npj Digital Medicine

https://doi.org/10.1038/s41746-025-02145-5

Article in Press

AI-guided digital intervention with physiological monitoring reduces intrusive memories after experimental trauma

Received: 30 June 2025

Accepted: 1 November 2025

Cite this article as: deBettencourt, M.T., Sakthivel, S., Holmes, E.A. *et al.* AI-guided digital intervention with physiological monitoring reduces intrusive memories after experimental trauma. *npj Digit. Med.* (2025). https://doi.org/10.1038/s41746-025-02145-5

Megan T. deBettencourt, Sruthi Sakthivel, Emily A. Holmes & Mark Chevillet

AI-guided digital intervention with physiological monitoring reduces intrusive memories after experimental trauma

Authors

Megan T. deBettencourt[1*], Sruthi Sakthivel[1], Emily A. Holmes[2,3], Mark Chevillet[1]

Affiliations

1. Ruby Neurotech, Redwood City, CA, USA
2. Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden
3. School of Psychology, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, United Kingdom

* Corresponding author, megan@ruby-neurotech.com

Keywords: Generative AI, large-language models, neurotechnology, pupillometry, PTSD

Abstract

Trauma prevalence is vast globally. Evidence-based digital treatments can help, but most require human guidance. Human guides provide tailored instructions and responsiveness to internal states, but limit scalability. Might generative AI and neurotechnology provide a scalable alternative? Here we provide a first test of ANTIDOTE, combining AI guidance and pupillometry to automatically deliver and monitor the Imagery Competing Task Intervention (ICTI). ICTI is a digital intervention developed by our group to reduce intrusive memories after psychological trauma, previously delivered with human guidance. One hundred healthy volunteers were exposed to videos of traumatic events and randomized to an intervention or active control condition. As predicted, intervention participants reported significantly fewer intrusive memories over the following week. Post-hoc assessments confirmed the AI guide delivered the intervention successfully. Pupil size tracked intervention engagement and was associated with symptom reduction, providing a candidate biomarker. These findings suggest a path towards developing AI-guided digital interventions with scalability potential.

Introduction

Traumatic events are unfortunately highly prevalent. Around 70% of people globally will experience a traumatic event during their lifetime[1]. In the United States (US) the lifetime prevalence of post-traumatic stress disorder (PTSD) has been estimated at 6.8% (standard error = 0.4) in the National Comorbidity Survey Replication[2]. A systematic literature review[3] estimated a lifetime prevalence of PTSD in the US from 3.4% to 26.9% in civilian populations and 7.7% to 17.0% in military populations. In terms of societal costs, for 2018 the annual economic burden of PTSD in the US was estimated at $232.2 billion, or $19,630 per individual with PTSD[4]. Despite this burden, treatment coverage remains limited, although systematic data remain sparse[5]. Data for veterans in the US suggest that only a third have received minimally adequate PTSD care[6]. The World Mental Health surveys across 21 countries have indicated a large treatment gap in that the majority of people (72.4%) with a 12-month anxiety disorder or PTSD do not receive any treatment[7]. Current evidence-based, psychological treatments include prolonged exposure, cognitive behavioural therapy with a trauma focus (CBT-TF), cognitive processing therapy, and eye movement desensitization and reprocessing (EMDR)[8]. However these treatments, even when digitized, can still require multiple sessions with highly trained clinicians[9], which limits the capacity to meet the enormous need. Further, there are ongoing concerns about high dropout from CBT-TF[10]. Current pharmacological treatments lend themselves to wider distribution, but reviews suggest they only have a low effect size for PTSD[8]. Thus there is a critical need for more scalable, efficient, and effective interventions after trauma. Digital therapeutics have emerged as a promising method that could be used for deploying mental health treatments at scale.

One appealing target for digital intervention development is intrusive memories of trauma — involuntary, distressing and sensory memories that repeatedly recur[11]. They are one of the hallmark symptoms of post-traumatic stress disorder (PTSD)[12]. Other symptoms of PTSD include persistent avoidance, negative alterations in cognition or mood, and marked arousal and reactivity. In a nationally representative sample of over 17,000 trauma-exposed adults in the US, 13% had PTSD[13]. This study found the prevalence of distressing memories (defined as recurrent, involuntary and intrusive memories of traumatic events) among those with PTSD was 95%, and 48% among those who were trauma exposed without PTSD. The sensitivity of distressing memories for the diagnosis of PTSD was 95.14% and the specificity was 51.91%. Intrusive memories are also reported by individuals with other diagnoses such as depression or anxiety[14]. Their widespread prevalence and broad clinical relevance make intrusive memories an interesting target for intervention.

To reduce the number of intrusive memories after trauma, our group has developed a digital intervention called ICTI (Imagery Competing Task Intervention) from lab to clinic. Recently, human-guided digital ICTI has been tested in two randomized controlled trials (RCTs) for healthcare staff who had experienced work-related trauma and led to a reduction in intrusive memories[15–17]. In this paper, we define and henceforth limit the term ICTI to refer to those studies using a protocol developed by our group alongside training in how to use that protocol in either lab or clinical settings. ICTI combines a brief memory reminder cue followed by a

demanding visuospatial task. One visuospatial task hypothesized to interfere with perceptual processing involves using mental rotation and mental imagery while playing the visual block puzzle computer game Tetris™. Both in our studies delivering ICTI soon after trauma and those conducted at longer time intervals after trauma have incorporated a brief reminder cue prior to the visuospatial task phase so that the relevant memory is actively held in working memory[18]. When the relevant memory is labile, the concurrent visuospatial working memory demands are presumed to compete with and weaken the perceptual features of the intrusive memory[19–21].

Early versions of the ICTI protocol were initially tested in laboratory settings on reducing intrusive memories in healthy adults using an experimental model of analogue trauma via the trauma film paradigm[22–25]. The trauma film paradigm has been used in numerous other studies to examine related experimental interventions and other approaches with variable effects[26,27]. In the first test of ICTI in a clinical sample, intrusive memories were reduced for the week following the intervention[28]. ICTI has since been shown to be safe and efficacious in reducing the number of intrusive memories one month post-intervention in three clinical RCTs of trauma-exposed individuals, including when delivered soon after trauma with emergency department patients[29] and in digital form when delivered at longer time intervals after trauma[15–17]. Early-stage, small scale studies suggest ICTI may also reduce intrusive memories in patients with PTSD[30,31], but we note these are case studies rather than RCTs and further that one crossover RCT with PTSD patients showed no benefit[32].

The digital forms of ICTI tested in two RCTs for trauma-exposed individuals[15–17] remain limited in scalability due to the dependence on trained human guides for the first session of the digital intervention. The guides provide interactive and personalized verbal instructions as well as monitor participants' non-verbal responses and task engagement. Training of human guides on the ICTI methods protocol has been part of the clinical ICTI RCTs to date[15,16,29]. Such training involves observation and corrective feedback, and can be time consuming.

Removing the reliance on a human guide could allow a more scalable deployment of a psychological treatment such as ICTI, and thus may better meet the needs of trauma-exposed populations globally. Generative AI systems now have sufficient capabilities to instruct participants and assess their comprehension in ways that instructional videos or static text instructions cannot, via interactive and individualized conversations[33,34]. Physiological signals, such as pupillometry (a putative index of cognitive effort)[35–37], are now measurable via lightweight and relatively low-cost devices that can be used outside of controlled laboratory settings[38,39]. This can enable monitoring of internal cognitive states and strategies during an intervention that are otherwise inaccessible via subjective observation or behavioral measures alone[40–44]. Incorporating these advances in generative AI and neurophysiology measures into ICTI could enable a more scalable solution that can both provide individualized instructions about how to perform the intervention (e.g. explaining how to emphasize mental rotation and imagery during gameplay) and observe pupil size during key portions of the protocol (e.g., memory reminder and gameplay) to infer cognitive effort.

