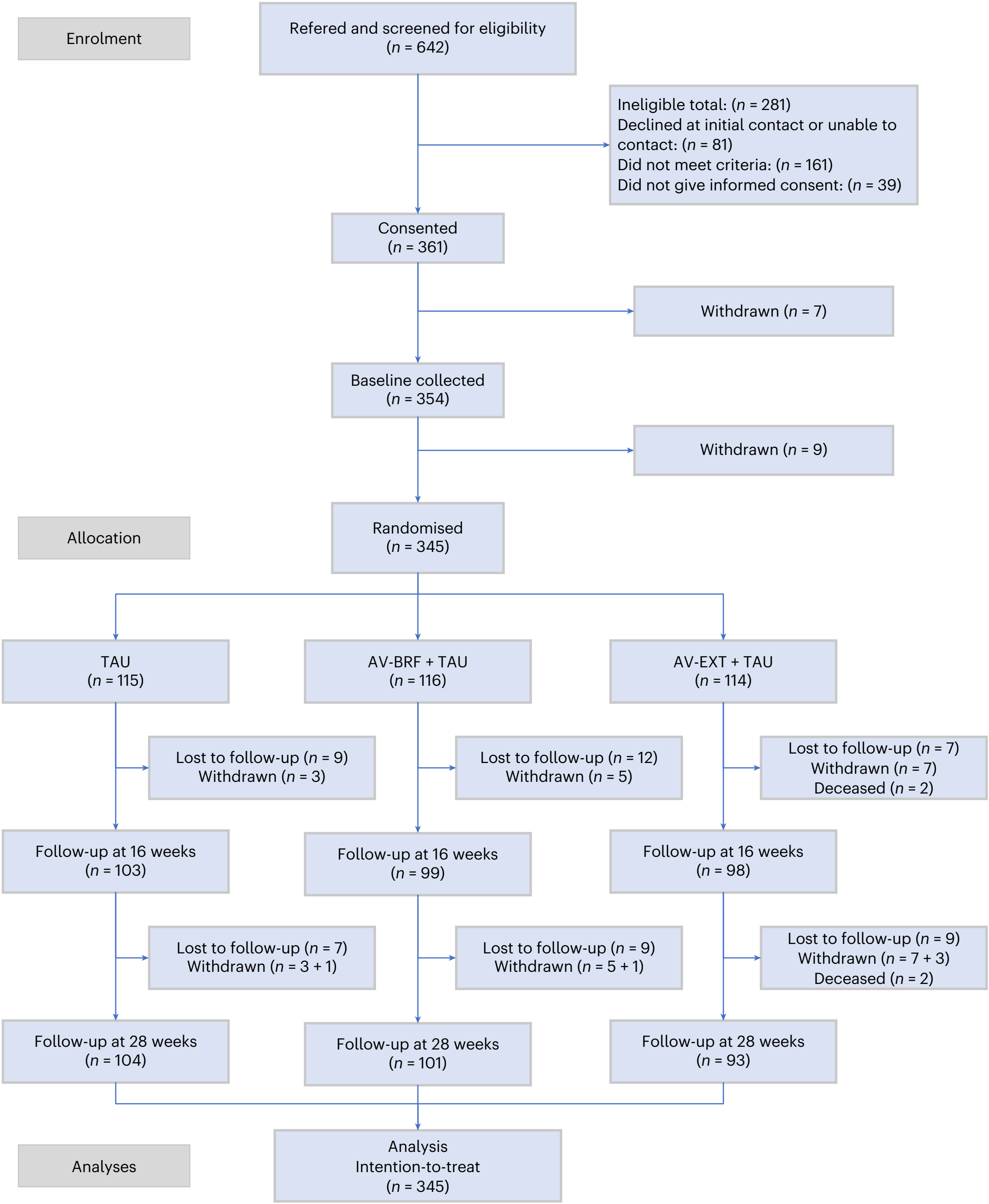


Fig. 1 | CONSORT diagram of all participants who were assessed for eligibility for the trial, randomized to AV-EXT + TAU, AV-BRF + TAU or TAU alone, and followed up to 28 weeks. Follow-up at 16 weeks: 16 weeks after baseline (post-treatment follow-up). Follow-up at 28 weeks: 28 weeks after the baseline.

Safety

Table 3 presents all serious adverse events (SAEs) according to arm and event type. There were 58 SAEs across 56 participants, with 51% of events occurring in the AV-EXT arm. Most events were admission to hospital for psychological health events and these occurred equally across arms. There were two deaths in the AV-EXT arm. One of the deaths was a suicide, which occurred in the context of a long-standing pattern of suicidality-related hospital admissions, with increased alcohol use identified as a key factor. The independent Data Monitoring and Ethic Committee (DMEC) deemed this to be unrelated to treatment. For the second death, it was not possible to establish a definite cause of death. However, a serious untoward incident review was conducted independently by the responsible NHS trust and concluded that there was no evidence of a relationship between the death and engagement with AVATAR therapy; therefore, it was determined to be unrelated to treatment or other trial procedures by the independent DMEC. No SAEs were related to trial procedures (treatment, device or assessment). Six events, involving five participants, were 'possibly related' to treatment. A 'possibly related' rating meant that the DMEC Chair did not determine that it was related but could not definitively rule out a relationship. The single 'possibly related' event for AV-BRF and four of the five for AV-EXT were hospital admissions (the other was a crisis team involvement). The main factor in the rating for each of these as 'possibly related' was the timing of the event being close to the AVATAR therapy course; however, in each case there were plausible unrelated contributory factors linked to admission identified by the

Table 1 | Demographic and clinical characteristics of the ITT population across trial arms at baseline

	Trial arm			
	TAU	**AV-BRF**	**AV-EXT**	**Total**
n	115 (33.3%)	116 (33.6%)	114 (33.0%)	345 (100.0%)
Age (years)	38.69 (±12.78)	39.35 (±13.31)	40.81 (±13.69)	39.61 (±13.26)
Age when they first started to hear voices (years)	24.39 (±11.99)	23.70 (±10.88)	25.93 (±12.10)	24.67 (±11.67)
Duration of contact with mental health services (years)	12.75 (±10.15)	13.24 (±11.61)	14.03 (±11.69)	13.33 (±11.15)
Gender				
Male	69 (60.0%)	72 (62.1%)	71 (62.3%)	212 (61.4%)
Female	44 (38.3%)	43 (37.1%)	42 (36.8%)	129 (37.4%)
Other	2 (1.7%)	1 (0.9%)	1 (0.9%)	4 (1.2%)
Ethnicity				
White	69 (60.0%)	63 (54.3%)	71 (62.3%)	203 (58.8%)
Black or mixed Black	19 (16.5%)	19 (16.4%)	19 (16.7%)	57 (16.5%)
South Asian or mixed South Asian	12 (10.4%)	9 (7.8%)	6 (5.3%)	27 (7.8%)
Other	15 (13.0%)	25 (21.6%)	18 (15.8%)	58 (16.8%)
Marital status				
Single	86 (74.8%)	87 (75.0%)	86 (75.4%)	259 (75.1%)
In a relationship	4 (3.5%)	11 (9.5%)	7 (6.1%)	22 (6.4%)
Cohabiting	10 (8.7%)	1 (0.9%)	3 (2.6%)	14 (4.1%)
Married or civil partnership	6 (5.2%)	8 (6.9%)	6 (5.3%)	20 (5.8%)
Divorced	9 (7.8%)	7 (6.0%)	12 (10.5%)	28 (8.1%)
Widowed	0 (0%)	2 (1.7%)	0 (0%)	2 (0.6%)
Living status				
Living alone (± children)	54 (47.0%)	54 (46.6%)	51 (44.7%)	159 (46.1%)
Living with husband/wife (± children)	5 (4.3%)	7 (6.0%)	5 (4.4%)	17 (4.9%)
Living together as a couple (± children)	10 (8.7%)	5 (4.3%)	6 (5.3%)	21 (6.1%)
Living with parents	31 (27.0%)	33 (28.4%)	31 (27.2%)	95 (27.5%)
Living with other relatives	4 (3.5%)	3 (2.6%)	6 (5.3%)	13 (3.8%)
Living with others	10 (8.7%)	12 (10.3%)	15 (13.2%)	37 (10.7%)
Not available or not applicable	1 (0.9%)	1 (0.9%)	0 (0%)	2 (0.6%)
Unknown	0 (0%)	1 (0.9%)	0 (0%)	1 (0.3%)
Highest level of schooling				
Primary school	1 (0.9%)	2 (1.7%)	2 (1.8%)	5 (1.4%)
Secondary, no exams qualifications	10 (8.7%)	9 (7.8%)	6 (5.3%)	25 (7.2%)
Secondary, O level or CSE equivalent	24 (20.9%)	20 (17.2%)	32 (28.1%)	76 (22.0%)
Secondary, A level equivalent	20 (17.4%)	14 (12.1%)	16 (14.0%)	50 (14.5%)
Vocational education or college	21 (18.3%)	35 (30.2%)	23 (20.2%)	79 (22.9%)
University degree or professional qualification	39 (33.9%)	34 (29.3%)	34 (29.8%)	107 (31.0%)
Not available or not applicable	0 (0%)	2 (1.7%)	1 (0.9%)	3 (0.9%)
Employment status				
Unemployed	85 (73.9%)	86 (74.1%)	86 (75.4%)	257 (74.5%)
Employed full-time	8 (7.0%)	12 (10.3%)	10 (8.8%)	30 (8.7%)
Employed part-time	8 (7.0%)	3 (2.6%)	5 (4.4%)	16 (4.6%)
Self-employed	0 (0%)	0 (0%)	2 (1.8%)	2 (0.6%)
Retired	1 (0.9%)	4 (3.4%)	2 (1.8%)	7 (2.0%)
Student	11 (9.6%)	9 (7.8%)	7 (6.1%)	27 (7.8%)
Housewife or husband	2 (1.7%)	2 (1.7%)	1 (0.9%)	5 (1.4%)
Not available or not applicable	0 (0%)	0 (0%)	1 (0.9%)	1 (0.3%)
Diagnoses (according to the ICD-10 codes)				
F20—Schizophrenia	54 (47.0%)	52 (44.8%)	45 (39.5%)	151 (43.8%)

Table 1 (continued) | Demographic and clinical characteristics of the ITT population across trial arms at baseline

	Trial arm			
	TAU	AV-BRF	AV-EXT	Total
F22—Persistent delusional disorders	2 (1.7%)	1 (0.9%)	0 (0%)	3 (0.9%)
F23—Acute and transient psychotic disorders	2 (1.7%)	0 (0%)	2 (1.8%)	4 (1.2%)
F24—Induced delusional disorder	0 (0%)	1 (0.9%)	0 (0%)	1 (0.3%)
F25—Schizoaffective disorders	6 (5.2%)	9 (7.8%)	12 (10.5%)	27 (7.8%)
F28—Other nonorganic psychotic disorders	4 (3.5%)	3 (2.6%)	1 (0.9%)	8 (2.3%)
F29—Unspecified nonorganic psychosis	31 (27.0%)	35 (30.2%)	41 (36.0%)	107 (31.0%)
F31—Bipolar affective disorder	3 (2.6%)	1 (0.9%)	4 (3.5%)	8 (2.3%)
F32.3—Severe depressive episode with psychotic symptoms	11 (9.6%)	14 (12.1%)	8 (7.0%)	33 (9.6%)
Not available or not applicable	2 (1.7%)	0 (0%)	1 (0.9%)	3 (0.9%)
Deprivation Index				
Most deprived	41 (35.7%)	54 (46.6%)	45 (39.5%)	140 (40.6%)
Second quintile	38 (33.0%)	33 (28.4%)	37 (32.5%)	108 (31.3%)
Third quintile	21 (18.3%)	13 (11.2%)	20 (17.5%)	54 (15.7%)
Fourth quintile	4 (3.5%)	6 (5.2%)	5 (4.4%)	15 (4.3%)
Least deprived	7 (6.1%)	8 (6.9%)	4 (3.5%)	19 (5.5%)
Unknown	4 (3.5%)	2 (1.7%)	3 (2.6%)	9 (2.6%)
PSYRATS-AH-Distress	15.70 (±2.78)	15.72 (±2.72)	15.89 (±2.77)	15.77 (±2.75)
PSYRATS-AH-Frequency	7.87 (±1.95)	7.39 (±2.11)	7.06 (±2.02)	7.44 (±2.05)
PSYRATS-AH-Total	30.64 (±4.42)	30.09 (±4.66)	30.11 (±4.42)	30.28 (±4.50)
Voice characterization				
More highly characterized (higher)	69 (60.0%)	71 (61.2%)	71 (62.3%)	211 (61.2%)
Less highly characterized (lower)	46 (40.0%)	45 (38.8%)	43 (37.7%)	134 (38.8%)

Values are presented as *n* (*n*%), indicating the frequency and its corresponding percentage of the total, while *n* (±*n*) indicates the mean value and s.d. ICD-10, International Statistical Classification of Diseases and Related Health Problems, 10th Revision.

clinical team and reviewed by the DMEC (for example, life stressors and substance use).

AV-EXT showed a higher number of SAEs compared to the two other arms. The category 'Other physical health event' contributed to this elevated number; it is unlikely to be treatment-specific in cause. Additional data tables are given in Supplementary Tables 5–8.

Moderation and compliance-adjusted analysis

We tested for moderation of treatment effects in a prespecified set of putative baseline moderators. There was no moderation according to low or high voice characterization for either comparison at either time point. The only moderation effects to meet the significance threshold related to comparisons for AV-EXT between Index of Multiple Deprivation quintiles (Q2 versus Q1 at 16 and 28 weeks; Q4 versus Q1 at 16 weeks) and an interaction for age at first hearing voices for AV-BRF, where an earlier onset of AHs may have been associated with larger treatment effects (Extended Data Tables 3 and 4). However, given the number of statistical tests, these findings may have occurred by chance. We estimated complier average causal effects (CACEs) using two definitions of treatment compliance and estimated larger CACE than intention-to-treat (ITT) effects for most comparisons. The overall pattern of findings is similar to that of the primary analysis, with larger between-group effects at 16 weeks and no between-group differences at 28 weeks (Extended Data Table 5).

Discussion

This multisite randomized trial of AVATAR therapy, investigated brief and extended forms, and tested delivery by a large cohort of therapists across geographically diverse sites. Voice distress mean scores and overall mean voice severity significantly improved in both AV-BRF and AV-EXT at end of treatment (16 weeks), compared to TAU alone. These improvements were maintained at 28 weeks, although they were no longer statistically significant. Voice frequency was reduced by AV-EXT (but not by AV-BRF) compared to TAU at the end of treatment (16 weeks); this improvement was sustained at the follow-up (28 weeks). The finding that AV-EXT demonstrated sustained reduction in the frequency of the occurrence of voices is relevant to research that shows that voice reduction (or cessation) is a clear priority for service users[14]. In summary, these findings meet the criteria prespecified in our statistical analysis plan for partial support of our main hypotheses, which stated that both versions of AVATAR therapy would be superior to TAU alone, at post-treatment and at follow-up, in reducing voice-related distress, voice severity and voice frequency.

The between-group effect sizes (that is, each version of therapy versus TAU) on voice-related distress post-treatment had a Cohen's $d = 0.58$ for AV-EXT and $d = 0.38$ for AV-BRF. These are greater than or equal to comparable post-treatment effect sizes of around 0.3–0.4 reported in recent meta-analyses for longer courses of cognitive behavioral therapy for psychosis (CBTp)[4,15], the current psychological treatment recommended by the National Institute for Clinical and Health Excellence (NICE)[16]. While AV-EXT treatment exceeded the threshold we prespecified for a clinically significant post-treatment change (that is, an effect size of 0.5 standard deviation), AV-BRF was slightly below this level with an associated *P* value just at the prespecified threshold for statistical significance ($P = 0.035$), suggesting some caution in its interpretation.