We developed an intelligent neurotech prototype ANTIDOTE (AI-guided Neurotherapy for Traumatic Intrusions in a Digital Therapeutic; Fig. **1**) to implement the ICTI in a trauma film paradigm. First, we incorporated generative AI to guide participants through the intervention, delivering structured and interactive instructional conversations. Second, we incorporated physiological monitoring to provide insight into participants' cognitive effort through the intervention. The goal was to develop a unified and automated system as an alternative to the critical roles of instruction and observation played by the human guide. In the current study, we evaluated ANTIDOTE using a widely used experimental model of analogue trauma — the trauma film paradigm[26,45]. Our primary objective was to test whether ANTIDOTE could produce a reduction in the number of intrusive memories, compared to an active control. We also explored the quality of intervention instruction provided by the AI guides and examined whether neurophysiological signals—specifically pupil size—tracked intervention engagement.

Results

Reduction of intrusive memories

The primary hypothesis, which was preregistered (https://doi.org/10.17605/OSF.IO/P56JV), was that participants in the intervention condition would report fewer intrusive memories relative to participants in the active control condition over seven days starting from the day of their in-person study session (see *Preregistration* in *Methods*). We obtained the total number of memory intrusions reported in the electronic log from each participant during the week following the experimental session (mean number of memory intrusions, m=16.31, 95% CIs [12.76, 20.34], n=100). Participants in the intervention condition recorded significantly fewer memory intrusions than participants in the active control condition (*intervention* m_i=11.62, [8.42, 15.56], n_i=50; *control* m_c=21.00, [15.04, 28.02], n_c=50; p=0.01; Cohen's d=0.49; Fig. **2a**). These results confirm the preregistered hypothesis and demonstrate the effectiveness of ANTIDOTE in delivering an automated psychological intervention (ICTI) to reduce intrusive memories after experimental trauma.

To examine how the intervention influenced the trajectory of intrusive memories over time, we conducted an exploratory analysis of the number of intrusive memories reported each day during the 7-day electronic log (Fig. **2b**). We observed significant differences between conditions on several individual days throughout the week following the session (p_0=0.081, p_1=0.117, p_2=0.024, p_3=0.006, p_4=0.047, p_5=0.015, p_6=0.050). To model this trajectory of the change in intrusive memories across days, we fit a mixed-effects linear model with random intercepts for participants (n=100; 700 observations). We observed a significant main effect of condition (β=−1.65, 95% CIs [−2.88, −0.43]; p=0.008), consistent with fewer intrusions throughout the time period in the intervention group compared to the active control group. There was also a significant main effect of day (β=−0.51, [−0.61, −0.40]; p<0.001), but the interaction between day and condition was not reliable (β=0.08, [−0.07,0.23]; p=0.30). This pattern suggests that ANTIDOTE exerted a consistent effect across the 7-day period.

We also conducted an additional exploratory examination of the number of intrusive memories reported at the end of the in-person session via a vigilance-intrusion task. The number of intrusive memories reported on this task correlated with the week-long electronic log (Spearman's ρ=0.23, p=0.02, n=98). There was no reliable difference in the mean number of intrusive memories on the vigilance-intrusion task between groups (intervention m_i=51.10 [41.90, 60.46], n=40; control m_c=48.04, [40.08, 66.46], n=48; p=0.63). Additional details are provided in the Supplementary Results.

Evaluating AI guidance

In contrast with previous ICTI studies where the intervention was led by trained human guides, here an AI guide delivered the instructions through text-based chat conversations with the human participant (see Methods). These human-AI conversations explained each of the key components of the experimental protocol (analogue trauma exposure, intervention condition cognitive task, the concept of intrusive memories, and the rationale and procedure for completing the electronic log) in a structured manner (Fig. **3a**). There was an overall high level of success in AI guidance, as all participants (n=100) completed multiple conversations with the AI guide.

To assess the quality of instructional delivery by the AI guide, we conducted a series of exploratory analyses to evaluate these conversations in four complementary ways: (1) participant survey feedback, (2) human grading of instructional quality, (3) AI-based grading as a scalable alternative, and (4) quality-control analysis of electronic log entries.

First, to understand the participants' experience of using an AI guide, we collected survey data about the experience with the AI guide. In general, participants rated the AI guidance highly (mean rating=4.41 of 5, 95% CIs [4.26, 4.54], n=72). To support responsible AI deployment and in alignment with AI safety guidelines, all conversation logs underwent manual post-hoc review for potentially harmful or offensive content, and no such instances were observed.

Second, two human raters manually graded more than 400 conversations between the AI guide and the human participant to evaluate instructional quality and participant understanding (Fig. **3b**). They applied a scoring rubric, originally developed for training and evaluating human guides on how to lead participants through the digital ICTI[15–17]. Each conversation received a consensus integer score between 0 (lowest) and 6 (highest). Overall, the human grading scored the human-AI conversations at a competent level across participants (mean score, s=4.01, 95% CIs [3.92, 4.10, n=100]; Fig. **3c**). Scores were consistent across the five different human-AI conversations: s_1=4.01, [3.88, 4.14]; s_2=4.00, [3.80, 4.20]; s_3=4.07, [3.91, 4.23]; s_4=3.72, [3.59, 3.87]; s_5=4.27, [4.08, 4.44]. These findings show that the AI guide effectively communicated the instructions.

Furthermore, there was no difference between the score for participants in the intervention (mean score s_i=4.02, [3.88, 4.16], n=50) versus control conditions (mean score s_c=4.00, [3.89, 4.12], n=50; p=0.86). These results indicate that the AI guide reliably delivered instructions with competence and impartiality across conditions.

Third, we examined AI grading of human-AI conversations. Evaluating the quality of human-AI conversations was a time-intensive task that required manual scoring by trained human graders. We therefore assessed whether an AI grader could produce conversation ratings consistent with human evaluations (Fig. **3b**). We provided the AI grader the same rubric as used by the human graders. The AI grader scored the human-AI conversations at an overall similar level as the human graders (mean score, s=4.08, 95% CIs [3.99, 4.17], n=100; Fig. **3c**). There was no reliable difference between the human and AI graders (p=0.28, MAE=0.34, RMSE=0.44). Furthermore, grades assigned by the AI were strongly correlated with human scores across participants (Spearman's ρ=0.52, n=100; p<0.001; Fig. **3c**). These findings suggest that AI-based grading offers a scalable alternative for evaluating the fidelity of AI guidance for the intervention.

Fourth, we also conducted a quality control analysis of the entries in the electronic logs of intrusive memories to assess whether participants demonstrated a clear understanding of the study definition of the intrusive memory (i.e., image-based descriptions of scenes from the videos watched during the experimental session) and how to successfully complete the log from their conversations with the AI guide. We manually reviewed all 1,631 entries submitted across all participants. Entries with blank descriptions or those that did not meet the study's definition of an intrusive memory were excluded, accounting for 8.03% of the entries. In some cases, single entries captured multiple intrusive memories, resulting in a small increase in the total count (0.06% of the entries). The total number of intrusive memories across participants did not reliably change (Δ=1.15 entries, 95% CIs [0.35, 2.15], n=100; p=0.68). The very low rate of modifications indicates that the AI guide successfully conveyed key instructions, enabling participants to understand and complete the electronic log appropriately.