Secondary outcomes included recognized priorities for voice-hearers[14] (Supplementary Materials and Supplementary Table 9) and as

Table 2 | Treatment effect estimates on primary and secondary outcomes

Outcome	*n*	Time	Comparison	Effect	s.e.	*P*	96.5% CI		Effect size
Primary outcome									
Voice-related distress									
PSYRATS-AH-Distress	314	16	TAU versus AV-BRF	−1.05	0.500	**0.035**	−2.110	0	0.38
			TAU versus AV-EXT	−1.60	0.729	**0.029**	−3.133	−0.058	0.58
		28	TAU versus AV-BRF	−0.62	0.615	0.316	−1.912	0.679	0.22
			TAU versus AV-EXT	−1.06	0.779	0.175	−2.700	0.586	0.38
Key secondary outcomes									
Voice frequency									
PSYRATS-AH-Frequency	314	16	TAU versus AV-BRF	−0.50	0.244	0.042	−1.012	0.018	0.24
			TAU versus AV-EXT	−0.62	0.246	**0.011**	−1.140	−0.104	0.30
		28	TAU versus AV-BRF	−0.65	0.323	0.044	−1.331	0.030	0.32
			TAU versus AV-EXT	−0.89	0.300	**0.003**	−1.525	−0.258	0.43
Voice severity									
PSYRATS-AH Total	314	16	TAU versus AV-BRF	−2.04	0.853	**0.017**	−3.836	−0.239	0.45
			TAU versus AV-EXT	−2.32	0.894	**0.009**	−4.208	−0.438	0.52
		28	TAU versus AV-BRF	−1.61	1.256	0.199	−4.260	1.036	0.36
			TAU versus AV-EXT	−1.87	1.138	0.100	−4.274	0.526	0.42
Other secondary outcomes									
Other voice-specific measures									
BAVQ									
Omnipotence	298	16	TAU versus AV-BRF	0.64	0.455	0.162	−0.324	1.595	0.18
			TAU versus AV-EXT	0.73	0.652	0.266	−0.649	2.101	0.20
		28	TAU versus AV-BRF	0.83	0.399	0.038	−0.014	1.668	0.23
			TAU versus AV-EXT	1.29	0.589	**0.028**	0.053	2.536	0.36
Malevolence	298	16	TAU versus AV-BRF	0.44	0.469	0.352	−0.552	1.426	0.10
			TAU versus AV-EXT	0.38	0.571	0.507	−0.826	1.584	0.09
		28	TAU versus AV-BRF	−0.14	0.480	0.772	−1.151	0.873	0.03
			TAU versus AV-EXT	0.26	0.711	0.712	−1.236	1.760	0.06
Benevolence	298	16	TAU versus AV-BRF	0.11	0.367	0.767	−0.665	0.882	0.03
			TAU versus AV-EXT	0.52	0.417	0.212	−0.359	1.402	0.13
		28	TAU versus AV-BRF	0	0.435	0.993	−0.921	0.913	0
			TAU versus AV-EXT	0.67	0.505	0.186	−0.397	1.731	0.17
Total	298	16	TAU versus AV-BRF	3.05	1.548	0.049	−0.214	6.313	0.26
			TAU versus AV-EXT	3.00	1.248	**0.016**	0.372	5.634	0.26
		28	TAU versus AV-BRF	1.71	1.641	0.296	−1.747	5.175	0.15
			TAU versus AV-EXT	3.01	1.819	0.098	−0.824	6.846	0.26
VAAS									
Acceptance	297	16	TAU versus AV-BRF	3.41	0.815	**<0.001**	1.688	5.125	0.49
			TAU versus AV-EXT	3.88	1.048	**<0.001**	1.672	6.089	0.56
		28	TAU versus AV-BRF	2.84	0.813	**<0.001**	1.126	4.555	0.41
			TAU versus AV-EXT	3.98	1.196	**0.001**	1.458	6.500	0.58
Action	297	16	TAU versus AV-BRF	1.98	0.813	**0.015**	0.262	3.689	0.22
			TAU versus AV-EXT	3.26	0.974	**0.001**	1.205	5.313	0.36
		28	TAU versus AV-BRF	2.25	0.924	**0.015**	0.297	4.194	0.25
			TAU versus AV-EXT	2.83	1.042	**0.007**	0.635	5.027	0.31
Full-scale	297	16	TAU versus AV-BRF	5.51	1.406	**<0.001**	2.550	8.477	0.39
			TAU versus AV-EXT	7.17	1.776	**<0.001**	3.420	10.910	0.51
		28	TAU versus AV-BRF	5.12	1.462	**<0.001**	2.039	8.204	0.36
			TAU versus AV-EXT	6.83	1.989	**0.001**	2.634	11.022	0.48

Table 2 (continued) | Treatment effect estimates on primary and secondary outcomes

Outcome	*n*	Time	Comparison	Effect	s.e.	*P*	96.5% CI		Effect size
VPDS	298	16	TAU versus AV-BRF	−0.35	0.160	**0.030**	−0.686	−0.010	0.28
			TAU versus AV-EXT	−0.36	0.144	**0.013**	−0.659	−0.053	0.29
		28	TAU versus AV-BRF	−0.17	0.136	0.205	−0.460	0.115	0.14
			TAU versus AV-EXT	−0.31	0.137	**0.022**	−0.601	−0.026	0.25
Hallucinations Remission Score	314	16	TAU versus AV-BRF	−0.11	0.094	0.233	−0.311	0.086	0.22
			TAU versus AV-EXT	−0.21	0.108	0.050	−0.439	0.016	0.41
		28	TAU versus AV-BRF	−0.02	0.157	0.894	−0.351	0.310	0.04
			TAU versus AV-EXT	−0.06	0.127	0.654	−0.324	0.210	0.11
Distressing persecutory beliefs (delusions)									
PSYRATS-Delusions	279	16	TAU versus AV-BRF	−1.85	0.776	**0.017**	−3.490	−0.216	0.37
			TAU versus AV-EXT	−2.86	1.220	**0.019**	−5.433	−0.289	0.57
		28	TAU versus AV-BRF	−1.24	0.780	0.111	−2.890	0.401	0.25
			TAU versus AV-EXT	−3.41	1.448	**0.019**	−6.459	−0.355	0.68
Well-being and recovery									
WEMWBS	291	16	TAU versus AV-BRF	2.19	0.949	**0.021**	0.185	4.186	0.20
			TAU versus AV-EXT	6.83	2.028	**0.001**	2.555	11.107	0.63
		28	TAU versus AV-BRF	2.03	1.084	0.061	−0.251	4.321	0.19
			TAU versus AV-EXT	5.10	2.280	**0.025**	0.294	9.909	0.47
CHOICE	291	16	TAU versus AV-BRF	7.39	2.529	**0.003**	2.061	12.723	0.33
			TAU versus AV-EXT	9.44	2.260	**<0.001**	4.673	14.204	0.42
		28	TAU versus AV-BRF	5.67	2.404	**0.018**	0.598	10.737	0.25
			TAU versus AV-EXT	6.40	2.473	**0.010**	1.181	11.610	0.29
Mood, anxiety, trauma									
DASS									
Anxiety	293	16	TAU versus AV-BRF	−4.50	0.953	**<0.001**	−6.512	−2.495	0.44
			TAU versus AV-EXT	−4.07	0.856	**<0.001**	−5.879	−2.269	0.40
		28	TAU versus AV-BRF	−3.26	0.936	**<0.001**	−5.232	−1.287	0.32
			TAU versus AV-EXT	−2.07	1.268	0.103	−4.740	0.605	0.20
Depression	293	16	TAU versus AV-BRF	−3.76	1.111	**0.001**	−6.102	−1.416	0.33
			TAU versus AV-EXT	−4.19	1.252	**0.001**	−6.829	−1.548	0.37
		28	TAU versus AV-BRF	−2.66	1.168	**0.023**	−5.122	−0.196	0.24
			TAU versus AV-EXT	−1.90	1.505	0.207	−5.073	1.274	0.17
Stress	293	16	TAU versus AV-BRF	−3.97	1.004	**<0.001**	−6.083	−1.848	0.39
			TAU versus AV-EXT	−3.57	1.239	**0.004**	−6.187	−0.962	0.35
		28	TAU versus AV-BRF	−3.57	0.965	**<0.001**	−5.606	−1.536	0.35
			TAU versus AV-EXT	−1.75	1.360	0.198	−4.616	1.118	0.17
BDI	291	16	TAU versus AV-BRF	−2.06	1.517	0.174	−5.261	1.137	0.14
			TAU versus AV-EXT	−3.52	1.665	**0.035**	−7.026	−0.005	0.24
		28	TAU versus AV-BRF	−1.73	1.547	0.264	−4.991	1.534	0.12
			TAU versus AV-EXT	−1.72	2.160	0.425	−6.279	2.831	0.12
ITQ									
DSO	257	16	TAU versus AV-BRF	−1.55	0.769	0.043	−3.177	0.068	0.25
			TAU versus AV-EXT	−1.37	0.949	0.149	−3.369	0.631	0.22
PTSD	259	16	TAU versus AV-BRF	−0.82	1.050	0.433	−3.038	1.391	0.12
			TAU versus AV-EXT	−0.73	0.844	0.387	−2.512	1.049	0.10
Anxiety (ESM)	139	16	TAU versus AV-BRF	−1.07	0.689	0.122	−2.520	0.387	0.25
			TAU versus AV-EXT	−2.10	0.731	**0.004**	−3.645	−0.563	0.50
		28	TAU versus AV-BRF	0.09	0.836	0.913	−1.671	1.855	0.02
			TAU versus AV-EXT	−0.29	0.861	0.734	−2.109	1.523	0.07

Effects are estimates of between-group mean difference after adjusting for site, voice characterization and baseline measurement of each outcome. DSO, disturbances in self-organization; PTSD, posttraumatic stress disorder; VAAS, Voices Acceptance and Action Scale. Time points: 16, 16 weeks after baseline (post-treatment follow-up); 28, 28 week follow-up after baseline. *n* represents the sample size for the longitudinal mixed model for each measure in the analysis. Bold denotes $P \leq 0.035$.

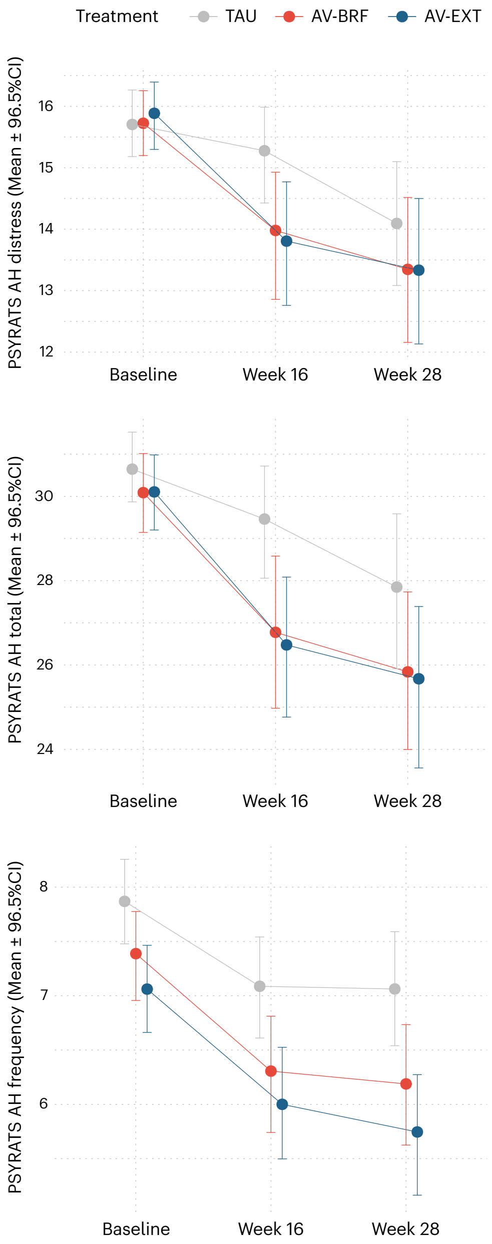


Fig. 2 | PSYRATS-AH-Distress, Total Severity and Frequency observed mean scores with 96.5% CIs. Week 16: 16 weeks after baseline (post-treatment follow-up). Week 28: 28-week follow-up after baseline. The center points represent the mean values with the 96.5% CIs. The sample size (*n*) for PSYRATS-AH-Distress, Total Severity and Frequency was 345 at baseline, 299 at week 16 and 298 at week 28.

such are worthy of some consideration. While the trial was not designed for a direct comparison between AV-BRF and AV-EXT, the different secondary outcome effects of each, compared to TAU alone, are also informative. AV-BRF delivered benefits at both time points (albeit with relatively small between-group effects) on important secondary outcomes, including personal recovery, voice acceptance and action and mood (anxiety, depression and stress), with the latter being the single domain where sustained effects were observed for AV-BRF but not AV-EXT. Engagement in AV-BRF was also strong, with treatment completion rates of over 80%. AV-EXT delivered benefits at both time points across well-being, personal recovery and empowerment, outcomes that were not demonstrated in earlier AVATAR studies[5,8]. There were also significant post-treatment between-group effects for AV-EXT on depression, stress and anxiety, with convergent evidence from ESM data for anxiety reduction in daily life. However, there were no significant differences for AV-EXT in mood outcomes at 28 weeks. With respect to the hypothesized effects on beliefs about voices, improvements were observed for AV-EXT in omnipotence but not malevolence. Finally, evidence of improvements in persecutory distressing beliefs linked to the voice showed a moderate-to-large effect size at 28 weeks ($d = 0.68$), which is approximately double the typical findings reported for delusions in meta-analyses of CBTp[4,15]. This evidence supports AV-EXT delivering important and sustained changes in the personal understanding of the voice, which were not observed in AV-BRF.

AVATAR therapy as delivered in this trial was safe. There were no SAEs rated by the independent DMEC as related to treatment or the medical device. There was a larger number of overall SAEs and adverse events (AEs) reported in AV-EXT, across a diverse range of categories; this is probably at least partially attributable to the opportunity for increased monitoring and reporting provided by therapists over more sessions in AV-EXT. In addition, the trauma focus within AV-EXT is a possible factor in the higher recording of affective changes (as AEs) that did not meet the threshold for SAEs (that is, transient increases in distress or voice-hearing that resolved over time). The findings from this multisite trial allied to those reported in the AVATAR1 trial provide evidence supporting the safety of AVATAR therapy.

Despite the range of significant sustained secondary outcomes, the lack of a significant effect on the primary outcome at 28 weeks in AV-EXT is to be acknowledged and considered. Treatment completion of just under 60% for AV-EXT (compared to just over 80% for AV-BRF) may be plausibly linked to the increased direct trauma-focused work within sessions and suggests the need for improved treatment engagement, adherence and, consequently, efficacy, based on learning from the current trial. Consistent with this, the compliance-adjusted analysis showed larger treatment effect estimates than the ITT findings in the subgroup of participants who complied with their allocation by fully completing treatment. Improved engagement may be delivered through use of a collaborative review around an optimal session number to ensure that the person retains a strong sense of control, particularly around trauma-focused work. Planned qualitative analysis, which will explore the experience of direct dialoguing with voice content and includes individuals who did not complete treatment, will inform this future optimization of treatment engagement. Another key challenge is how to optimize and sustain the real-world impact from the (often) powerful change in distress observed within dialogues. The current method (provision of complete dialogue recordings to listen to at home) was subject to variable engagement and could be enhanced. Work is soon to commence on innovation in artificial intelligence (AI)-powered virtual conversational agents capable of delivering avatar dialogues (Wellcome ref. no. 227721/Z/23/Z). In addition to boosting future scalability, AI integration with mobile technology would transform between-session practice, potentially boosting long-term efficacy. There is also interest in the use of immersive virtual reality to enhance AVATAR therapy delivery and effects. The evidence is currently limited, but the results of AVATAR VRSocial in Germany (ISRCTN35980117) and independent trials in Denmark and Canada will be informative[17,18] Finally, we plan to examine hypotheses concerning the mechanisms of change in a future analysis, which will guide further refinement of the approach. Candidate mechanisms include reduced anxiety and sense of threat, and increased empowerment and voice acceptance.

Although previous work found that duration of AVATAR voice dialogues and everyday behavioral engagement with voices were related to more complex characterization[19], the moderation analysis did not support our hypothesis that greater baseline complexity of voice characterization would moderate treatment effects. Furthermore, there were very few demographic variables that moderated treatment effects; given the number of tests, we caution that these findings may have occurred by chance. Overall, the results suggest no robust

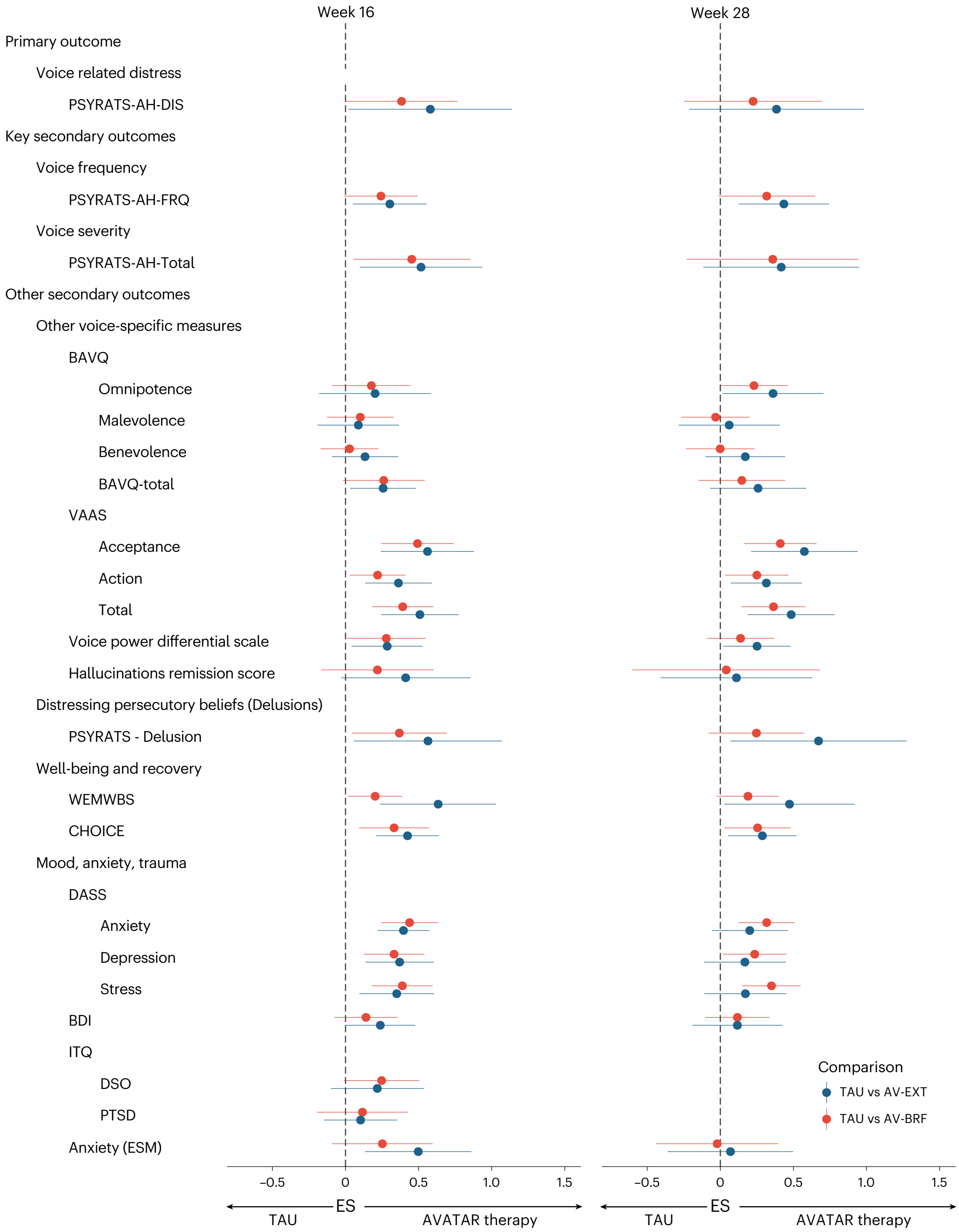


Fig. 3 | Effect size estimates with 96.5% CIs for primary and secondary outcomes at 16 and 28 weeks. Week 16: 16 weeks after baseline (post-treatment follow-up). Week 28: follow-up 28 weeks after baseline. The effect sizes (center points) with 96.5% CIs (error bars) for each outcome are shown. Effect sizes were calculated by dividing the estimated treatment effects from the mixed model and its 96.5% CI by the baseline s.d. of that outcome. The sample size (*n*) for each mixed-model outcome is provided in Table 2.

evidence of differential effectiveness for either AV-BRF or AV-EXT across clinical or demographic variables.

The design of this trial had some limitations. First, the use of a TAU control meant that we could not determine the benefits of AVATAR therapy compared to another psychological treatment. The current frontline psychological therapy (CBTp), recommended by NICE as a minimum of 16 sessions, is notably longer in duration than even AV-EXT. In our previous trial, we adapted a form of brief supportive counseling as a control of comparable duration, but this is not routinely available and was outperformed by AVATAR therapy; therefore, a TAU comparison was used for this larger, pragmatic, multisite study[20]. The ICC for the therapist clustering effect indicated that around 5% of the residual variance in primary outcome was at the therapist level. Based on our previous trial, which showed a smaller therapist ICC, we did not account for therapist clustering effects in the sample size calculation; however, this was considered within the analysis models. Finally, the trial was not fully powered for a comparison of AV-BRF with AV-EXT because the sample size required was impracticable. Health economic

Table 3 | Serious adverse events according to treatment type across the trial arms

	TAU P, E (%) 14,14 (24.1%)	AV-BRF P, E (%) 14, 14 (24.1%)	AV-EXT P, E (%) 28, 30 (51.7%)
Serious adverse events			
Distress associated with completion of assessment measures	–	–	–
Significant distress during AVATAR therapy	–	–	–
Admission to hospital for psychological health event	8, 8 (57.1%)	8, 8 (57.1%)	8, 9 (30.0%)
Admission to hospital for physical health event	1, 1 (7.1%)	2, 2 (14.3%)	5, 5 (16.7%)
Referral to crisis team	2, 2 (14.3%)	1, 1 (7.1%)	6, 6 (20.0%)
Violent incident necessitating police involvement (victim)	–	–	–
Violent incident necessitating police involvement (accused)	1, 1 (7.1%)	–	–
Deliberate self-harm	–	2, 2 (14.3%)	3, 4 (13.3%)
Other psychological health event	1, 1 (7.1%)	–	3, 3 (10.0%)
Other physical health event	1, 1 (7.1%)	1, 1 (7.1%)	1, 1 (3.3%)
Death	–	–	2, 2 (6.7%)

P, E (%) represents participant and events (% of events).

analysis, to be reported separately, will offer relevant information on the cost-effectiveness of both versions.

AVATAR therapy is one of several evidence-based digital health interventions emerging for psychosis and schizophrenia[21–24]. AVATAR therapy offers the experience of a powerful digital 'sense of presence' of a distressing voice, shared with the therapist and enabling rapid change and reduced frequency[12]. The delivery of the trial, across a geographically and demographically diverse sample and including therapists from a range of disciplines in routine clinical settings, strengthens the real-world relevance of these findings. In the context of adaptation to the coronavirus disease 2019 (COVID-19) pandemic, remote delivery of AVATAR therapy has been shown to be feasible and acceptable, which is promising for future scalability (see also the AMETHYST trial; ClinicalTrials.gov registration: NCT05982158). A recently published NICE-Early Value Assessment of digital therapies for psychosis recommended AVATAR therapy for NHS deployment while further real-world evidence is generated[25]. The data reported in this article on efficacy and safety provide evidence to inform this ongoing evaluation. Forthcoming trial outputs (to be reported separately) will provide cost-effectiveness analysis and include qualitative studies of diverse patient and clinician perspectives on AVATAR therapy. Building on the AVATAR2 data, real-world evidence of clinical and cost-effectiveness, safety and acceptability of AVATAR therapy when implemented in routine care is now required to support a full NICE submission and facilitate widespread NHS adoption.

In conclusion, this study has provided partial support for the primary hypothesis, in that there were superior effects on the primary outcome of voice-related distress of AVATAR therapy over TAU alone at 16 weeks, in both AV-BRF and AV-EXT versions, and that AV-EXT met our threshold for clinically significant change; however, the effects were no longer statistically significant at 28 weeks. In addition, we have provided indications of a wider range of sustained improvements in outcomes prioritized by voice-hearers, of the longer formulation-based AV-EXT version, which connects dialogues to the person's life history[7]. Treatment completion of AV-BRF was high, while comparable rates for AV-EXT suggested the need for refinement to improve engagement. Based on these trial findings, we recommend that future development and provision of AVATAR therapy is primarily guided by the AV-EXT protocol.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03252-8.

References

1. Hollis, C. et al. Identifying research priorities for digital technology in mental health care: results of the James Lind Alliance Priority Setting Partnership. *Lancet Psychiatry* **5**, 845–854 (2018).
2. Bucci, S., Schwannauer, M. & Berry, N. The digital revolution and its impact on mental health care. *Psychol. Psychother.* **92**, 277–297 (2019).
3. Leucht, S. et al. Sixty years of placebo-controlled antipsychotic drug trials in acute schizophrenia: systematic review, Bayesian meta-analysis, and meta-regression of efficacy predictors. *Am. J. Psychiatry* **174**, 927–942 (2017).
4. Turner, D. T. et al. Efficacy and moderators of cognitive behavioural therapy for psychosis versus other psychological interventions: an individual-participant data meta-analysis. *Front. Psychiatry* **11**, 402 (2020).
5. Leff, J., Williams, G., Huckvale, M., Arbuthnot, M. & Leff, A. P. Avatar therapy for persecutory auditory hallucinations: what is it and how does it work? *Psychosis* **6**, 166–176 (2014).
6. Hayward, M., Jones, A.-M., Bogen-Johnston, L., Thomas, N. & Strauss, C. Relating therapy for distressing auditory hallucinations: a pilot randomized controlled trial. *Schizophr. Res.* **183**, 137–142 (2017).
7. Longden, E. et al. A psychological intervention for engaging dialogically with auditory hallucinations (Talking With Voices): a single-site, randomised controlled feasibility trial. *Schizophr. Res.* **250**, 172–179 (2022).
8. Craig, T. K. et al. AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. *Lancet Psychiatry* **5**, 31–40 (2018).
9. du Sert, O. P. et al. Virtual reality therapy for refractory auditory verbal hallucinations in schizophrenia: a pilot clinical trial. *Schizophr. Res.* **197**, 176–181 (2018).
10. Ward, T. et al. AVATAR therapy for distressing voices: a comprehensive account of therapeutic targets. *Schizophr. Bull.* **46**, 1038–1044 (2020).
11. Garety, P. et al. Optimising AVATAR therapy for people who hear distressing voices: study protocol for the AVATAR2 multi-centre randomised controlled trial. *Trials* **22**, 366 (2021).
12. Rus-Calafell, M. et al. The role of sense of voice presence and anxiety reduction in AVATAR therapy. *J. Clin. Med.* **9**, 2748 (2020).
13. Woodward, T. S. et al. Symptom dimensions of the psychotic symptom rating scales in psychosis: a multisite study. *Schizophr. Bull.* **40**, S265–S274 (2014).
14. Longden, E., Branitsky, A., Sheaves, B., Chauhan, N. & Morrison, A. P. Preferred treatment outcomes in psychological therapy for voices: a comparison of staff and service-user perspectives. *Psychosis* **16**, 107–117 (2024).
15. Sitko, K., Bewick, B. M., Owens, D. & Masterson, C. Meta-analysis and meta-regression of cognitive behavioral therapy for psychosis (CBTp) across time: the effectiveness of CBTp has improved for delusions. *Schizophr. Bull. Open* **1**, sgaa023 (2020).

16. National Institute for Health and Care Excellence (NICE). *Psychosis and Schizophrenia in Adults: Prevention and Management. Clinical Guideline [CG178]* https://www.nice.org.uk/guidance/cg178 (2014).
17. Smith, L. C. et al. The CHALLENGE trial: the effects of a virtual reality-assisted exposure therapy for persistent auditory hallucinations versus supportive counselling in people with psychosis: study protocol for a randomised clinical trial. *Trials* **23**, 773 (2022).
18. Dellazizzo, L., Potvin, S., Phraxayavong, K. & Dumais, A. One-year randomized trial comparing virtual reality-assisted therapy to cognitive–behavioral therapy for patients with treatment-resistant schizophrenia. *NPJ Schizophr.* **7**, 9 (2021).
19. Ward, T. et al. The role of characterisation in everyday voice engagement and AVATAR therapy dialogue. *Psychol. Med.* **52**, 1–8 (2021).
20. Gold, S. M. et al. Control conditions for randomised trials of behavioural interventions in psychiatry: a decision framework. *Lancet Psychiatry* **4**, 725–732 (2017).
21. Torous, J. et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* **20**, 318–335 (2021).
22. Garety, P. et al. Effects of SlowMo, a blended digital therapy targeting reasoning, on paranoia among people with psychosis: a randomized clinical trial. *JAMA Psychiatry* **78**, 714–725 (2021).
23. Gumley, A. I. et al. The EMPOWER blended digital intervention for relapse prevention in schizophrenia: a feasibility cluster randomised controlled trial in Scotland and Australia. *Lancet Psychiatry* **9**, 477–486 (2022).
24. Cella, M. et al. Virtual reality therapy for the negative symptoms of schizophrenia (V-NeST): a pilot randomised feasibility trial. *Schizophr. Res.* **248**, 50–57 (2022).
25. National Institute for Health and Care Excellence (NICE). *Digital Health Technologies to Help Manage Symptoms of Psychosis and Prevent Relapse in Adults and Young People: Early Value Assessment* https://www.nice.org.uk/guidance/hte17 (2024).