We next assessed whether the reduction in intrusive memories remained after applying data quality control procedures. We still observed reliably fewer intrusive memories in the intervention group (intervention: m_i=10.70, 95% CIs [7.50, 14.68], n_i=50) relative to the control group (m_c=19.62, 95% CIs [14.16, 26.14], n_c=50; p=0.01, Cohen's d=0.49).

Imagery competing cognitive task gameplay and pupillometry

During the cognitive task component of the experimental protocol, participants in the intervention group played a block puzzle game that dynamically varied in difficulty. The game difficulty started at level 1, the slowest and easiest level. When participants successfully cleared a line, the game difficulty increased in a stepwise manner until level 12, the fastest and most difficult level. If the pieces piled up to the top of the game field, the game reset back to level 1. Thus, each participant experienced an individualized trajectory contingent to their game play.

A key aspect of the intervention task is that participants are instructed to engage in mental rotation and imagery during gameplay. We conducted exploratory analyses to examine how neurophysiological measures, specifically pupil size, a putative signature of cognitive effort[35], track these internal mental states during gameplay. We compared pupil size during the cognitive task (intervention or control, each 15 minutes) versus the 10-minute rest period which occurred

after watching the videos. During the intervention cognitive task (i.e., mental rotation during gameplay), the average pupil size was larger than during rest (mean difference Δ=0.46, 95% CIs [0.38, 0.54], n=47; p<0.001; Fig. **4a**). During the control cognitive task (i.e., listening to the podcast), the difference in the average pupil size from rest was trending but not reliable (mean difference Δ=0.05, 95% CIs [0.00, 0.10], n=49; p=0.07). The interaction between groups was reliable (p<0.001).

To more directly link pupil size and cognitive effort, we leveraged the simultaneous dynamics of the game difficulty (Fig. **4b**). For each game piece that fell for every participant, we calculated the difficulty level as well as the mean pupil size. We fit a linear mixed-effects model to examine the relationship between difficulty level and pupil size (n=48 participants; 7956 total pieces). Pupil size differences were de-meaned within participants, and both variables (difficulty level and pupil size) were standardized. We included participants as a random effect, with varying intercepts and slopes. There was a reliably positive relationship between the difficulty level and pupil size (β=0.26, 95% CIs [0.12, 0.39]; p<0.001; Fig. **4c**). That is, pupil size increased with increasing game difficulty.

Memory reminder behavior and pupillometry

A critical component of ANTIDOTE (and the ICTI intervention) is a memory reminder, when participants are instructed to briefly list their "worst moments" that they remember from the film, prior to the cognitive task. In this paradigm, participants were asked to provide brief written descriptions of the key moments that they found most distressing in the videos. Participants listed a variable number of entries (mean number of entries #=5.91, 95% CI=[5.48, 6.35], n=100). A manual review confirmed that all 591 entries (100%) were related to the video content. There was no significant difference between the number of entries for the control vs. intervention participants ($\#_i$=6.02 [5.40, 6.66]; $\#_c$=5.80, [5.22, 6.38]; p=0.66). That is, participants were successful at recalling distressing moments from the films, and the behavioral measure of memory (i.e., number of moments recalled) did not differ between the conditions.

In addition to the memory behavior, we were interested in the internal memory state, which we assessed via pupil size (Fig. **5a**). We conducted exploratory analyses, examining whether the memory reminder engaged cognitive effort, indicated by a larger pupil size. We compared pupil size versus the 10-minute rest period which occurred after watching the videos and prior to the memory reminder. Indeed, pupil size was larger during the memory reminder versus rest, for both participants in the control group (mean difference Δ=0.28, 95% CIs [0.21, 0.36], n=49; p<0.001) and the intervention group (mean difference Δ=0.24, 95% CIs [0.17, 0.31], n=45; p<0.001). Consistent with the fact that the memory reminder occurred before the intervention and control groups diverged, there was no significant difference in pupil size between groups (p=0.41).

We also examined the pupil dynamics during active memory recall, specifically examining the time between the time at which the reminder screen initially appeared and the time at which the first text entry was submitted. The duration of the entire memory reminder period varied across participants, based on factors including the latency of memory recall, number of entries, and

typing speed. We aligned the pupil size data from all participants to the onset of the memory reminder screen (Fig. **5b)**. To ensure consistent data length, we truncated each trace to the shortest duration of the memory reminder period (*t*=29 sec) across all participants with available data (*n*=93). At the onset of the memory reminder screen, there was no reliable difference from baseline (*p*=0.79 at t=0 sec). Following a brief initial dip, pupil size rose reliably above baseline (*p*=0.006 at *t*=2.05 sec after the reminder screen appeared) and remained elevated.

Participants provided a variable number of entries during the memory reminder phase (between 1 and 15), modeled after clinical implementations of ICTI for patients with PTSD[16,17]. Beyond examining the dynamics up until the first entry, we also assessed the effect of subsequent entries. We conducted a linear mixed-effects model relating entry number (2 and above, as the time period up until the first entry involved the ramp up from memory reminder onset, see Fig. **5b**) to mean pupil size during the entry. The model included all participants with available pupil data from eligible entries (*n*=88 participants, 477 entries total) and incorporated random intercepts and slopes to account for within-subject variability. Both the entry index and pupil size were standardized prior to modeling. As the entry number increased, the pupil size decreased (β=−0.18, 95% CI [−0.29, −0.06]; *p*=0.002; Fig. **5c**). According to our interpretation of pupil size as a putative index of cognitive effort, these results suggest that less cognitive effort was expended when reporting later entries.

Physiological predictors of intervention success

To assess whether our neurophysiological markers of cognitive effort were related to the success of the intervention, we conducted additional exploratory analyses investigating whether there was a relationship between pupil size and the number of intrusive memories for two key phases of the experimental session: the cognitive task (either gameplay or listening) and the memory reminder.

First, we analyzed participants in both the intervention and control groups combined, investigating the cognitive task phase. We explored whether greater cognitive effort, indexed by larger pupil size during the cognitive tasks (either gameplay or listening), was associated with fewer intrusive memories. Pupil size during the cognitive task was negatively correlated with the number of intrusive memories (Spearman's ρ=−0.31, *n*=96; *p*=0.002). This relationship was further quantified by a linear regression of pupil size during the cognitive task predicting the number of intrusive memories (β=−17.73, [−30.27, −5.19], *n*=96; *p*=0.007; Fig. **6a**). That is, our measure of greater cognitive effort during the cognitive task (i.e., measured during the experimental session) was associated with fewer intrusive memories in the real world over the next week. When examined separately within intervention and control groups, these associations were not reliably predictive (control: β=−28.12, [−63.41, 7.17], *n*=49; *p*=0.12; intervention: β=−6.42, [−20.29, 7.45], *n*=47; *p*=0.36).

Next, we included pupil size measured during the memory reminder period as an additional predictor for the number of intrusive memories. Specifically, we fit a linear model (in both

intervention and control groups) in which pupil size during both the cognitive task and during the memory reminder period predicted the number of intrusive memories. This allowed us to assess the unique contributions of task engagement and memory recall when considered simultaneously. The modeling results revealed that while the influence of the cognitive task remained a significant predictor of the number of intrusive memories (β=−18.51, [−32.89, −4.13], n=94; p=0.01), the cognitive effort measured during memory reminder phase was not reliably predictive (β=0.75, [−18.58, 20.08], n=94; p=0.94).