Philippa A. Garety[1,2] ✉, **Clementine J. Edwards**[1,2], **Hassan Jafari**[3], **Richard Emsley**[3], **Mark Huckvale**[4], **Mar Rus-Calafell**[1,5], **Miriam Fornells-Ambrojo**[6,7], **Andrew Gumley**[8,9], **Gillian Haddock**[10,11], **Sandra Bucci**[10,11], **Hamish J. McLeod**[8,9], **Jeffrey McDonnell**[6,7], **Moya Clancy**[8,9], **Michael Fitzsimmons**[10,11], **Hannah Ball**[10,11], **Alice Montague**[6,7], **Nikos Xanidis**[8,9], **Amy Hardy**[1,2], **Thomas K. J. Craig**[2,12,13] & **Thomas Ward**[1,2,13] ✉

[1]Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. [2]South London & Maudsley NHS Foundation Trust, London, UK. [3]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [4]University College London, London, UK. [5]Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Bochum, Germany. [6]Research Department of Clinical, Educational and Health Psychology, University College London, London, UK. [7]North East London NHS Foundation Trust, London, UK. [8]School of Health and Wellbeing, University of Glasgow, Glasgow, UK. [9]NHS Greater Glasgow & Clyde, Glasgow, UK. [10]Division of Psychology and Mental Health, School of Health Sciences, University of Manchester and the Manchester Academic Health Sciences Centre, Manchester, UK. [11]Greater Manchester Mental Health NHS Foundation Trust and the Manchester Academic Health Sciences Centre, Manchester, UK. [12]Department of Health Service and Population Research, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. [13]These authors contributed equally: Thomas K. J. Craig, Thomas Ward. ✉e-mail: Philippa.Garety@kcl.ac.uk; Thomas.Ward@kcl.ac.uk

Methods

Study design and oversight

This multisite, parallel-group, assessor-blinded, randomized controlled trial assessed the efficacy and safety of two forms of AVATAR therapy, AV-BRF (six sessions with a standardized focus on exposure, assertiveness and self-esteem) plus TAU or AV-EXT (12 sessions, with an initial phase mirroring AV-BRF followed by a personalized, developmentally focused second phase) plus TAU compared to TAU alone on reducing voice-related distress (primary outcome), voice-related frequency and severity (key secondary outcomes), and other mood, well-being and voice-related outcomes. The study received ethical approval (Camberwell St. Giles Research Ethics Committee: no. 20/LO/0657; Integrated Research Application System no. 277118) and was prospectively registered with the ISRCTN registry at which the published trial protocol[11] and statistical analysis plan can also be accessed (ISRCTN55682735). The trial complied with the International Conference on Harmonization Good Clinical Practice guidelines and the 2013 Declaration of Helsinki. The study was overseen by an independent trial steering committee and a separate independent data monitoring and ethics committee. All participants provided written informed consent.

Trial protocol deviations. Before participant recruitment commenced, the COVID-19 pandemic began. Face-to-face contact was restricted intermittently. The study start was delayed for 3 months, as mandated nationally, and the protocol and procedures were adapted to allow for remote delivery of the trial. The final trial protocol (v.1.2) was approved before participant recruitment commenced and no changes were made to the protocol for the conduct of the trial subsequent to the start of the trial.

Participants

Between 1 January 2021 and 30 November 2022, we assessed 642 people for eligibility, recruiting 345 participants. Participants were randomized at four study sites, each recruiting from two mental health service providers, in the United Kingdom (three in England: South London, North London, Manchester; one in Scotland: Glasgow) and were randomly allocated to three parallel arms: 116 to AV-BRF, 114 to AV-EXT and 115 to TAU control (Fig. 1).

Participants were referred by a clinician at the participating clinical sites. Other routes to participation included contact through institutional research registers or self-referral.

The inclusion criteria were as follows: (1) aged 18 years or over; (2) currently under the care of a specialist mental health team; (3) current frequent and distressing voices (as measured by a score of at least one on each of the intensity of distress and frequency items of the PSYRATS-AH (Voices) Scale[26]), persisting for at least 6 months and spoken in English; (4) speak and read English to a sufficient level to provide consent and complete the assessment procedures; (5) a clinical diagnosis of schizophrenia spectrum disorder (ICD-10 F20–29) or affective disorder with psychotic symptoms (ICD-10 F30–39, subcategories with psychotic symptoms) as determined through clinical records and additional consultation with the clinical team, if required. Criteria for exclusion included: (1) primary diagnosis of substance disorder, personality disorder or learning disability; (2) lacking capacity to consent; (3) profound visual or hearing impairment or insufficient comprehension of English to be able to engage in assessment or treatment; (4) currently undertaking individual psychological treatment for voices; (5) currently experiencing an acute mental health crisis.

Randomization and masking

After baseline assessment, we randomly assigned (1:1:1) eligible participants via a secure independent web-based service hosted by the King's Clinical Trials Unit, using randomly varying sized blocks (three and six), stratified according to site and baseline voice characterization (more or less) as defined by meeting the threshold for more highly characterized voices (score > 7) on the Voice Characterisation Checklist[27].

Research assessors were masked to allocation and procedures were followed to maintain their masking (assessors did not have access to clinical records after the baseline (pre-randomization) assessment or access to the treatment database at any stage); all assessments were done at sites remote from the clinic and participants were reminded before each assessment not to disclose their allocation. It is not possible to mask psychological treatment participants or therapists to their allocation; site coordinators were unmasked and informed participants. Therapists were allocated at each site based on availability. Breaks in assessor masking were recorded; if unmasking occurred, reallocation to another rater occurred. All primary and key secondary outcomes (PSYRATS-AH Scale) were assessed by blinded assessors. Unmasking occurred in 29 assessments (8.4%) at 16 weeks and 15 assessments (4.3%) at 28 weeks. All assessments of these individuals were scored by blinded assessors after these instances of unmasking.

Procedures

The intervention. AVATAR therapy is a digital treatment in which the person engages in face-to-face dialogues with a personalized digital embodiment of the voice ('the avatar'). The avatar is presented to the person on a two-dimensional computer screen.

In the AVATAR2 trial, AVATAR therapy was delivered in two versions, according to the randomized condition, that is, AV-BRF and AV-EXT, according to a comprehensive clinical manual.

Therapy structure. Both versions commence with an initial clinical assessment session, which includes creation of the avatar. Approximately 20 min are dedicated to making the avatar of the person's main distressing voice. This is to create a tangible representation of the voice, with a face, to whom the person can directly address their resistance. The aim is to create a voice and an image that is a 'good enough' representation of the voice for the person; the created avatar tends to achieve a surprisingly good match. However, in practice, there is a balance between creating a workable representation of the voice while ensuring that the person does not feel overburdened or pressured to achieve 'a perfect match'—in this context, as a rule of thumb, 70% is considered a good match. Where possible the avatar should represent the dominant persecutory voice as identified by the person. While some may experience a rotating gallery of characters, the guiding principle is to create an avatar that best represents the group of voices and recommend that the person tries out what works with the avatar for other distressing voice(s) they experience.

AV-BRF consists of six individual, face-to-face sessions delivered by trained therapists using the AVATAR therapy software. AV-BRF is designed to include the core aspects of AVATAR therapy, specifically the use of the avatar to deliver a realistic enactment of the voice (including exposure to verbatim voice content) and a treatment focus on increasing power and control and self-esteem. AV-EXT consists of 12 individual, face-to-face sessions and consists of two phases. The first phase mirrors AV-BRF. The aim of phase 2 is to develop an understanding of the voice(s) within the broader context of the person's life and relationship history, informing a series of dialogues that flexibly target a wider range of treatment targets[10] (Extended Data Fig. 1). For both versions, sessions could be increased or reduced by a maximum of two for treatment completion, guided by the clinical judgment of the therapist and in collaboration with the participant.

Each session (in both versions) consists of three parts: (1) pre-dialogue discussion; (2) active avatar dialogue; and (3) post-dialogue debrief. The whole takes 45–60 min. Pre-dialogues involve a review of the previous week (changes in the voice and other progress), identification of the main themes to be tackled in the forthcoming dialogue and, as necessary, practice role-play focused on the anticipated challenges within the dialogue. For AV-EXT, the pre-dialogue (particularly from the mid-treatment review onward) is also used to explore and formulate the possible contribution of previous traumatic experience to the voice-hearing experience, including instances of abuse, bullying, racial and sexual discrimination, or other forms of social exclusion and marginalization. During active dialogues, the therapist and the voice-hearer sit in separate rooms, communicating digitally, with the therapist remotely viewing the participant using a webcam. The post-dialogue session discusses the dialogue experience, commenting on the strengths shown by the person, discussing emerging content and finally giving a recording of the dialogue session and encouragement for the week ahead.

Therapy delivery. Therapy was intended to be delivered in person, at a participant's local mental health clinic. However, because of the COVID-19 pandemic, the software was adapted to support remote treatment delivery using video conferencing software. This allowed participants to have treatment sessions from home, joining their therapist via remote web link. Eighty-seven percent of participants (n = 200) had treatment in person at the clinic (99 of 114 (87%) for AV-EXT and 101 of 116 (87%) for AV-BRF). Further data on face-to-face and remote delivery can be found in Supplementary Tables 3 and 4.

Therapist training. Of the 19 therapists who participated in the trial, 12 were qualified clinical or counseling psychologists, five were psychiatrists (three specialist trainee and two consultants) and two were nurse therapists; 11 were female and eight were male. The mean years of experience in delivering psychological treatment before commencing AVATAR therapy was 11.6 years (s.d. = 10.3, range = 1–40); 18 of 19 had more than 6 months' experience of psychosis intervention at the start of their involvement in the trial. Two of the therapists were expert AVATAR therapists and trainers and delivered treatment in the previous AVATAR1 trial[8]. All other therapists were trained for the study. Training involved a combination of direct teaching and self-directed learning (including access to live treatment reference material), followed by closely supervised training cases. After the training period, the treatment supervision model included 1:1 (typically weekly) and group-based peer supervision (typically monthly). Sharing of live audio was a crucial aspect of supervision to inform discussions around the key treatment processes to be targeted (both in terms of enacting the avatar and suggested consolidation work before and after dialogue).

Adherence, fidelity and competence. Treatment adherence was assessed by the number of sessions attended. Fidelity to the clinical manual was assessed by the therapist completing a session-by-session checklist. An a priori checklist of therapist fidelity to protocolized components of treatment was developed based on earlier AVATAR clinical trials with specific additions for AV-EXT. Fidelity was predefined as completion of 80% of the specified components for each session. For both AV-BRF and AV-EXT, the mean self-reported fidelity for each session was more than 90%, with an overall mean rating (across all sessions) for AV-BRF of 92.46 (s.d. = 9.57; minimum = 19.64; maximum = 100) and for AV-EXT a mean of 93.38 (s.d. = 8.61; minimum = 17.31; maximum = 100) for AV-EXT.

Therapist competence was assessed by an expert in AVATAR therapy for both general and clinical and AVATAR-specific skills. Each newly trained trial therapist was rated for competence based on the review of early, mid and late session treatment delivery for at least one completed intervention. Ratings were conducted for two cases for therapists who delivered completed treatment with more than five participants. Cases were selected at random for each therapist, but excluding any cases where audio recordings were not available (for example, because of technical issues or the participant not consenting to a full recording). The rating tool was adapted from AVATAR1 to allow for different skill requirements for each level of treatment. For AVATAR-BRF, it included five items for session one and six (each) for mid and later sessions (17 items). The AVATAR-EXT rating tool mirrored this with the key difference being one additional item rated at the mid and last sessions to capture 'promoting an understanding of voice within broader autobiographical and person-specific context' (total items = 19). Each item was rated 1–5 with a total possible score across the three sessions of 85 for AVATAR-BRF (17 items) and 95 for AVATAR-EXT (19 items), with a benchmark of 3/5 per item for competent delivery (or 60% for the total score across all items). The mean competence rating for AV-BRF (n = 10 cases) was 79.8% (s.d. = 13.5); for AV-EXT (n = 13 cases), it was 76.8% (s.d. = 13.5).

AVATAR hardware and software. The Avatar Therapy System facilitates the delivery of AVATAR therapy for voice-hearing through a mix of commodity computer hardware and custom software. The software supports both enrollment of an avatar for the voice-hearer and real-time communication between the therapist and the voice-hearer using the avatar as a third party in a treatment session. The computing platform consists of two Windows laptops (or a laptop and desktop and a tablet) connected over a network. These can either be located within two rooms in the clinic (local delivery), or can be located at the therapist's office and the client's home (remote delivery). The key technical elements of the software include voice enrollment, face enrollment, real-time voice conversion, real-time lip synchronization and real-time animation.

Voice enrollment is the process by which the client chooses a voice for the avatar. The therapist makes a recording of the client's normal voice and the software manipulates that voice along dimensions of pitch, vocal tract size, spectral tilt and temporal roughness. Slider controls on the interface allow the client to hear many different variations of the therapist's voice until a good match to the 'voice' is found (Extended Data Fig. 2). These control settings are chosen and saved.

Face enrollment is the process by which the client chooses a face for the avatar. The underlying technology for creating and modifying faces is called FaceGen and is licensed from Singular Inversions. In face enrollment, a set of faces that match the basic attributes of the heard voice is generated and the closest one is chosen by the client. The software then allows the manipulation of facial shape, color and texture, as well as the addition of hair. The software supports different ethnicities and some nonhuman characters, such as a devil, witch and robot.

Real-time voice conversion is a technology for converting the therapist's voice to the avatar's voice within a live treatment session. The stored voice transformation settings chosen during voice enrollment are applied to the therapist's voice recorded from a headset microphone; the converted voice is then communicated to the client's computer over the network.

Real-time lip synchronization is the process by which representative mouth shapes and jaw positions of the avatar are chosen from an acoustic analysis of the speech signal being produced by the therapist when speaking as the avatar. This mapping between acoustic signal and visemes is performed by a neural network classifier.

Real-time animation is the process by which the three-dimensional model of the avatar is animated during the treatment session to make it look like the avatar is engaging in a dialogue. This is achieved by morphing the three-dimensional graphical model of the avatar according to the viseme output of the lip synchronization component while it is speaking. In addition, the avatar looks around and blinks occasionally so that it looks more alive.

The Avatar Therapy System also acts as a database of therapists, clients, avatars and sessions. It keeps recordings of treatment sessions, which can be shared with clients. Facilities also exist for backup and synchronization between a group of laptops at one site. The Avatar Therapy System has been registered as a class 1 medical device by Avatar Therapy Ltd.

Patient and public involvement

Patient and public involvement (PPI) in which experts by experience supported the study, had a major role at all stages of the AVATAR2 trial, including design, recruitment of staff and participants, analysis and dissemination through supporting the development of accessible plain English summaries and visual representations. An active and creative group of people was established, including members from different backgrounds, with lived experience of mental health conditions and recovery, including carers. In total, the AVATAR2 PPI group included over 20 members across the four sites, with at least four PPI consultants at each site. The local groups met approximately every 2 months for the duration of the trial, with specific activities planned between meetings. There was also coordination of PPI input from sites to the AVATAR2 whole-team events, which took place approximately every 6 months. Group members were reimbursed for their time at a rate of GBP20 per hour; travel expenses were covered for attendance at meetings. Individual members contributed to a wide range of activities: planning events, supporting recruitment, reviewing documents, joining interview and recruitment panels, training research assistants, feeding back on content for the website, reviewing the results and their importance and interpretation, involvement in other public-facing work and attending wider team meetings. Each member was buddied with a named research worker who facilitated flexible and tailored involvement. Personal development plans were a helpful tool to support learning and development within the role. In keeping with principles of open and collaborative involvement, the activities of the PPI group extended beyond those specified within the trial protocol. For example, a creative space was identified as important during PPI meetings and a creative writing workshop emerged organically over time. While independent from the core deliverables of the trial, this regular creative workshop became highly valued and impactful across all aspects of the project (further details of this work will be the focus of a separate publication).

A formal facilitated series of meetings was held with our PPI consultants concerning the outcomes they considered important and reflections on the results. The outcomes considered important are shown in Supplementary Table 6. These are set alongside the most relevant trial outcome measures and whether significant effects were found.

Concomitant care

Throughout the post-randomization period, participants in all three arms continued with their usual care (TAU). TAU was delivered according to UK national and local service guidelines, typically involving antipsychotic medication, contact with a mental health worker and outpatient psychiatric appointments. Participation did not alter pharmacological or psychosocial treatment decisions. As expected, the TAU-alone arm showed higher levels of other psychological interventions (further data are provided in Supplementary Table 10); 326 (94%) participants were prescribed antipsychotic medication of whom a quarter were prescribed clozapine. Dosages were converted to chlorpromazine equivalents and were broadly comparable between the three arms of the study (Supplementary Table 11).

Assessment procedures

Assessments were conducted at 0 weeks (baseline) before randomization, 16 weeks (after baseline; follow-up after treatment) and 28 weeks (after baseline; follow-up after 28 weeks). Blinded assessors conducted recruitment and consent procedures and assessments remotely, or at clinics or the participants' homes. Unblinded site coordinators reviewed electronic clinical notes for the period of participation to collect health economic data.

Assessors were trained to administer the assessment battery by the lead trial coordinator, completed practice assessments and were observed by site coordinators before working independently. Scoring fidelity meetings for the PSYRATS-AH were conducted repeatedly and all assessors attended weekly supervision with coordinators to maintain scoring accuracy and consistency.

Participants were invited to provide additional consent to take part in the ESM assessments at their initial consent into the trial. If they provided consent, they were invited to complete the assessments at every time point (baseline, 16 and 28 weeks). The ESM assessment week consisted of ten questionnaires a day for 6 days and was delivered through the m-Path smartphone application (https://m-path.io/landing/). This provides a self-report of mental state in the flow of daily life. Participants could use their own phones or borrow one from the study team to complete the study. The items contributing to the anxiety score were as follows: right now, I feel: relaxed: 1 not at all, 7 very much so; safe: 1 not at all, 7 very much so; stressed: 1 not at all, 7 very much so; wound up: 1 not at all, 7 very much so; scared: 1 not at all, 7 very much so.

Participants in all three trial arms were compensated GBP20, with an additional GBP15 for the experience sampling assessment, at each time point.

Study hypotheses

The study investigated the following hypotheses: (1) AV-BRF will be more effective in reducing voice-related distress, total voice severity and voice frequency than TAU after treatment (16 weeks) and at the follow-up (28 weeks); (2) AV-EXT will be more effective in reducing voice-related distress, total voice severity and voice frequency than TAU after treatment (16 weeks) and at the follow-up (28 weeks); (3) greater baseline complexity of voice characterization will moderate the treatment effects of AV-BRF and AV-EXT compared to TAU. Other clinical characteristics will be explored as potential moderators. The following additional study hypotheses will be reported in subsequent publications: (1) AV-EXT will reduce perceived omnipotence and malevolence compared to TAU and these improvements will mediate change in the primary outcome; (2) in both AV-BRF and AV-EXT, the treatment effects on the primary outcome will be mediated by anxiety reduction, as measured by the ESM in daily life; (3) AV-BRF and AV-EXT will both have favorable incremental cost-effectiveness ratios compared to routine care.

We prespecified the interpretation of our results as follows: for each comparison of AVATAR-BRF versus TAU, and AVATAR-EXT versus TAU, if the estimated between-group difference at 16 weeks is statistically significant, we will conclude that there is a treatment effect on the outcome at the end of the intervention period. This will constitute partial support for our hypothesis; if the estimated between-group difference at 28 weeks is statistically significant, we will conclude that there is a treatment effect on the outcome at the follow-up. If there is a statistically significant between-group difference at 28 weeks but not at the earlier 16-week time point, this will constitute partial support for our hypothesis; if there is a statistically significant between-group difference at both time points, we will conclude that the treatment effect is sustained and this will constitute full support for our hypothesis; for the primary outcome of PSYRATS voice-related distress, we will assess the magnitude of the between-group difference against the plausible effect sizes in the sample size calculations.

Outcomes

The prespecified primary outcome for the study was reduction in distress associated with voices at end of treatment (16 week) and at the follow-up (28 weeks), as measured by the distress dimension of the PSYRATS-AH (five items, distress (two items), negative content (two items) and control[13]. The PSYRATS-AH is a dimensional, semistructured, assessor-rated clinical interview assessing AHs, consisting of 11 items, each item scored from zero (voices not present) to four[26]. Voice-related distress was selected as the primary outcome because it is the central target of the therapy approach and valued as an outcome by experts by experience (Supplementary Table 9).

Key secondary outcomes, as specified in the primary hypotheses, were reductions in the voice frequency scale score (three items: frequency, duration and disruption) and the total severity score (all 11 items) on the PSYRATS-AH Scale at 16 and 28 weeks.

Other secondary outcomes were a mix of assessor-rated and self-reported measures, with effects estimated at 16 and 28 weeks. These included distressing beliefs (PSYRATS-Delusions[26]), well-being (WEMWBS[28]), psychological recovery (CHOICE[29]), fearful attachment (Relationships Questionnaire item[30]), VAAS[31], measuring acceptance-based attitudes and actions in relation to voice-hearing experiences, mood (DASS[32] and BDI[33]), anxiety in daily life (using the ESM), voice power (VPDS item[34]) and BAVQ (omnipotence, malevolence and benevolence, total, BAVQ-R[35]), and trauma-related symptoms (ITQ[36]) (16 weeks only).

The clinical characteristics of participants were further assessed at baseline with the Clinical Assessment Interview for Negative Symptoms (CAINS[37]) and Scale for Assessment of Positive Symptoms[38] (further details of all measures are provided in Supplementary Table 12).

Safety

All AEs and SAEs were recorded according to the trial standard operating procedure for AEs, following CONSORT guidance, with the extension for social and psychological interventions, and the extension for reporting of harms. The chief investigator reviewed all reports and notified the independent DMEC Chair of any SAEs as they occurred. The DMEC Chair was responsible for reviewing all SAEs and determining the relatedness, if any, of SAEs to the trial procedures (rating as yes, no, possibly related). AEs were recorded for the duration of each participant's involvement in the trial, from the date on which they signed the consent form until the date of the final assessment or contact with the trial team if they withdrew before their final assessment. Monitoring was conducted by therapists and research assistants, supervised by trial coordinators throughout their contact with participants. After the conclusion of the final assessment for each participant, the trial coordinator reviewed the electronic clinical notes and logged any AEs during their participation in the trial.

All AEs were discussed weekly in trial coordinator meetings and monthly at clinical trial management committee meetings to ensure accurate and consistent monitoring across sites. Where an event was determined to be serious by the site trial coordinator and principal investigator, this form was sent to the DMEC Chair for further review and to determine the rating of the relatedness of the event to any trial procedure.

The criteria for determining whether an incident should be considered a serious or nonserious AE are shown below and were included within a standard operating procedure for AE reporting, followed by all staff during the trial. An AE was defined as: any untoward medical occurrence, unintended disease or injury, or untoward clinical signs in participants that lead to significant increased distress and interference with daily life such that intervention from the clinical team was required.

This included AEs related to both intervention arms of the AVATAR2 trial and to the TAU group, and to all research procedures involved. It was anticipated that participants might experience some distress in relation to the assessment measures or treatment processes. If this distress was managed by the trial team and did not require additional support from clinical services, then this was not classified as an AE.

An AE was defined as serious (that is, an SAE) by the ISO 14155:2011 guidelines for medical device trials if it: resulted in death OR was a life-threatening illness or injury OR required (voluntary or involuntary) hospitalization or prolongation of existing hospitalization OR resulted in persistent or significant disability or incapacity OR medical or surgical intervention was required to prevent any of the above OR led to fetal distress, fetal death or consisted of a congenital anomaly or birth defect OR was otherwise considered medically significant by the investigator.

Life-threatening in the definition of an SAE refers to an event in which the individual was at risk of death at the time of the event; it does not refer to an event that might hypothetically have caused death if it were more severe. Events that are not immediately life-threatening or do not result in death or hospitalization but may jeopardize the individual or may require intervention to prevent one or the other outcomes listed, should be considered serious.

A planned hospitalization for a preexisting condition, without a serious deterioration in health, is not considered an SAE.

Statistical analysis and sample size calculations

We powered the study to detect plausible effect sizes based on our previous AVATAR therapy trial[8]. There we found a clinically meaningful reduction in PSYRATS-AH distress of 4.8 points, with an effect size of approximately $d = 0.8$, but we reduced this for the current trial to take into consideration the increase in number of centers, the follow-up comparison (not only the end of treatment) and a more pragmatic trial design. We are accounting for two formal comparisons: AV-EXT versus TAU—plausible effect size = 0.6; and AV-BRF versus TAU—plausible effect size = 0.5. The study was powered for an overall treatment effect at a 5% significance level, accounting for two multiple group comparisons in which the tests are correlated because of shared control data (at $r = 0.5$), giving an alpha level for each group-specific test of 0.035. Accordingly, a sample size of 92 per group or 276 in total in the analysis dataset had 90% power to detect a minimum clinically significant difference (effect size) of 0.5 standard deviations. We sought to recruit 345 participants at baseline (87 per site), with $n = 115$ per treatment arm, allowing for a conservative attrition rate of 20%.

We report the findings in line with the most recent relevant CONSORT guidelines, the 2018 extension for reporting social and psychological intervention trials[39]. No interim analysis was performed. All analyses were conducted in Stata v.18.1 (ref. 40). To visualize the data, R v.4.33 (ref. 41) and ggplot2 v.3.5.0 (ref. 42) were used. The senior statistician (R.E.) was unblinded only after completion of the initial analyses and presentation of these results to our external advisory committees. The junior statistician (H.J.) was unblinded during the study after preparing the first closed DMEC report; the statistical analysis was performed unblinded owing to the need to account for therapist effects in the AVATAR arms.

The primary estimand is the treatment policy estimand. The primary analyses were carried out using the ITT sample: participants were analyzed in the group they were randomized to; available data from all participants are included, including those who did not complete treatment.

The primary analyses of the hypotheses of between-group differences in the AV-EXT versus TAU and AV-BRF versus TAU in voice distress as measured using the PSYRATS-AH distress score were analyzed using a mixed-effects (random) model at all post-randomization time points (weeks 16 and 28). Fixed effects were the center, baseline assessment for the outcome under investigation, voice characterization (low or / high), treatment, time (categorical, 16 or 28 weeks) and time × treatment interactions. Marginal treatment effects were estimated for the outcomes at each time point and reported separately as adjusted mean differences in scores between the randomized groups with 96.5% CIs and two-sided *P* values.

To account for the partial nested design, we included a random intercept for therapist in the treatment arms only, with the participants in the TAU arm considered as being in individual clusters of size one. The same therapists delivered both AV-EXT and AV-BRF. Participants in the intervention arms who did not attend any sessions with a therapist were nominally allocated to a single therapist ID for ITT purposes. Participant was included as a random intercept nested within therapist to account for repeated measures of outcomes.

For the continuous secondary outcomes we followed the same model as the primary analysis: linear mixed models, including the outcome measures at all post-randomization time points, with a time by treatment interaction to allow the estimation of the between-arm difference at each time point.

All statistical models were estimated using maximum likelihood estimation, which allows for missing outcome data under the missing at random assumption. In addition, we report estimates for Cohen's *d* effect sizes at 16 and 28 weeks as the adjusted mean difference of the outcome divided by the sample s.d. of the outcome at baseline. CIs for Cohen's *d* were calculated by dividing the 96.5% confidence limits by the sample s.d. of the outcome at baseline. These are displayed in a forest plot with the primary outcome at the top, followed by key secondary and other outcomes, with a separate plot for each time point.

The moderation analysis investigated how a prespecified set of putative baseline moderators affected the efficacy of the treatment interventions (TAU, AV-BRF, AV-EXT) in reducing the distress associated with AHs over time at 16 and 28 weeks. Specifically, the interaction effects between treatment groups, time and moderator were analyzed to understand the differential influence of moderator on the treatment effects between TAU versus AV-EXT and TAU versus AV-BRF. We investigated the specific hypothesis that greater baseline complexity of voice characterization would moderate the treatment effects of AVATAR-BRF and AVATAR-EXT compared to TAU. The following measures of baseline clinical and cognitive characteristics were also considered as potential moderators of treatment effects: PSYRATS-AH-Distress, trauma-related symptoms: PTSD, DSO, negative symptoms (CAINS for motivation and pleasure, and CAINS for expressiveness), duration of mental health services (early versus not early), duration of hearing voices, age voices started and attachment (Relationship Questionnaire). We also examined demographic variables as moderators: age, gender and self-defined ethnicity.

For a continuous moderator, the difference in treatment effect between the unit levels of the moderator can be interpreted as the difference in the estimated treatment effect between a participant with a moderator value at baseline of $a + 1$ and a participant with a moderator value at baseline of a. For a binary moderator (for example, low versus high voice characterization), the difference in treatment effect can be interpreted as the difference in the estimated treatment effect between participants with low and those with high voice characterization.

CACE compares the average outcome between those in the AV-EXT and AV-BRF groups who meet the definition of receiving a minimal treatment dose and the latent subgroup of people in the control group who would have received this dose had they been randomized to the respective intervention group (that is, this is a hidden counterfactual). For each outcome, we would expect to see an increased effect estimate for CACE relative to the ITT effect because we are systematically excluding people who do not receive the treatment dose of AV-EXT or AV-BRF. However, the statistical significance does not necessarily increase because the instrumental variable method used to calculate the CACE estimates produces larger standard errors as a consequence of accounting for the selection effects between those who receive and those who do not receive a treatment dose in the intervention arm. As such, this is best seen as a bias correction that answers the question 'what is the effect of offering AVATAR compared to TAU in those who would receive a treatment dose of AVATAR if offered?'

We estimated the CACE for each comparison of AV-EXT versus TAU, and AV-BRF versus TAU, separately; this means that we excluded from the model the AVATAR intervention group not being used in the comparison. We used an instrumental variable method with two-stage least squares estimation and robust standard errors to account for clustering by therapist. We estimated the effect for each time point using separate analyses. Randomization was used as the instrument; receipt of a minimal treatment dose was the endogenous (treatment received) variable. We included the same set of covariates as the primary analysis models in both stage regressions.

Further details of the statistical methods, including the treatment of missing data, are provided in the statistical analysis plan (published in the ISRCTN registry with the identifier ISRCTN55682735). A post hoc sensitivity analysis for missing data in the primary outcome is provided in Supplementary Table 13 and Supplementary Fig. 1).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Open access information on the AVATAR2 trial, such as the trial protocol and statistical analysis plan, including the example analysis code, has been published in the ISRCTN registry with the identifier ISRCTN55682735; the final trial protocol (v.1.2) was also published in *Trials*[11]. Individual participant data have been deposited in the King's Open Research Data System, but access is restricted due to privacy reasons and general data protection regulations, and can only be accessed after review. Data will be made accessible after the publication of this paper. A request can be made by academic or clinical researchers to research.data@kcl.ac.uk for the purpose of conducting noncommercial, ethically approved research. The research data team will review the request against the conditions set out in the Data Access Agreement. An initial response to requests will be formulated within a month. A Data Access Agreement will be drawn up before data can be shared.

Code availability

In accordance with the data availability protocol, the corresponding statistical code will be provided as part of the Data Access Agreement.