Finally, we specifically investigated the participants within the intervention group. The previous analyses examined participants in both the intervention and control groups. We repeated these analyses, restricted to just the participants in the intervention condition. The cognitive task effect coefficient replicated the effect found in the full sample, that a larger pupil size during the cognitive task (here gameplay) predicted fewer intrusions (β=−28.41, [−52.33, −4.49], n=45; p=0.02; Fig. **6b**). Whereas, the coefficient for memory reminder phase was reliably positive — a larger pupil size predicted more memory intrusions (β=33.21, [2.14, 64.28], n=45; p=0.04; Fig. **6b**). That is, both pupil size during the memory reminder and during the mental rotation gameplay task predicted the intervention success, albeit in different directions. This suggests a conceptual model where the ideal approach may be to expend low cognitive effort during memory recall, followed by high cognitive effort during the intervention (Fig. **6c**).

Discussion

We investigated whether combining advances in generative AI and neurotechnology could allow effective delivery of an emerging evidence-based digital mental health treatment in a controlled experimental model of trauma to reduce intrusive memories, and thus enable future scalability. We developed and tested ANTIDOTE, an intelligent neurotech prototype, which combined three key elements: (1) an evidence-based digital treatment for intrusive memories, the Imagery Competing Task Intervention (ICTI) developed from our group, (2) an AI guide to provide interactive instruction and assess participant comprehension, and (3) pupillometry to monitor cognitive effort during key phases of the intervention. We conducted a randomized controlled experimental study to evaluate whether ANTIDOTE would reduce intrusive memories reported by healthy participants after viewing videos of traumatic events. As hypothesized and preregistered, participants in the intervention group reported significantly fewer intrusive memories of experimental trauma over the following week compared to an active control group. This finding is notable given conditions were well matched, as each included a memory reminder phase and differed primarily in the type of task (visuospatial versus auditory) that followed.

We also conducted a series of exploratory analyses to examine how the AI guides and neurophysiological monitoring supported core functions previously fulfilled by human guides when administering the digital form of the ICTI intervention. In our prior work on ICTI, human guides were extensively trained through instructional sessions, supervision, and corrective

feedback, and their competency was both trained and evaluated using a rubric to help standardize ICTI delivery across human guides and reduce variability within and between studies. A key motivation for our development of an AI guide was to provide standardized delivery without the variability introduced by human guides and the need for time consuming human training. The AI guides successfully delivered individualized instruction, as scored on a clinical rubric first by human graders and next by AI graders, albeit not to the top score of the rubric. Additional evidence of AI instructional effectiveness included favorable participant survey feedback and successful completion of electronic logs containing valid intrusive memory entries. Pupillometry, used to monitor cognitive states during the intervention, provided objective insight into the cognitive effort required during key phases (i.e., memory reminder and gameplay with mental rotation) and were associated with intervention outcomes.

There are several possible advantages to the future scalability of digital mental health interventions through the use of generative AI tools[46], specifically large language models (LLMs), to deliver evidence-based intervention protocols. In our study, the AI guide delivered interactive and individualized instructions to participants, providing a consistent and standardized framework for administering an experimental version of this digital intervention. This approach offers advantages over unblinded human guides, particularly in improving instructional consistency, increasing methodological rigor, and reducing bias. Importantly, the AI guide was unaware of the existence of different treatment groups across participants and was even blinded to condition assignment during different conversations with the same participant. Future research with the trauma film paradigm might use an AI guide to explore alternative hypotheses such as around different protocols, intervention components or control conditions.

Here, our AI guide as an instructional interface neither prompted for nor required disclosure of sensitive personal or health information. This was achieved by careful engineering of the system prompt to constrain the AI guide's function, without therapeutic or diagnostic intent. To further protect participant privacy, there was no involvement of the AI during the portions of the protocol where participants might be most likely to disclose personally identifiable or sensitive information, i.e., details of the intrusive memory symptom. Therefore, for both the memory reminder and as participants completed the electronic log of intrusive memories over the following week, the protocol deliberately used static instructions and webforms, rather than AI-guided conversations.

Alongside these advantages, there are also specific limitations and potential future improvements of our use of AI. First, although the guide was not designed to solicit personal disclosures, participants were not explicitly prevented from sharing sensitive information. Future systems could incorporate additional safeguards, such as the use of retrieval augmented generation or automated content moderation, to prevent the risk of unintended disclosures [47–49]. Second, AI guidance was not used in the instructions for the podcast listening task (though AI guidance was in all other sections of the control condition), which may have introduced an imbalance in instructional engagement between conditions for the task. Future work should consider adding an instructional conversation to the task component of the control condition to better balance the use of AI across groups.

Third, while all participants in the current study interacted effectively with the AI guide, future iterations could improve instructional quality and enhance accessibility for diverse populations and individuals with lower digital literacy. Although exploratory analyses that graded the human-AI conversations based on a clinical rubric indicated the overall competency of the AI guides (an average score of approximately 4), they did not reach the top score of the rubric ideally expected of trained human guides[15–17]. Future improvements could reduce the pedantry of the AI guide to increase tolerance for paraphrased input (and discourage parroting or direct copying of the AI's instructions by the participants) and encourage higher-level responses to support comprehension.

A notable innovation of ANTIDOTE is that it incorporated neurotechnology to observe participants during the intervention. In this study, pupil size provided insight into internal cognitive states hypothesized to relate to cognitive effort, such as mental rotation gameplay and trauma memory recall that might otherwise be inaccessible through behavior alone [40,50]. If human guidance is not available, physiological monitoring may in the future provide a sensitive metric for treatment compliance. Like motion capture in physical therapy, it can go beyond self-report of intervention completion to confirming that participants executed the intervention as intended and potentially enhancing intervention success[51]. Pupillometry results, albeit preliminary, also suggested a conceptual model for the various cognitive mechanisms underlying intervention success: low levels of memory recall effort during the memory reminder phase[20,52,53], followed by cognitively effortful gameplay involving mental rotation and mental imagery[54–56]. While this relationship between intrusive memories and pupillometry was exploratory and may be impacted by analytic decisions (e.g., choice of whether pupil size is normalized against the 3-minute rest period used here or some other period, or whether the cognitive task and memory reminder periods are analyzed individually or jointly) or sample characteristics, future work could aim to test replication, as well as fractionate the cognitive effort or cognitive tasks into subcomponents, while also considering other factors known to influence pupil size, such as emotional arousal and luminance(Pan et al. 2022; Pan et al. 2024).

While the lighting in the room was held constant during and across sessions, we did not control for low-level visual features on the experimental display that can influence pupil size, such as luminance differences between the tasks used in the intervention and control groups or luminance fluctuations within the gameplay. This choice was made to preserve consistency with prior studies and to maintain the visual design of the intervention, which may be important for engagement. However, if pupil size primarily reflected visual features rather than cognitive effort, this would likely have weakened rather than strengthened the observed relationships with task difficulty and intervention outcomes—suggesting that the pupillometry signal likely retained meaningful cognitive information despite any potential luminance confounds.

Future development to further scale the neurophysiological monitoring approach will require more accessible and flexible hardware and software. Although the screen-mounted eye tracker used here was portable, relatively low-cost, and did not require a chin rest, it still limits scalability as it is specialized hardware not typically integrated into standard consumer devices.

Emerging methods for eye-tracking using standard webcams or smartphone cameras may offer scalable alternatives in the future[57,58]. These advances could also enable deployment in more naturalistic settings, while still maintaining measurement fidelity across diverse populations. Finally, although pupil size in this study was monitored in real time, analyses were conducted post hoc. The observed relationships between pupil size at the group and individual level motivate the design of intelligent closed-loop systems that adapt dynamically to optimize cognitive engagement[59–61], by adjusting difficulty during the gameplay or the number of memories listed during the memory reminder phase, to potentially further improve intervention outcomes.

Our findings provide an initial proof-of-principle that an AI-guided and physiologically-monitored digital implementation of an evidence-based human-guided digital treatment (ICTI) can reduce the number of intrusive memories in healthy participants after exposure to analogue trauma. The magnitude of this reduction—approximately 45%—is similar to previous human-guided ICTI laboratory studies using a similar trauma film paradigm (54%[62], 52% Experiment 1[25], and 70% Experiment 2[25]). While the trauma film paradigm remains a preclinical model, it offers utility for intervention development prior to conducting clinical studies due to the opportunity for strong experimental control[45]. Additionally, some interventions developed using this experimental trauma model have now been tested for intrusive memories after real-world trauma, including human-guided ICTI. Compared to ICTI lab studies, similar reductions in mean number of intrusive memories were observed in some clinical studies of ICTI from our group (e.g. 62% versus active control at 1 week[28], 48% versus active control at 1 week[29], and 68% versus waitlist control at 4 weeks[17]). Taken together, this evidence motivates continued pre-clinical development to test whether future iterations of the approach taken by ANTIDOTE can be extended. If successful, future steps could include tests with real-world trauma populations and aim to reduce the reliance on trained human guides. Many questions remain that could be explored using the trauma film paradigm that have translational interest. For example, while our current analyses focused on the total number of intrusive memories and did not link intrusions to specific scenes in the analogue trauma film, future studies could examine whether intervention effects are related to the specific film scene recalled prior to the visuospatial task.

Beyond the encouraging empirical findings, ANTIDOTE also reflects two emerging directions in digital mental health care. First, the use of LLMs to deliver structured guidance within evidence-based protocols parallels growing efforts to incorporate the use of LLMs in medicine[34] and to define and implement the role of digital navigators—human technology coaches increasingly integrated into clinical care teams[63]. Second, the inclusion of real-time physiological monitoring aligns with increasing interest in integrating objective measures—such as digital phenotyping[64] and digital biomarkers[65]—into mental health care and other areas of clinical care where insight into internal state is desirable (e.g. pain)[66]. While our use of pupillometry was exploratory, it illustrates how physiological signals might eventually support intervention fidelity, cognitive engagement tracking, or personalization of treatment delivery. These trends, though still early in clinical adoption, are beginning to inform implementation models and signal a shift toward more responsive, data-informed approaches in digital mental health[67].

This study presents initial evidence that a fully automated, AI-guided digital intervention can conceptually replicate the intrusion reduction effects of a human-guided intervention after trauma (i.e., ICTI) in a controlled experimental model of trauma. The use of trained human guides may pose a bottleneck to treatment scalability. By demonstrating that both instruction and engagement monitoring can be delivered through AI and without human involvement in an experimental model of trauma, ANTIDOTE represents a meaningful step towards developing scalable, low-cost mental health care. As the use of AI tools and digital phenotyping gain traction in clinical care, developing approaches like ANTIDOTE that operationalize these concepts in structured, evidence-based interventions could help close the gap between research and real-world impact. Continued development and clinical validation will be critical to determine whether such systems can extend access to effective care for the millions affected by trauma worldwide.

Methods

Experimental protocol overview

This study developed and evaluated ANTIDOTE, an AI-guided digital neurotech intervention to reduce intrusive memories after exposure to an experimental model of trauma. The study was designed as a digital experimental medicine implementation of the Imagery Competing Task Intervention (ICTI) to reduce intrusive memories developed by our group[15,16] and formerly reliant on trained human guides. ICTI includes a memory reminder and cognitive task (including mental rotation and imagery during computer game play).

Using a between-subjects design, participants were randomly assigned to either an intervention or active control condition. Both the intervention and active control groups underwent similar procedures, receiving standardized instructions by an AI guide while physiological measures were continuously recorded. Both groups watched a film containing traumatic content, followed by a brief rest period, and a memory reminder designed to briefly orient the participant to hotspots in the film. The key experimental manipulation occurred during the next phase, the cognitive task. Participants in the intervention condition completed a visuospatial block puzzle game, emphasizing visual imagery and the use of mental rotation while playing the game. Participants in the active control condition completed an auditory task, listening to a podcast about classical music unlikely to engage visual imagery. Following the in-person experimental session, all participants used their own personal electronic devices (e.g., smartphone or computer) to remotely log their intrusive memories to the trauma film over the following week.

Participants

One-hundred participants (55 female, 42 male, 3 other/declined to provide sex; mean age = 41.56 years, SD = 14.25, 1 declined to provide age, range 18-65 years) were recruited from the San Francisco Bay Area community via targeted electronic and physical advertisements. In terms of self-identified ethnicity, 34% identified as Asian or Asian American, 7% as Black or African American, 7% as Hispanic or Latino, 41% as White or European American, 8% as Multiracial, 1% as Other, and 2% preferred not to respond. Participants' educational attainment was as follows: 23% reported a high school diploma or GED, 14% an associate degree, 33% a bachelor's degree, 21% a master's degree, 1% a doctoral degree, 5% a professional degree beyond bachelor's (e.g., JD, MD, PsyD), and 3% preferred not to respond. This sample size meets and exceeds the preregistered target of 80 ("More than 80 complete datasets may be collected if time and circumstances allow"). Two additional participants started the study but voluntarily disenrolled prior to completion. Inclusion criteria were (a) aged between 18 and 65 years old, (b) English fluency, (c) access to an internet-enabled smartphone or computer, (d) not having previously participated in similar studies, and (e) no self-reported recent or planned stress-inducing or traumatic experiences during the week of study participation. Data collection started in June 2024 and was completed in October 2024. All participants provided their written consent after being informed that the study involved watching emotionally distressing video content and would include both physiological and behavioral measurement. Ethical approval for the study was granted by the Advarra Institutional Review Board (Protocol Reference ID Pro00073795).

Preregistration

The study was preregistered on the Open Science Framework (OSF) using the template from AsPredicted.org prior to any data collection (https://doi.org/10.17605/OSF.IO/P56JV). The primary hypothesis, which was preregistered, was that "*participants receiving the intervention will report fewer intrusive memories relative to participants who receive a control task*". The preregistered analysis was to "*compare the total number of intrusive memories reported by participants in the Intervention condition versus those in the Control condition via their entries in the electronic diary over seven days starting from the day of their in-person study session.*" This is the only preregistered hypothesis and analysis; all other analyses reported are exploratory. The preregistration also included a series of exploratory analyses, many of which are beyond the scope of this paper. However, we report a subset of exploratory analyses, including whether pupil size differed between groups and whether it predicted the number of intrusive memories reported.

Condition assignment

Each participant was randomly and independently allocated to either the intervention (*n*=50) or active control (*n*=50) condition with equal probability. Note, the 50/50 split between conditions occurred by chance, as no stratification or balancing was applied. Randomization was implemented using Python-based random number generation upon each participant's arrival for the experimental session. All condition-specific instructions were standardized, delivered by the static text within the web-based software platform and the AI guide (see details on how the AI guide was blinded to condition below). This ensured that instructions were consistent across conditions, except for condition-specific instructions related to the cognitive tasks for the intervention or active control group. The AI was unaware of group assignment or even the existence of multiple groups. To help maintain participant blinding, the informed consent form stated that participants would be randomly assigned to one of two cognitive tasks (intervention or control) but not what the cognitive tasks were. After randomization participants received condition-specific task instructions from the AI-guide, but were not told whether the task was intervention or control in order to maintain blinding. The recruiting materials did not include any depiction of the cognitive tasks.