References

26. Haddock, G., McCarron, J., Tarrier, N. & Faragher, E. B. Scales to measure dimensions of hallucinations and delusions: the psychotic symptom rating scales (PSYRATS). *Psychol. Med.* **29**, 879–889 (1999).
27. Edwards, C. J. et al. The Voice Characterisation Checklist: psychometric properties of a brief clinical assessment of voices as social agents. *Front. Psychiatry* **14**, 1192655 (2023).
28. Tennant, R. et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health Qual. Life Outcomes* **5**, 63 (2007).
29. Greenwood, K. E. et al. CHoice of Outcome In Cbt for psychosEs (CHOICE): the development of a new service user-led outcome measure of CBT for psychosis. *Schizophr. Bull.* **36**, 126–135 (2010).
30. Bartholomew, K. & Horowitz, L. M. Attachment styles among young adults: a test of a four-category model. *J. Pers. Soc. Psychol.* **61**, 226–244 (1991).
31. Shawyer, F. et al. The voices acceptance and action scale (VAAS): pilot data. *J. Clin. Psychol.* **63**, 593–606 (2007).
32. Henry, J. D. & Crawford, J. R. The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. *Br. J. Clin. Psychol.* **44**, 227–239 (2005).
33. Beck, A. T., Steer, R. A. & Brown, G. K. *BDI-II: Beck Depression Inventory* (Pearson, 1996).
34. Birchwood, M., Meaden, A., Trower, P., Gilbert, P. & Plaistow, J. The power and omnipotence of voices: subordination and entrapment by voices and significant others. *Psychol. Med.* **30**, 337–344 (2000).
35. Chadwick, P., Lees, S. & Birchwood, M. The revised Beliefs About Voices Questionnaire (BAVQ–R). *Br. J. Psychiatry* **177**, 229–232 (2000).
36. Cloitre, M. et al. The International Trauma Questionnaire: development of a self-report measure of ICD-11 PTSD and complex PTSD. *Acta Psychiatr. Scand.* **138**, 536–546 (2018).
37. Kring, A. M., Gur, R. E., Blanchard, J. J., Horan, W. P. & Reise, S. P. The Clinical Assessment Interview for Negative Symptoms (CAINS): final development and validation. *Am. J. Psychiatry* **170**, 165–172 (2013).

38. Andreasen, N. C *Scale for the Assessment of Positive Symptoms (SAPS)* (Univ. of Iowa, 1985).
39. Grant, S. et al. CONSORT-SPI 2018 Explanation and Elaboration: guidance for reporting social and psychological intervention trials. *Trials* **19**, 406 (2018).
40. *Stata Statistical Software: Release 18* (StataCorp LLC, 2023).
41. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2024).
42. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).

Acknowledgements

This research was funded by a Wellcome grant no. 215471/Z/19/Z. The study was supported by the National Institute for Health and Care Research (NIHR), the Wellcome Trust King's Clinical Research Facility, the NIHR Maudsley Biomedical Research Centre and Maudsley NHS Foundation Trust, King's College London (P.A.G. and R.E., no. NIHR203318) and the Manchester Biomedical Research Centre (S.B. and G.H., no. NIHR203308). S.B. discloses support for publication of this work from the NIHR research professorship (no. NIHR300794). R.E. discloses support for publication of this work from the NIHR research professorship (no. NIHR300051). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. This study was supported by NHS Research Scotland through the Chief Scientist Office and the NHS Scotland Mental Health Network (A.G. and H.J.M., no. NRSMHN/2021/01). M.R.-C. discloses support for the publication of this work from the Sofja Kovalevskaja Award from the Alexander von Humbold Foundation and the Ministry of Education and Research in Germany (SKP2020RusCalafell). The funding bodies did not have any role in study design, data collection and analysis, data interpretation or writing of the manuscript. We thank the following people who contributed to the AVATAR2 trial but are not listed as authors. First and foremost, we thank all the trial participants who consented to take part and generously provided their data, and the clinical teams who referred and supported the participants. We also thank the following Trial Steering Committee members: M. Broome, D. McDaid, R. Dudley, N. Sirey, M. Arbuthnott and F. Lobban. We thank the following members of the Data Monitoring & Ethics Committee: K. Greenwood, R. Ogollah and J. Hodgekins. We thank the following collaborators: E. Peters, I. Myin-Germeys, P. McCrone, M. Wampers, S. Apers, E. Gkaintatzi and the NHS co-principal investigators S. Johnson (Camden and Islington NHS Foundation Trust), K. Livingstone (NHS Lanarkshire), A. Hill (Oxleas NHS Foundation Trust), J. Ellis and S. Kaye (Pennine Care NHS Foundation Trust). We thank the following PPI consultants: A. Grant, Alison, N. Sirey, S. Gogan, J. Shaylor, J. Harling, S. Joveini, F. Yanga, L. Rankin, J. Adeshina, L. Christie, L. Joachim, C. Ashe, C. Mackenzie-Nash, A. Phiri, C.-L. Burns, Z. Howarth, the McPin foundation and other PPI consultants who made an important contribution but have chosen not to be named. We thank S. Syrett (Peer Researcher), C. Myrie and J. Hall (Ethnicity Diversity and Inclusion Consultants), R. Underwood for his contribution in covering the KCL site coordination during the lead coordinator's maternity leave, and the following research workers: L. Williams, N. Hill, O. Owrid, L. Miller, T. Sowemimo, R. Patel, S. Aslam, S. Allan, B. Rudd, L. Richardson, K. O'Hare, C. Matrunola, L. Grice, R. Wadsworth, R. Bailey, D. Sher, C. Graves, A. McGee and D. Ramsell. Finally, we thank the following therapists (trained in AVATAR therapy, who saw either the trial participants or those allocated to TAU who chose to receive treatment after completion of their involvement): A. Adams, Z. Brittenden, T. Stevens, O. Collier, C. Lankester, E. Primrose, P. Cawthorne, B. Brandrett, A. Scholes, A. Larkin, R. Cook, C. O'Brien, G. Blanchard, C. Trifonov, E. Quinlan, H. Ferner, F. Pearce, M. O'Brien, N. Neath, J. Dix, J. Wilson and S. Sluman.

Author contributions

P.A.G., T.W., T.K.J.C., R.E., M.H. and M.R.-C. conceived and designed the study. P.A.G., T.K.J.C., R.E., M.H., A.G., G.H. and M.F.-A obtained the funding. P.A.G. was the study's principal investigator. T.K.J.C. was the study's co-principal investigator. T.W. led and T.K.J.C. co-led treatment provision and training. C.J.E. coordinated, with the local site leads J.McD., M.C. and M.F., all aspects of trial management, recruitment, data collection and processing. M.H. designed the software. M.F.-A., A.G., G.H., S.B. and H.J.M. were the local site principal investigators. H.B., A.M. and N.X. led the local site treatment training and provision. A.H. led trauma treatment development. R.E., C.J.E. and H.J. had full access to all the data. R.E. and H.J. were responsible for the data analyses. P.A.G., T.W., C.J.E., T.K.J.C., M.H. and R.E. wrote the first draft of the paper. All authors edited subsequent revisions of the draft and approved the final paper. The authors agree to be accountable for the work.

Competing interests

M.H. is a shareholder of Avatar Therapy Ltd. T.K.J.C. and P.A.G. are unpaid scientific advisers to Avatar Therapy Ltd. S.B. is Director and shareholder of CareLoop Health Ltd, which develops and markets digital therapeutics for schizophrenia and a digital screening app for postnatal depression. The other authors declare no competing interests.

Additional information

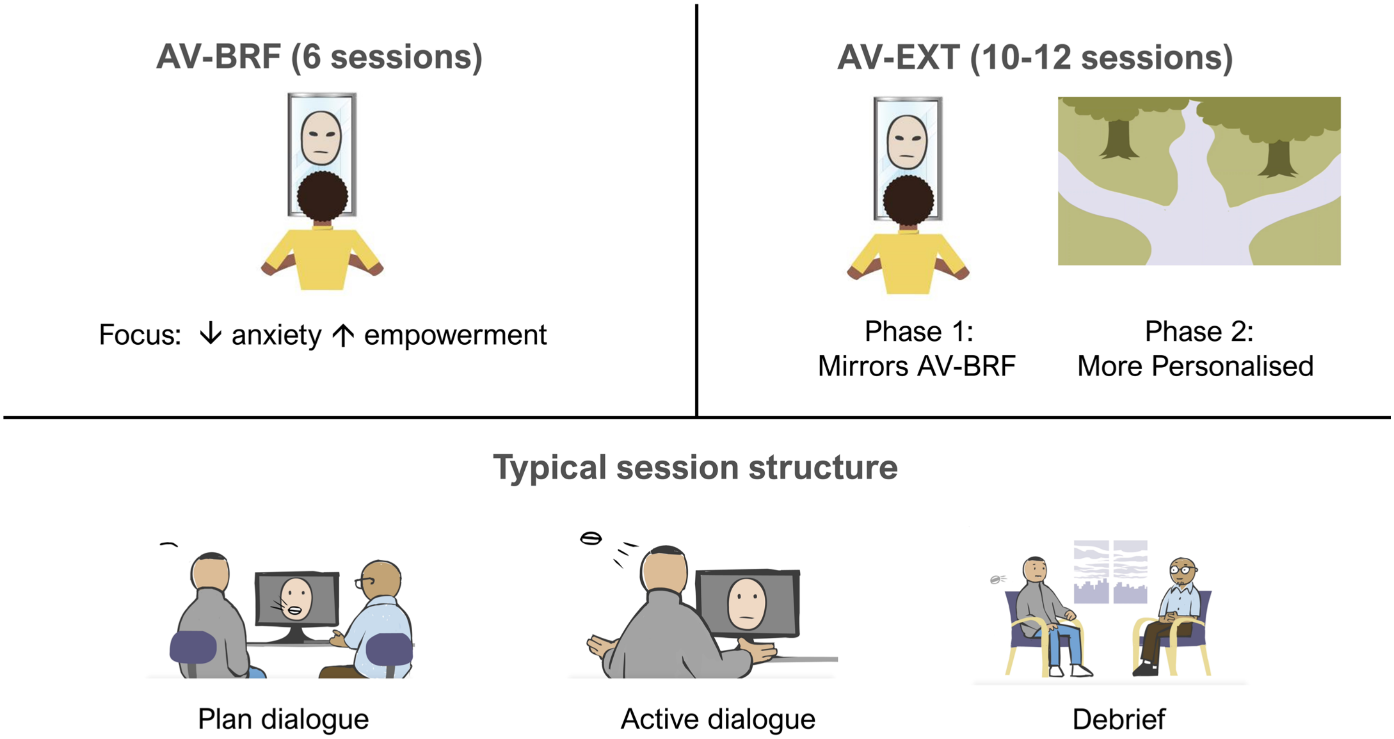
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03252-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03252-8.

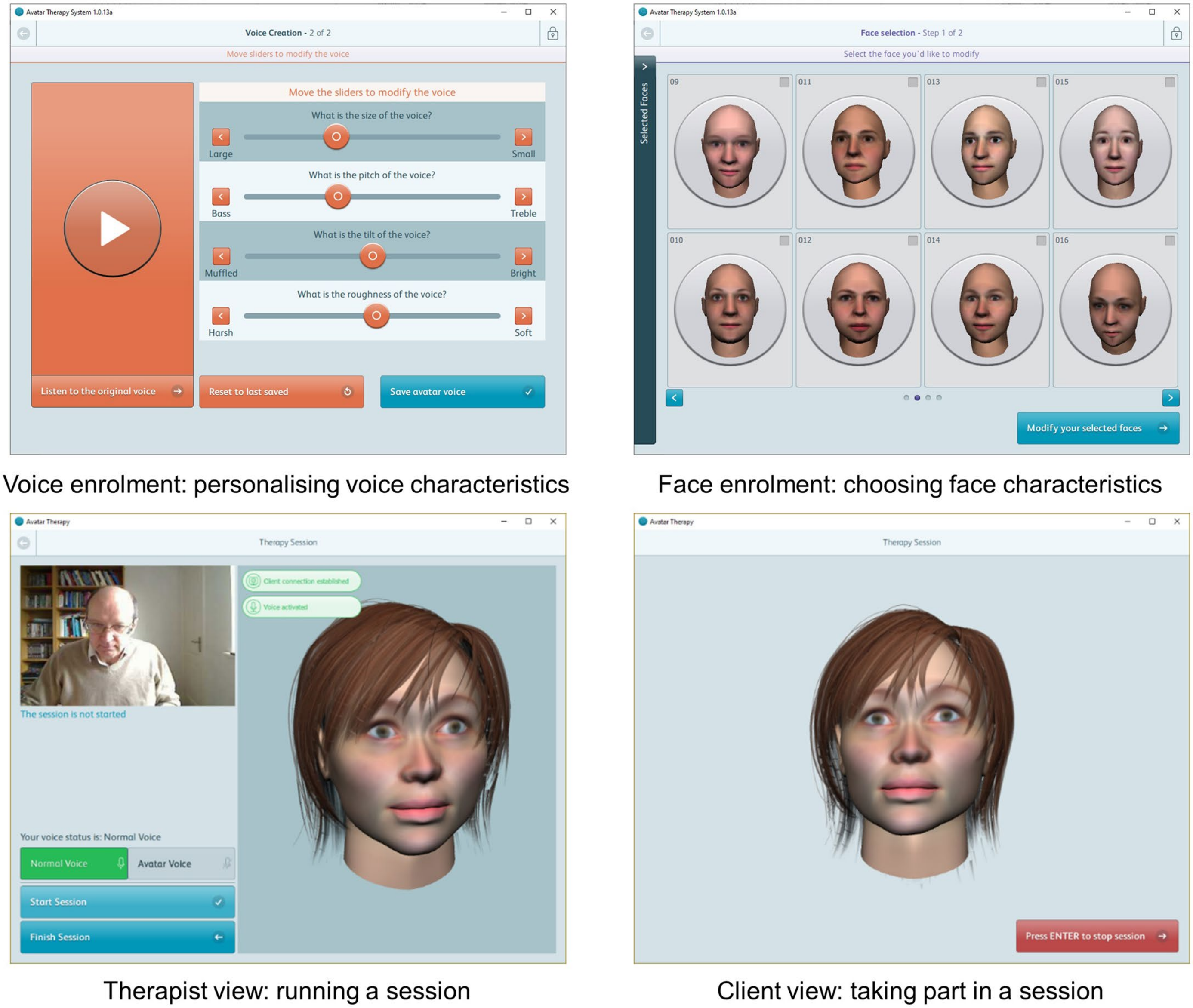
Correspondence and requests for materials should be addressed to Philippa A. Garety or Thomas Ward.

Peer review information *Nature Medicine* thanks Tao Chen, Toshi Furukawa and Thole Hoppen for their contribution to the peer review of this work. Primary Handling Editor: Ming Yang, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Illustration of AV-BRF and AV-EXT and typical structure for an active dialogue session. Treatment foci and session structure.


Avatar Therapy System 1.0.13a
Voice Creation - 2 of 2
Move sliders to modify the voice
Move the sliders to modify the voice
What is the size of the voice?
Large
Small
What is the pitch of the voice?
Bass
Treble
What is the tilt of the voice?
Muffled
Bright
What is the roughness of the voice?
Harsh
Soft
Listen to the original voice
Reset to last saved
Save avatar voice
Avatar Therapy System 1.0.13a
Face selection - Step 1 of 2
Select the face you'd like to modify
Selected Faces
09
011
013
015
010
012
014
016
Modify your selected faces
Avatar Therapy
Therapy Session
Client connection established
Voice activated
The session is not started
Your voice status is: Normal Voice
Normal Voice
Avatar Voice
Start Session
Finish Session
Avatar Therapy
Therapy Session
Press ENTER to stop session

Voice enrolment: personalising voice characteristics

Face enrolment: choosing face characteristics

Therapist view: running a session

Client view: taking part in a session

Extended Data Fig. 2 | Illustration of AVATAR Therapy therapist and client interface. AVATAR therapy software interface.