AI guide

A central feature of this study was the use of an AI guide in place of the human guidance used in our prior ICTI studies. The AI guide, implemented as a structured chatbot, engaged participants in a structured, multi-turn, multi-phase conversation using only a custom prompt provided to OpenAI's GPT-4 model (i.e., without any fine tuning or retrieval-augmented generation). Each instructional conversation began with a short segment explaining the upcoming task, followed by asking the participant to summarize the instructions in their own words. The AI compared the summary to the original instructions and provided corrective feedback if key points were missing and asked participants to revise their response. Once the summary was deemed complete, the guide proceeded to the next segment.

Participants could ask questions at any point, which the AI would answer before resuming the instructional sequence. Five instructional conversations were interleaved throughout the experimental session, each corresponding to a different phase of the protocol:

(1) Analogue trauma exposure film viewing: Before viewing the video clips, participants completed the first instructional conversation with the AI guide, which instructed participants to immerse themselves in the scenes, and imagine the events happening to themselves or someone they care about. (2) Intervention condition cognitive task: Prior to gameplay, the AI guide instructed participants on how to control the game and the cognitive strategies participants should use. Participants were asked to focus on using mental rotation and mental imagery to imagine different placements of each piece rather than maximizing their score. (3) The concept and definition of intrusive memories: After the cognitive task, participants received instructions on intrusive memories being visual images from the film that might unintentionally pop back into their mind. (4) The rationale for intrusive memories log: why keeping the intrusive memory log was important to the study and (5) The procedure for logging intrusive memories: The participant watched audiovisual tutorials on how to log intrusive memories in the diary using their personal electronic device.

The AI guide was blinded to condition assignment: prompt content was identical across groups, except for the task-specific prompt for the intervention phase. The AI guide retained memory within each conversation to allow for coherent interaction, but had no access to information from previous conversations with the same participant. There was also no access to conversations with other participants. This preserved blinding of the AI guide across all other conversations in the protocol.

Experiment Protocol

The experimental protocol consisted of a baseline period, film viewing, rest, memory reminder, cognitive task (intervention condition or active control condition), vigilance-intrusion task, and intrusive memory reporting.

Baseline

At the start of the study, participants completed a 3-minute baseline rest period, during which a fixation cross was displayed on the screen. They were instructed to sit quietly and let their mind wander, but not to close their eyes for an extended period of time or fall asleep. This rest period provided a baseline measure of pupil size across individuals.

Film viewing

All participants viewed a compilation of 10 video clips of a distressing nature (approximately 11.5 minutes total duration). The videos included depictions of actual or threatened death and serious injury. Films included public service films about car accidents in the context of importance of wearing a seatbelt, dangers of drinking alcohol and driving, about risks of drowning when swimming after drinking alcohol,footage concerning war, and an animal on the rampage, and medical procedures (laser eye surgery and open leg fracture), eight clips of which had been used previously[25,62]. Before viewing the video clips, participants completed the first instructional conversation with the AI guide, which instructed participants to immerse themselves in the scenes, and imagine the events happening to themselves or someone they care about. Participants rated their sadness, depression and hopelessness on a 10-point scale before and after watching the video clips to check for mood change.

Rest

Following the video, all participants completed a 10-minute rest period, during which a fixation cross was displayed on the screen. They were instructed to sit quietly and let their mind wander, but not to close their eyes for an extended period of time or fall asleep. This rest period was comparable with previous work[25] and here also provided a point of comparison for analyzing physiological signatures of cognitive effort during the memory reminder and cognitive task.

Memory reminder

All participants were instructed to recall their "worst moments" from the video. These memories have been referred to as hotspots and associated with future intrusive memories[68,69]. To enhance AI safety, AI tools were intentionally not used in this task, as it is the one portion within ICTI where participants are asked to disclose potentially sensitive information (i.e. their brief descriptions of intrusive memories). Instead, participants were instructed via static text on the screen to picture the scenes that stood out in their mind and briefly describe the visual details (5-7 words). Participants typed brief descriptions of each scene into a list of entries, so that we could verify whether they had successfully retrieved memories from the videos. Participants could provide a variable number of entries. This design was modeled after clinical implementations of ICTI for trauma-exposed individuals[17] and intended to enhance ecological validity and enhance the translational relevance of ANTIDOTE.

Cognitive task

Participants then completed a 15-minute cognitive task, which differed by condition.

Intervention condition

Participants in the intervention condition completed an imagery competing task. Similar to previous studies, this task involved playing a visual falling-block puzzle game (a game genre popularized by Tetris). Prior to gameplay, the AI guide instructed participants on how to control the game and the cognitive strategies participants should use. Participants were asked to focus on mental rotation and imagining different placements of each piece rather than maximizing their score. The game layout consisted of the game field on the left, with the upcoming three pieces displayed on the upper right, the current game level in the middle right, and the game score in the lower right. Participants controlled the pieces using the arrow keys: left/right arrows to move the pieces horizontally, the up arrow to rotate them, and the down arrow to accelerate their descent. Gameplay lasted 15 minutes, with difficulty (i.e., block drop speed) increasing after each cleared line and resetting when the blocks filled the game field. The software for the game was adapted from publicly available open-source code (https://github.com/mpirescarvalho/react-tetris, MIT license). This approach provides precise logging of game states and participant behavior, synchronized with physiological recordings.

Active control condition

Participants in the control group listened to the first 15 minutes of an episode[70] of Fresh Air by the US National Public Radio about classical piano. This auditory task was selected to provide neutral, non-aversive content that did not rely heavily on visual imagery nor mental rotation. Instructions for this task were simply to listen to the podcast, and were delivered as static text

on the webpage rather than by an instructional conversation with the AI guide. This control condition was selected as a structured, standardized digital task of the type that we have used in our prior clinical research with ICTI[15,29]. The podcast task was developed to control for digital delivery, expectation effects, and attention demands, while avoiding visuospatial elements such as mental imagery and mental rotation[15,29].

Vigilance-intrusion task

After the cognitive task, all participants (n=100) received instructions from the AI guide explaining the concept of a visual intrusive memory. Following this, most participants (n=98 of 100; 2 excluded due to time constraints) completed a vigilance-intrusion task similar to prior work[25]. This task provided an opportunity to report intrusive memories during the experimental session by combining a vigilance task (a sustained attention to response task, SART) with concurrent intrusion reporting. Task-specific instructions were provided as static text. During the task, numbers (0-9) appeared on the screen for 250 ms, followed by a fixation cross for 1500 ms. For 33% of the trials, a blurred still from a video appeared behind the number. In total, participants completed 270 trials. For the vigilance component (i.e., SART), participants were instructed to press the "j" key in response to every digit except the number 3, which occurred on 10% of the trials. Thus, correctly responding to the number 3 required inhibiting the prepotent response. To report an intrusive memory, participants were instructed to press the "f" key. Unlike previous implementations of this task[25], participants were not asked to provide written descriptions of each intrusion.

Intrusive memory reporting

At the end of the experimental session, participants completed two final AI-guided instruction conversations: (1) why keeping the intrusive memory log was important to the study, and (2) how to complete the intrusive memory log. For the next seven days, participants logged any intrusive memories of scenes in the videos they experienced by making entries in a Google Sheets spreadsheet using their personal smartphone or computer. The spreadsheet contained multiple tabs: Intro, Example, and then tabs for each day (1-7). The "Intro" tab contained excerpts from the AI guide instruction conversation on intrusion reporting, including the importance of keeping an accurate record for the study, a description of visual intrusions, and instructions for how to complete the electronic log. The Example tab was completed by participants during the AI guide instruction conversation. Daily email reminders linked to the relevant tab for each day. If no intrusions occurred, participants selected "No Intrusions." Otherwise, they selected "Visual Intrusion" and typed a short description in the adjacent column. Multiple intrusions were entered as separate rows. Intrusions were counted individually even if repeated. The total number of intrusive memories across all days was used as the primary outcome measure.

Apparatus

Participants completed the in-person experimental session alone in a quiet testing room within a converted office building with low ambient lighting. The experimental protocol was presented in a Google Chrome web browser that was displayed on the full screen of an external LCD monitor (15.6 inches; 2560x1440 resolution). Participants used an external keyboard and mouse. The

experiment was implemented as a JavaScript-based React application using jsPsych (version 7.3.3) alongside custom-built components for the memory reminder and block puzzle game. This React app was hosted on a local server accessed by the client testing machine.

Physiological data

Physiological data were collected continuously throughout the experimental session. Data were time-locked to the experimental protocol via WebSocket communication between the front-end React app and the client machine.

Pupil size and eye gaze were recorded using a Tobii Pro Spark eyetracker (60 Hz sampling rate) mounted directly below the LCD monitor. Participants completed a 5-point eyetracker calibration and validation procedure. Due to technical issues, calibration data were not recorded for 2 participants, and eye-tracking data were not recorded for 1 participant. The mean viewing distance during calibration was 68.80 cm (95% CIs [67.34, 70.24], n=97). Data analysis of pupillometry data is described in the *Pupillometry Analysis* section below. No chin rest was used.

Cardiovascular data were recorded using a pulse oximetry ear clip sensor (Nonin Xpod 8000Q2, 75 Hz sampling rate) placed on the participant's left ear lobe. These data were collected to explore heart rate and heart rate variability (HRV) as indicators of cognitive load, analogous to pupil size. However, analysis of these data is beyond the scope of the current paper.

Video recordings were captured using a USB webcam (Logitech Brio, 30 Hz sampling rate) to enable post-hoc assessment of participant compliance with the protocol, since there was no human experimenter in the room with the participant. For example, behaviors such as falling asleep.

Analysis of the total number of intrusive memories

The primary hypothesis, which was preregistered, was that participants in the intervention condition would report fewer intrusive memories than those in the control condition. The primary outcome measure was the total number of intrusive memories recorded by each participant in electronic logs completed during the seven days following the experimental session.

Following an intention-to-treat approach, we first tested this hypothesis using the raw total number of intrusive memory entries from the electronic log from all participants (n=100). This is consistent with the preregistration, which states "The primary analysis will compare the total number of intrusive memories reported by participants in the Intervention condition versus those in the Control condition via their entries in the electronic diary over seven days starting from the day of their in-person study session."

We conducted a between-groups comparison of the total number of intrusive memories. As the data did not meet the assumption of normality, we used a non-parametric test (see *Statistics*). Although the preregistered hypothesis was directional, we present two-tailed comparisons in the Results section. For completeness, we report both parametric results in the Supplementary Results.

To assess the robustness of our findings, we conducted two additional exploratory analyses, which were not preregistered. First, we evaluated the results after applying quality control procedures to all intrusive memory entries (see *Evaluating AI guidance* below). Second,

we evaluated the results excluding any participants who deviated from the protocol or were outliers (see *Supplementary Results*). For both of these analyses, there were reliably fewer intrusive memories in the intervention versus active control groups.

Temporal pattern of intrusive memories

To examine the trajectory of intrusive memories across days, we analyzed between-group differences in the number of the intrusive memories reported per day during the 7-day electronic log. We also modeled these data using a linear mixed-effects model including Day (0-6) and Condition (Intervention or Control) as fixed effects, and participant as a random effect to account for repeated measures.

Evaluating AI guidance

All participants completed instructional conversations with the AI guide, which were assessed in 4 ways: (1) self reported surveys from the participants, (2) human grading of the human-AI conversations according to a rubric, (3) AI grading of the human-AI conversations according to the same rubric, and (4) quality control analysis of electronic log entries.

First, time permitting, most participants (n=72) completed a brief survey at the end of their experimental session. Participants rated several statements on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Four statements focused on their experience with the AI guide: “The AI chatbot provided instructions that were easy to understand”, “The AI chatbot was easy to interact with”, “The AI chatbot did a good job ensuring I understood the instructions”, and “The AI chatbot produced unexpected or inappropriate content.”. The final statement was reverse-scored so that, for all items, higher values consistently reflected a more positive experience with the AI guide. Two additional statements assessed their general experience during the session (“I felt physically comfortable throughout the study session.”, “The software ran smoothly without any apparent bugs.”) and were not included in the analysis of the AI-specific ratings.

Second, every completed human-AI conversation with each participant was evaluated by human grading. Two human graders applied a rubric closely modeled from one that had previously been used to train human guides to administer ICTI[15–17]. Note that the rubric was not used in the development of the AI prompts, but was used for exploratory post-hoc analysis of the five instructional conversations. Human graders were blind to participants’ condition assignment for four of the five instructional conversations (analogue trauma exposure film viewing, the concept and definition of intrusive memories, rationale for intrusive memories log, and procedure for logging intrusive memories). For the intervention condition cognitive task, blinding was not possible because the content explicitly referred to the computer game. To minimize any potential bias from this partial blinding, all conversations of the same type were graded together. The two human graders each reviewed each conversation and agreed upon a final consensus score for each conversation. Participant ratings were obtained by averaging the scores from all of their conversations.

The original rubric was adapted from the revised cognitive therapy scale that is used in the training of (human) cognitive behavioural therapists[71] and incorporated the Dreyfus system for denoting competence[72]. The rubric rated the conversations on a seven-point Likert scale, ranging from 0 (absence: no explanation given to participants or explanation is

incomprehensible), 1 (major problems: key elements are omitted), 2 (novice: lacking detail or clarity, difficult to understand), 3 (advanced beginner: interferes with user understanding, overly strict or overly permissive), 4 (competent: minor problems, slightly strict or slightly permissive), 5 (proficient: accurate with minimal issues), and 6 (excellence: accurate and efficient explanation even in the face of participant difficulties).

Third, we investigated whether an AI model could reliably grade human-AI conversations in a manner consistent with human raters, when provided with the same chat logs and grading rubric. The AI was instructed to evaluate each conversation using the same rubric as the human graders (Open AI, model version gpt-4o). The model was prompted with the grading rubric along with the conversation text, and instructed to assign a numeric score (0-6) along with a brief justification for the rating. Each conversation was evaluated independently, and the prompt remained fixed across all conversations. Participant ratings were obtained by averaging the scores from all of their conversations.

Some participants (n=5 of 100) did not complete the fifth and final chat (i.e., the procedure for logging intrusive memories) due to time limitations (n=4) or lack of engagement and falling asleep (n=1). Two of these participants started but did not complete the fourth chat (i.e., the rationale for logging intrusive memories). All incomplete and missing conversations were excluded from scoring by both the human raters and the AI.

Fourth, we also conducted an exploratory post-hoc systematic quality control review of all entries in the electronic log of intrusive memories. Each entry was reviewed in accordance with the study's definition of an intrusive memory according to prior work[25,62], which had three requirements: (1) image-based descriptions of scenes (2) from the videos watched during the experimental session that (3) unintentionally popped into mind. The number of intrusive memories for a participant decreased if the description of the intrusive memory was blank, did not match our definition of intrusive memories (e.g., a verbal rumination), or could not be mapped to a video shown in the experimental session. Conversely, the number of intrusive memories increased if the participant selected “No Intrusions” from the dropdown menu but provided a description of an intrusive memory, if a single log entry could be mapped to multiple intrusions (e.g., “Man shaving and cutting and bleeding x 3”) or multiple videos (e.g., “Crushed leg video and elephant”). All adjustments were discussed and agreed upon as a team without consideration of a participant’s condition assignment.