Extended Data Table 1 | Descriptive statistics for primary and other secondary voice-related outcomes across three treatment groups at each time point

		TAU			AV-BRF			AV-EXT		
Outcome	**Time**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
PSYRATS-AH-Distress	Baseline	115	15.70	2.785	116	15.72	2.721	114	15.89	2.768
Range: 0-20	Week 16	103	15.28	3.835	98	13.98	4.863	98	13.81	4.757
	Week 28	104	14.09	5.063	101	13.35	5.292	93	13.33	5.265
PSYRATS-AH-Total	Baseline	115	30.64	4.420	116	30.09	4.658	114	30.11	4.423
Range: 0-44	Week 16	103	29.46	6.751	98	26.78	8.419	98	26.48	8.274
	Week 28	104	27.85	8.836	101	25.84	9.503	93	25.67	8.987
PSYRATS-AH-Frequency	Baseline	115	7.87	1.949	116	7.39	2.113	114	7.06	2.023
Range: 0-12	Week 16	103	7.09	2.210	98	6.31	2.522	98	6.00	2.479
	Week 28	104	7.06	2.562	101	6.19	2.763	93	5.75	2.600
BAVQ Omnipotence	Baseline	114	6.71	3.485	116	6.89	3.451	113	6.99	3.853
Range: 0-18	Week 16	94	7.46	3.988	94	8.34	4.403	88	8.36	4.289
	Week 28	95	7.62	3.912	92	8.53	4.364	83	8.94	4.511
BAVQ Malevolence	Baseline	114	6.24	4.231	116	6.74	4.229	113	6.71	4.529
Range: 0-18	Week 16	94	7.06	4.979	94	7.79	4.996	88	7.91	4.484
	Week 28	95	7.68	4.783	92	7.54	4.386	83	7.98	4.362
BAVQ Benevolence	Baseline	114	3.52	4.260	116	3.32	3.666	113	3.02	3.787
Range: 0-18	Week 16	94	3.01	3.542	94	3.04	3.821	88	3.48	4.288
	Week 28	95	3.43	4.184	92	3.18	4.038	83	3.58	4.208
BAVQ Total	Baseline	114	34.13	12.225	116	33.97	12.109	113	34.38	10.671
Range: 0-105	Week 16	94	34.55	11.994	93	37.74	13.576	88	38.43	11.506
	Week 28	95	36.72	13.713	92	37.97	12.942	83	39.35	12.480
VAAS Acceptance	Baseline	114	47.55	6.885	116	47.07	6.544	112	48.28	7.329
Range: 16-80	Week 16	94	48.15	6.952	92	50.84	8.435	88	52.46	8.503
	Week 28	95	48.87	7.916	92	51.48	8.281	83	53.12	7.774
VAAS Action	Baseline	114	46.43	7.846	116	46.17	9.539	112	48.09	9.458
Range: 15-75	Week 16	94	47.75	8.914	92	49.09	9.460	88	51.95	10.116
	Week 28	95	47.92	8.816	92	50.02	10.067	83	51.69	9.657
VAAS Full Scale	Baseline	114	93.99	12.810	116	93.25	14.363	112	96.38	14.935
Range: 31-155	Week 16	94	95.90	14.197	92	99.92	16.830	88	104.41	17.006
	Week 28	95	96.79	15.066	92	101.50	16.698	83	104.81	15.681
Voice Power Differential Scale	Baseline	112	3.36	1.222	114	3.40	1.287	113	3.31	1.247
Range: 1-5	Week 16	98	3.29	1.324	94	3.02	1.328	91	2.87	1.222
	Week 28	99	3.18	1.265	93	3.09	1.248	86	2.81	1.101
Hallucinations Remission Score	Baseline	115	3.61	0.541	116	3.66	0.527	114	3.66	0.477
Range: 0-4	Week 16	104	3.49	0.848	99	3.38	0.804	97	3.28	0.875
	Week 28	104	3.26	1.052	101	3.27	1.019	93	3.20	1.099

AH, auditory hallucinations; AV-BRF, Brief AVATAR therapy; AV-EXT, Extended AVATAR Therapy; BAVQ, Beliefs About Voices Questionnaire; N: Number, PSYRATS: Psychotic Symptom Rating Scales; TAU, treatment as usual; VAAS, Voices Acceptance And Action Scale. Time points: 16, 16 weeks after baseline (after-treatment follow-up); 28, 28 weeks after the baseline follow-up.

Extended Data Table 2 | Descriptive statistics for other secondary outcomes across three treatment groups at each time point

		TAU			AV-BRF			AV-EXT		
Outcome	**Time**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
PSYRATS Delusions	Baseline	109	14.86	4.900	105	14.03	5.740	108	15.02	4.386
Range: 0-24	Week 16	95	13.01	6.213	86	10.55	7.184	88	10.68	6.617
	Week 28	95	13.04	6.145	83	10.86	7.050	84	10.11	7.051
WEMWBS	Baseline	112	35.90	11.030	115	37.37	10.360	112	37.05	10.983
Range: 14-70	Week 16	93	34.35	10.827	93	37.60	10.930	87	41.13	11.849
	Week 28	93	34.80	11.141	91	39.04	12.860	82	40.23	12.667
CHOICE	Baseline	113	45.96	23.321	115	51.18	21.456	110	46.99	21.756
Range: 0-100	Week 16	93	43.83	21.492	93	54.21	24.144	87	55.76	24.459
	Week 28	93	48.45	23.491	91	58.47	25.352	81	57.10	21.400
DASS Anxiety	Baseline	114	20.50	10.390	115	18.78	10.152	112	19.21	10.294
Range: 0-42	Week 16	93	22.74	10.263	93	16.99	10.235	87	17.54	10.044
	Week 28	94	20.06	10.414	91	15.83	9.450	82	17.58	10.784
DASS Depression	Baseline	114	24.04	10.630	115	22.10	11.754	112	22.57	11.536
Range: 0-42	Week 16	93	25.05	11.122	93	19.84	10.290	87	19.40	11.147
	Week 28	94	23.64	11.031	91	18.99	11.689	82	20.10	11.537
DASS Stress	Baseline	114	24.07	9.567	115	22.22	10.399	112	22.55	10.614
Range: 0-42	Week 16	93	25.34	9.169	93	19.76	9.380	87	20.34	10.500
	Week 28	94	23.98	9.479	91	19.16	9.629	82	21.65	11.144
BDI	Baseline	113	32.29	14.595	115	27.72	13.965	112	30.61	15.489
Range: 0-63	Week 16	94	31.68	14.415	91	25.50	15.428	89	25.98	14.580
	Week 28	94	30.63	15.295	91	24.51	15.097	83	27.11	15.462
Anxiety (ESM)	Baseline	68	15.00	3.688	67	14.12	4.293	74	13.83	4.592
Range: 4-32	Week 16	40	15.44	5.101	35	13.08	4.907	49	11.77	4.388
	Week 28	38	13.83	4.645	30	13.28	5.557	38	12.50	5.256
ITQ DSO	Baseline	104	14.91	5.910	112	13.75	6.580	108	14.46	6.330
Range: 0-24	Week 16	87	14.51	6.200	89	12.13	6.420	82	12.44	6.510
ITQ PTSD	Baseline	105	11.80	7.060	112	10.45	6.950	106	11.65	7.240
Range: 0-24	Week 16	88	11.19	6.960	89	9.77	6.440	83	10.44	6.680

AV-BRF, Brief AVATAR therapy; AV-EXT, Extended AVATAR Therapy; BDI, Beck Depression Inventory; CHOICE, CHoice of Outcome In Cbt for PsychosEs; DASS, Depression Anxiety Stress Scales; DSO, Disturbances in self-organization; ESM, Experience Sampling Method; ITQ, International Trauma Questionnaire; PTSD, Post-Traumatic Stress Disorder; TAU, treatment as usual; WEMWBS, Warwick-Edinburgh Mental Well-being Scale. Time points: 16, 16 weeks after baseline (after-treatment follow-up); 28, 28 weeks after the baseline follow-up.

Extended Data Table 3 | Moderation analysis for demographic variables: estimated margins of interaction effect between moderator variable and treatment group, including 96.5% confidence intervals

Moderator	Comparison	Time	Treatment	*β*	SE	Z	*P*	96.5% CI	
Age	Continuous	16	AV-BRF vs TAU	-0.052	0.033	-1.571	0.116	-0.1208	0.0176
	Continuous	16	AV-EXT vs TAU	-0.041	0.056	-0.733	0.463	-0.1582	0.0766
	Continuous	28	AV-BRF vs TAU	-0.042	0.041	-1.022	0.307	-0.1276	0.0443
	Continuous	28	AV-EXT vs TAU	-0.008	0.073	-0.110	0.912	-0.1617	0.1457
Gender	Female vs Male	16	AV-BRF vs TAU	0.704	1.174	0.600	0.549	-1.7705	3.1789
	Female vs Male	16	AV-EXT vs TAU	0.791	0.940	0.842	0.400	-1.1899	2.7717
	Female vs Male	28	AV-BRF vs TAU	-1.884	1.627	-1.158	0.247	-5.3141	1.5452
	Female vs Male	28	AV-EXT vs TAU	-1.093	1.358	-0.805	0.421	-3.9552	1.7700
Ethnicity	Black vs White	16	AV-BRF vs TAU	0.447	1.499	0.298	0.765	-2.7141	3.6087
	Black vs White	16	AV-EXT vs TAU	-1.895	1.623	-1.167	0.243	-5.3171	1.5275
	Asian vs White	16	AV-BRF vs TAU	2.730	1.301	2.098	0.036	-0.0138	5.4728
	Asian vs White	16	AV-EXT vs TAU	2.259	1.436	1.573	0.116	-0.7684	5.2866
	Other vs White	16	AV-BRF vs TAU	0.881	1.626	0.542	0.588	-2.5466	4.3091
	Other vs White	16	AV-EXT vs TAU	-0.040	1.628	-0.025	0.980	-3.4722	3.3920
	Black vs White	28	AV-BRF vs TAU	0.804	1.818	0.442	0.658	-3.0298	4.6371
	Black vs White	28	AV-EXT vs TAU	-3.391	1.622	-2.090	0.037	-6.8113	0.0290
	Asian vs White	28	AV-BRF vs TAU	2.359	1.683	1.401	0.161	-1.1901	5.9088
	Asian vs White	28	AV-EXT vs TAU	2.575	1.809	1.424	0.154	-1.2380	6.3884
	Other vs White	28	AV-BRF vs TAU	2.678	2.112	1.268	0.205	-1.7757	7.1315
	Other vs White	28	AV-EXT vs TAU	2.068	1.914	1.080	0.280	-1.9684	6.1040
IMD Quintile	Q2 vs Q1	16	AV-BRF vs TAU	2.033	1.297	1.567	0.117	-0.7019	4.7681
	Q2 vs Q1	16	AV-EXT vs TAU	2.404	1.108	2.169	**0.030**	0.0671	4.7402
	Q3 vs Q1	16	AV-BRF vs TAU	0.540	0.983	0.549	0.583	-1.5331	2.6132
	Q3 vs Q1	16	AV-EXT vs TAU	0.494	1.517	0.325	0.745	-2.7048	3.6925
	Q4 vs Q1	16	AV-BRF vs TAU	1.624	4.136	0.393	0.695	-7.0967	10.3448
	Q4 vs Q1	16	AV-EXT vs TAU	9.023	3.395	2.657	**0.008**	1.8641	16.1814
	Q5 vs Q1	16	AV-BRF vs TAU	1.409	1.451	0.971	0.331	-1.6495	4.4675
	Q5 vs Q1	16	AV-EXT vs TAU	2.061	1.764	1.168	0.243	-1.6588	5.7805
	Q2 vs Q1	28	AV-BRF vs TAU	2.142	1.494	1.434	0.152	-1.0083	5.2926
	Q2 vs Q1	28	AV-EXT vs TAU	3.274	1.323	2.474	**0.013**	0.4839	6.0642
	Q3 vs Q1	28	AV-BRF vs TAU	1.576	1.176	1.340	0.180	-0.9033	4.0547
	Q3 vs Q1	28	AV-EXT vs TAU	-2.333	1.718	-1.358	0.174	-5.9555	1.2891
	Q4 vs Q1	28	AV-BRF vs TAU	5.476	4.294	1.275	0.202	-3.5764	14.5291
	Q4 vs Q1	28	AV-EXT vs TAU	1.418	3.870	0.367	0.714	-6.7410	9.5778
	Q5 vs Q1	28	AV-BRF vs TAU	2.033	1.627	1.250	0.211	-1.3966	5.4630
	Q5 vs Q1	28	AV-EXT vs TAU	-1.488	3.723	-0.400	0.689	-9.3367	6.3601

AV-BRF, Brief AVATAR therapy; AV-EXT, Extended AVATAR therapy; IMD, Index of Multiple Deprivation; TAU, treatment as usual. Time points: 16, 16 weeks after baseline (after-treatment follow-up); 28, 28 weeks after the baseline follow-up. Mixed-effects regression models were used for the moderation analysis, with robust standard error estimation. Two-sided tests were applied, with adjustments for multiple comparisons using a significance level of 0.035 and 96.5% confidence intervals reported. The *β* coefficients represent the difference in treatment effect estimates attributable to changes between the levels of the moderator, as specified in the 'Comparison' column, for each treatment condition shown in the 'Treatment' column. Bold denotes $P \leq 0.035$.

Extended Data Table 4 | Moderation analysis for clinical variables: estimated margins of interaction effect between moderator variable and treatment group, including 96.5% confidence intervals

Moderator	Comparison	Time	Treatment	*β*	SE	Z	*P*	96.5% CI	
VoCC	Low vs High	16	AV-BRF vs TAU	-0.597	1.181	-0.506	0.613	-3.0870	1.8920
	Low vs High	16	AV-EXT vs TAU	0.189	1.091	0.173	0.863	-2.1122	2.4895
	Low vs High	28	AV-BRF vs TAU	-1.067	1.194	-0.894	0.371	-3.5836	1.4499
	Low vs High	28	AV-EXT vs TAU	-2.067	1.192	-1.733	0.083	-4.5801	0.4469
PSYRATS AH Distress	Continuous	16	AV-BRF vs TAU	0.103	0.131	0.783	0.434	-0.1738	0.3791
	Continuous	16	AV-EXT vs TAU	-0.450	0.226	-1.990	0.047	-0.9270	0.0267
	Continuous	28	AV-BRF vs TAU	0.152	0.207	0.732	0.464	-0.2855	0.5893
	Continuous	28	AV-EXT vs TAU	-0.093	0.258	-0.362	0.717	-0.6363	0.4497
PSYRATS AH Total	Continuous	16	AV-BRF vs TAU	0.007	0.090	0.074	0.941	-0.1840	0.1975
	Continuous	16	AV-EXT vs TAU	-0.184	0.132	-1.395	0.163	-0.4609	0.0938
	Continuous	28	AV-BRF vs TAU	0.062	0.148	0.417	0.676	-0.2507	0.3744
	Continuous	28	AV-EXT vs TAU	-0.095	0.150	-0.634	0.526	-0.4120	0.2216
ITQ-PTSD	Continuous	16	AV-BRF vs TAU	0.085	0.078	1.083	0.279	-0.0804	0.2503
	Continuous	16	AV-EXT vs TAU	-0.125	0.089	-1.401	0.161	-0.3120	0.0629
	Continuous	28	AV-BRF vs TAU	0.014	0.121	0.117	0.907	-0.2406	0.2688
	Continuous	28	AV-EXT vs TAU	-0.074	0.106	-0.692	0.489	-0.2981	0.1508
ITQ-DSO	Continuous	16	AV-BRF vs TAU	0.100	0.084	1.193	0.233	-0.0771	0.2780
	Continuous	16	AV-EXT vs TAU	-0.018	0.100	-0.181	0.857	-0.2290	0.1929
	Continuous	28	AV-BRF vs TAU	-0.005	0.133	-0.041	0.968	-0.2860	0.2752
	Continuous	28	AV-EXT vs TAU	-0.003	0.114	-0.030	0.976	-0.2446	0.2376
CAINS MAP	Continuous	16	AV-BRF vs TAU	0.021	0.058	0.357	0.721	-0.1015	0.1430
	Continuous	16	AV-EXT vs TAU	0.002	0.058	0.041	0.968	-0.1194	0.1241
	Continuous	28	AV-BRF vs TAU	-0.109	0.090	-1.202	0.229	-0.2994	0.0819
	Continuous	28	AV-EXT vs TAU	-0.040	0.079	-0.499	0.618	-0.2069	0.1277
CAINS EXP	Continuous	16	AV-BRF vs TAU	0.000	0.187	0.001	0.999	-0.3941	0.3944
	Continuous	16	AV-EXT vs TAU	0.233	0.216	1.080	0.280	-0.2218	0.6880
	Continuous	28	AV-BRF vs TAU	-0.049	0.202	-0.242	0.808	-0.4745	0.3766
	Continuous	28	AV-EXT vs TAU	0.308	0.177	1.742	0.082	-0.0648	0.6801
Duration MH service	Early vs Not Early	16	AV-BRF vs TAU	-0.617	1.014	-0.608	0.543	-2.7539	1.5204
	Early vs Not Early	16	AV-EXT vs TAU	-0.472	1.562	-0.302	0.763	-3.7651	2.8220
	Early vs Not Early	28	AV-BRF vs TAU	1.546	1.013	1.526	0.127	-0.5903	3.6817
	Early vs Not Early	28	AV-EXT vs TAU	1.301	1.573	0.827	0.408	-2.0160	4.6173
Relationship Q	Continuous	16	AV-BRF vs TAU	0.255	0.358	0.713	0.476	-0.4991	1.0093
	Continuous	16	AV-EXT vs TAU	-0.368	0.232	-1.585	0.113	-0.8580	0.1216
	Continuous	28	AV-BRF vs TAU	0.056	0.369	0.151	0.880	-0.7214	0.8329
	Continuous	28	AV-EXT vs TAU	-0.292	0.324	-0.902	0.367	-0.9759	0.3911
Age voice started	Continuous	16	AV-BRF vs TAU	-0.081	0.036	-2.250	**0.024**	-0.1580	-0.0050
	Continuous	16	AV-EXT vs TAU	-0.078	0.042	-1.860	0.063	-0.1660	0.0100
	Continuous	28	AV-BRF vs TAU	-0.117	0.058	-2.020	0.043	-0.2390	0.0050
	Continuous	28	AV-EXT vs TAU	-0.096	0.052	-1.842	0.065	-0.2060	0.0140
Duration of hearing voices	Continuous	16	AV-BRF vs TAU	0.011	0.038	0.281	0.779	-0.0700	0.0920
	Continuous	16	AV-EXT vs TAU	0.017	0.040	0.426	0.670	-0.0680	0.1020
	Continuous	28	AV-BRF vs TAU	0.049	0.056	0.888	0.374	-0.0680	0.1660
	Continuous	28	AV-EXT vs TAU	0.073	0.068	1.073	0.283	-0.0700	0.2160

AH, auditory hallucinations; AV-BRF, Brief AVATAR therapy; AV-EXT, Extended AVATAR therapy; DSO, Disturbances in self-organization; ITQ, International Trauma Questionnaire; MH, mental health; PSYRATS, Psychotic Symptom Rating Scales; PTSD, Post-Traumatic Stress Disorder; TAU, treatment as usual; VoCC, Voice Characterization Checklist. Time points: 16, 16 weeks after baseline (after-treatment follow-up); 28, 28 weeks after the baseline follow-up. Mixed-effects regression models were used for the analysis, with robust standard error estimation. Two-sided tests were applied, with adjustments for multiple comparisons using a significance level of 0.035 and 96.5% confidence intervals reported. The *β* coefficients represent the difference in treatment effect estimates attributable to changes between the levels of the moderator, as specified in the 'Comparison' column, for each treatment condition shown in the 'Treatment' column. Bold denotes $P \leq 0.035$.

Extended Data Table 5 | Compliance-adjusted analysis

Compliance definition	Outcome	AV-BRF vs TAU	AV-EXT vs TAU
	PSYRATS-AH	CACE (SE); p-value (96.5% CI)	CACE (SE); p-value (95% CI)
No therapy vs. some therapy + full therapy	16 Weeks	-1.23 (0.56); 0.029 (-2.42, -0.04)	-1.61 (0.63); 0.010 (-2.94, -0.29)
	28 Weeks	-0.67 (0.71); 0.347 (-2.17, 0.83)	-0.85 (0.77); 0.271 (-2.48, 0.78)
No therapy + some vs. full therapy	16 Weeks	-1.26 (0.57); 0.029 (-2.47, -0.05)	-2.20 (0.85); 0.009 (-3.98, -0.42)
	28 Weeks	-0.70 (0.74); 0.347 (-2.26, 0.87)	-1.17 (1.05); 0.268 (-3.39, 1.05)

AH, auditory hallucinations; AV-BRF, Brief AVATAR therapy; AV-EXT, Extended AVATAR Therapy; CACE, Compliance Adjusted Causal Effects; PSYRATS, Psychotic Symptom Rating Scales; TAU, treatment as usual.

nature portfolio

Corresponding author(s): Thomas Ward

Last updated by author(s): Aug 6, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

An online data collection system for clinical trials (MACRO; InferMed Ltd, Version 4.0) was used for data entry and storage. This is hosted on a dedicated server at King's College London (KCL) and managed by the KCL Clinical Trial Unit.

Data analysis

Data description and the inferential analysis were performed using Stata version 18.0. For visualising the data, R version 4.3.3 and the ggplot2 package version 3.5.0 were utilised.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Open access information on the AVATAR2 trial such as the trial protocol and statistical analysis plan, including example analysis code, is published in the ISRCTN registry with the identifier ISRCTN55682735; the final trial protocol (V1.2) was also published in Trials 11. Individual participant data have been deposited in King’s

Open Research Data System (KORDS), but access is restricted due to privacy reasons and general data protection regulations and can only be accessed after review. The data will be made accessible following the publication of this paper. A request can be made by academic or clinical researchers to <research.data@kcl.ac.uk> for the purpose of conducting non-commercial ethically approved research. The research data team will review the request against the conditions set out in the Data Access Agreement. An initial response to requests will be formulated within a month. A Data Access Agreement will be drawn up before data can be shared.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender

In compliance with the reporting requirements for sex and gender in clinical trials, this trial collected gender data based on self-reporting from participants, ensuring an inclusive approach to gender identity. Therapy completion rates have been reported in a disaggregated manner by gender. Furthermore, we conducted a moderation analysis to assess the impact of gender as a moderating variable on the effectiveness of the therapy.

Reporting on race, ethnicity, or other socially relevant groupings

In line with the standards for reporting on demographics in clinical trials, this study has collected data on participants' ethnicity and assessed socio-economic status using the Index of Multiple Deprivation (IMD). We have reported outcomes related to therapy completion rates in a disaggregated manner by ethnicity. Additionally, a moderation analysis was performed, with ethnicity and IMD Quintile serving as moderators.

Population characteristics

Our study systematically collected demographic population characteristics as presented in Table1 within the manuscript. We investigated the specific hypothesis that greater baseline complexity of voice characterisation will moderate the treatment effects of AVATAR-Brief and AVATAR-Extended compared to TAU. The following measures of baseline clinical and cognitive characteristics were also considered as potential moderators of treatment effects: PSYRATS-AH distress, Trama related symptoms: PTSD (Post-Traumatic Stress Disorder), DSO (Disturbances in Self-Organisation), Negative Symptoms (Clinical Assessment Interview for Negative Symptoms - motivation and pleasure, and expressive, CAINS MAP, CAINS EXP), duration of mental health services (early vs. not early), duration of hearing voices, age voice started, attachment (Relationship Questionnaire). We also examined demographic variables as moderators: age, gender and self-defined ethnicity.

Recruitment

Participants were recruited between 1st January 2021 and 30th November 2022, across four UK main University trial sites: the Institute of Psychiatry, Psychology & Neuroscience (King's College London), University College London, the University of Manchester and the University of Glasgow. Recruitment was conducted via referrals from clinicians at mental health services based within two (National Health Service (NHS)) providers per site, ensuring a diverse sample with respect to demography and geography. The four main NHS recruitment sites were South London and Maudsley NHS, North East London NHS Foundation Trust, Greater Manchester Mental Health NHS Foundation Trust and NHS Greater Glasgow & Clyde. The four additional NHS trusts were: Oxleas NHS Foundation Trust, Camden & Islington NHS Foundation Trust, Pennine Care NHS Foundation Trust and NHS Lanarkshire. The recruitment process was as follows: Participants were identified through close liaison with clinical staff based across specialist mental health services (inpatient and outpatient settings) in the NHS Trusts. Settings (and how these were named) varied across sites but included Early Intervention Psychosis Teams, Community Mental Health Teams (CMHTs), Rehabilitation and Recovery teams etc. After clinical staff had confirmed that a potential participant was suitable to be approached (i.e. meets study criteria and no clinical contra-indications), research workers met each potential participant to discuss the study, provide written information and time to consider it, respond to questions and seek written informed consent. Other routes to participation included contact through institutional research registers, or self-referral.

With 642 individuals assessed for eligibility and 345 successfully enrolled and randomized into three parallel study arms, our recruitment strategy was designed to minimize any potential self-selection or other biases. We have thoroughly reviewed our recruitment process and do not identify any self-selection biases or other biases that could impact the findings.

Ethics oversight

The study received ethical approval (Camberwell St. Giles Research Ethics Committee: (20/LO/0657; IRAS (Integrated Research Application System) 277118) and was prospectively registered with the ISRCTN registry at which the published trial protocol (11) and statistical analysis plan can also be accessed (ISRCTN55682735).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

We powered the study to detect plausible effect sizes based on our previous AVATAR therapy trial (8) There we found a clinically meaningful reduction in PSYRATS-AH distress of 4.8 points, with an effect size of approximately d=0.8, but we conservatively reduced this for the current trial, to take into consideration the increase in number of centres, the follow-up comparison (not only end of treatment) and a more pragmatic trial design. We are accounting for two formal comparisons: AV-EXT vs. TAU – plausible effect size 0.6; and AV-BRF vs. TAU – plausible effect size 0.5. The study was powered for an overall treatment effect at a 5% significance level, accounting for 2 multiple comparisons in which the tests are correlated (at r=0.5), giving an alpha level for each test of 0.035. Accordingly, a sample size of 92 per group or 276 in total in the analysis set had 90% power to detect a minimum clinically significant difference (effect size) of 0.5 standard deviations. We sought to recruit 345 participants in total at baseline (87 per site), with n=115 per treatment arm, allowing for conservative attrition rates of 20%.

Data exclusions

The primary analyses were carried out using the intention to treat sample: participants were analysed in the group they are randomised to, and available data from all participants is included, including those who do not complete therapy.

Replication

This clinical trial was not a replication study. However it was designed and delivered using an approach which supports future replication.

Randomization

After baseline assessment, we randomly assigned (1:1:1) eligible participants, via a secure independent web-based service hosted by King's Clinical Trials Unit, using randomly varying sized blocks (3 and 6), stratified by site and baseline voice characterisation (more/less) as defined by meeting the threshold for more highly characterised voices (score>7) on the Voice Characterisation Checklist (27).

Blinding

Research assessors were masked to allocation, and procedures were followed to maintain their masking (assessors did not have access to clinical records after the baseline (pre-randomisation) assessment or access to the therapy database at any stage), all assessments were done at sites remote from the clinic, and participants were reminded before each assessment not to disclose their allocation. It is not possible to mask psychological therapy participants or therapists to their allocation; site co-ordinators were unmasked and informed participants. Therapists were allocated at each site based on availability. Breaks in assessor masking were recorded, and if unmasking occurred, re-allocation to another rater occurred. All primary and key secondary outcomes (PSYRATS-AH scale) were assessed by blinded assessors. Unmasking occurred in 29 people (8.4%) at 16-weeks and 15 people (4.3%) at 28-weeks. All assessments of these people were scored by blinded assessors following these instances of unmasking.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

ISRCTN55682735

Study protocol

The trial protocol is published in the ISRCTN registry with the identifier ISRCTN55682735; the final trial protocol prior to recruitment (V1.2) was also published in Trials journal.

Data collection

Between 1st January 2021 and 30th November 2022, we assessed 642 people for eligibility, recruiting 345 participants. Participants were randomised at four study sites, each recruiting from two mental health service providers, in the United Kingdom (3 England: South London, North London, Manchester; one Scotland: Glasgow) and were randomly allocated to three parallel arms: 116 to AV-BRF, 114 to AV-EXT and 115 to TAU control. Participants were referred by a clinician in the participating clinical sites. Other routes to participation included contact through institutional research registers, or self-referral.

The four UK main University trial sites: the Institute of Psychiatry, Psychology & Neuroscience (King's College London), University College London, the University of Manchester and the University of Glasgow. The four main NHS recruitment sites were South London and Maudsley NHS, North East London NHS Foundation Trust, Greater Manchester Mental Health NHS Foundation Trust and NHS Greater Glasgow & Clyde. The four additional NHS trusts were: Oxleas NHS Foundation Trust, Camden & Islington NHS Foundation Trust, Pennine Care NHS Foundation Trust and NHS Lanarkshire.

Participants were identified through close liaison with clinical staff based across specialist mental health services (inpatient and outpatient settings) in the NHS Trusts. Settings (and how these were named) varied across sites but included Early Intervention Psychosis Teams, Community Mental Health Teams (CMHTs), Rehabilitation and Recovery teams etc.

Outcomes

The pre-specified primary outcome for the study was reduction in distress associated with voices at end of treatment (16weeks) and follow up (28weeks), as measured by the distress dimension of the Psychotic Symptoms Rating Scale (PSYRATS-AH) (5 items, distress (2 items), negative content (2 items) and control. The PSYRATS-AH is a dimensional semi-structured assessor-rated clinical interview assessing auditory hallucinations, comprising in total 11 items, each item scored from 0 (voices not present) to 4.

Key secondary outcomes, as specified in the primary hypotheses, were reductions in the voice frequency scale score (3 items: frequency, duration, and disruption items) and the total severity score (all 11 items) on the PSYRATS-AH scale at 16 and 28 weeks.

Other secondary outcomes were a mix of assessor-rated and self-reported measures, with effects estimated at 16 and 28 weeks.

These included distressing beliefs (PSYRATS-Delusions), Wellbeing (Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS), Psychological recovery (Choice of Outcome in CBT for Psychosis (CHOICE),Fearful attachment (Relationships Questionnaire Item), Voices Action and Acceptance Scales (VAAS), measuring acceptance-based attitudes and actions in relation to voice-hearing experiences, Mood (Depression, Anxiety and Stress Scales (DASS), and Beck Depression Inventory, Anxiety in daily life (using Experience Sampling Measure), Voice power (Voice Power Differential Scale item) and Beliefs about Voices Questionnaire (omnipotence, malevolence and benevolence, total , BAVQ-R), and Trauma-related symptoms (International Trauma Questionnaire) (16 weeks only).

Clinical characteristics of participants were further assessed at baseline with the Clinical Assessment Interview for Negative Symptoms (CAINS) and Scale for Assessment of Positive Symptoms (SAPS). (Further details of all measures are provided in Supplementary Materials)

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A