Pupillometry

Binocular eye tracking was used, and pupil size was averaged across valid samples of left and right eyes. Pupil sizes were baselined to the mean from a 3-minute baseline period at the start of the session. Eye-tracking data were not recorded for one participant due to technical issues, and five others were missing data for specific phases due to either absent synchronization timestamps (from technical or network issues) or if neither eye provided any usable samples, (typically due to tracking loss).

In total, eye-tracking data were recorded for 99 of 100 participants (control: n_c=50, intervention: n_i=49). We analyzed pupil sizes during four phases of the experiment: baseline (3 minutes), rest (10 minutes), memory reminder (variable length, due to varying number of entries across participants), and cognitive task (15 minutes). Pupil data were available from the baseline for 97 participants (n_c=49, n_i=48), rest for 99 (n_c=50, n_i=49), memory reminder for 96

(n_c=50, n_i=46), and cognitive task for 98 (n_c=50, n_i=48). Statistical analyses comparing baselined pupil size between components (e.g., cognitive task vs. rest) were restricted to participants with valid data in all three components (baseline, cognitive task, and rest), ensuring consistent within-subject comparisons. For the mixed-effects model that related game difficulty to pupil size across pieces, we included participants with valid baselined pupil size data during the intervention cognitive task (n=48). For mixed-effects models spanning both phases (cognitive task and memory reminder), we included all participants with valid baselined pupil size data from both phases. Exact sample sizes are reported for each analysis.

Statistics

Summary statistics are reported as the mean with 95% Confidence Intervals (CIs). The preregistration stated that *"an independent samples t-test will assess the difference between groups, assuming data normality"* for the primary hypothesis. However, the data used to test the primary hypothesis (i.e., the total number of intrusive memories) violated the assumption of normality, as indicated by the Shapiro-Wilk Test (p<0.001) and confirmed through visual inspection. Therefore, statistical tests were conducted using non-parametric permutation tests (100,000 iterations). Despite the non-normality, the results were robust: a parametric test of the primary hypothesis (i.e., independent samples t-test) also yielded a consistent and statistically significant pattern of findings. The results of the parametric statistical tests are included in the *Supplementary Results*. Although the primary hypothesis was directional in nature ("*participants receiving the intervention will report fewer intrusive memories relative to participants who receive a control task*"), we report two-tailed p-values for all statistical tests. The test of the primary analysis in the Results section used the raw total number of intrusive memories from the electronic logs for all participants (n=100). We conducted two additional exploratory analyses of the primary hypothesis: first, following data quality control of the intrusive memory log entries; second after excluding participants with protocol deviations or who were identified as outliers (see *Supplementary Results*). These adjustments had minimal influence on summary statistics and did not change the outcome of the primary hypothesis. Correlations were computed using the Spearman rank correlation, which is appropriate for non-parametric data. Analysis of pupillometry data were conducted on all participants with available eye-tracking data from the relevant portions of the study, in order to maximize data inclusion.

We fit linear mixed-effects models using the *statsmodels* package (version 0.14.1), to account for individual differences. We report parametric estimates and associated confidence intervals for the mixed-effects models, as some permutation-based models failed to converge. All analyses were conducted in Python, and all analysis scripts are available in the code repository on OSF.

Data availability

The data supporting the findings of this study are available in the OSF repository at https://osf.io/anvuk/?view_only=9a0b4ba867d44edea4914c14c836f628, and will be made publicly available upon acceptance. This includes the behavioral data (e.g., intrusive memory counts from the electronic logs, text entries during the memory reminder phase, vigilance-intrusion task performance, intervention gameplay metrics), as well as physiological data (e.g., pupillometry recordings). Data that may include identifiable information and materials that

constitute proprietary intellectual property are described in the Materials and Methods section but are not available for distribution.

Code availability

All code used to reproduce the analyses and figures in this study is available in the OSF repository at https://osf.io/anvuk/?view_only=9a0b4ba867d44edea4914c14c836f628, and will be made publicly available upon acceptance.

Acknowledgements

This work is supported by Wellcome Leap. AI tools assisted with language editing and phrasing suggestions; all content was reviewed and edited by the authors. We also acknowledge the British Film Institute National Archives, LyleBailie International, Kino International, and Frontline for film materials.

Author Contributions

All authors conceived the study. S.S. collected the data, M.T.dB. and M.C. supported data collection. M.T.dB. and S.S. analyzed the data and drafted the figures and the manuscript. All authors (M.T.dB., S.S., E.A.H., M.C.) interpreted the data, reviewed, revised, and approved the final version.

Competing Interests

M.T.dB., S.S., M.C. hold equity in Ruby Neurotech, a company whose interests may be affected by the research reported in this article. E.A.H. developed the imagery-competing task intervention for intrusive memories and holds the trademark (ANEMONE™) through Afterimagery.AB. E.A.H. receives book royalties from Guildford Press and Oxford University Press and receives occasional honoraria for conference keynotes and clinical workshops. E.A.H. also receives funding from the Swedish Research Council (2020–00873). E.A.H. is on the Board of Trustees of the MQ Foundation.

References

Bibliography

1. Kessler, R. C. *et al.* Trauma and PTSD in the WHO World Mental Health surveys. *Eur. J. Psychotraumatol.* **8**, 1353383 (2017).
2. Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **62**, 593–602 (2005).
3. Schein, J. *et al.* Prevalence of post-traumatic stress disorder in the United States: a systematic literature review. *Curr. Med. Res. Opin.* **37**, 2151–2161 (2021).
4. Davis, L. L. *et al.* The economic burden of posttraumatic stress disorder in the United States from a societal perspective. *J. Clin. Psychiatry* **83**, (2022).
5. Stein, D. J. *et al.* Determinants of effective treatment coverage for posttraumatic stress disorder: findings from the World Mental Health Surveys. *BMC Psychiatry* **23**, 226 (2023).
6. Spoont, M. R., Murdoch, M., Hodges, J. & Nugent, S. Treatment receipt by veterans after a PTSD diagnosis in PTSD, mental health, or general medical clinics. *Psychiatr. Serv.* **61**, 58–63 (2010).
7. Alonso, J. *et al.* Treatment gap for anxiety disorders is global: Results of the World Mental Health Surveys in 21 countries. *Depress. Anxiety* **35**, 195–208 (2018).
8. Bisson, J. I. & Olff, M. Prevention and treatment of PTSD: the current evidence base. *European Journal of Psychotraumatology* (2021) doi:10.1080/20008198.2020.1824381.
9. Ehlers, A. *et al.* Therapist-assisted online psychological therapies differing in trauma focus for post-traumatic stress disorder (STOP-PTSD): a UK-based, single-blind, randomised controlled trial. *Lancet Psychiatry* **10**, 608–622 (2023).
10. Wright, S. *et al.* Predictors of study dropout in cognitive-behavioural therapy with a trauma focus for post-traumatic stress disorder in adults: An individual participant data meta-analysis. *BMJ Ment. Health* **27**, e301159 (2024).
11. Iyadurai, L. *et al.* Intrusive memories of trauma: A target for research bridging cognitive science and its clinical application. *Clin Psychol Rev* **69**, 67–82 (2019).
12. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5 (R))*. (American Psychiatric Association Publishing, Arlington, TX, 2013).
13. Martalek, A., Dubertret, C., Fovet, T., Le Strat, Y. & Tebeka, S. Distressing memories: A continuum from wellness to PTSD. *J. Affect. Disord.* **363**, 198–205 (2024).
14. Holmes, E. A. & Mathews, A. Mental imagery in emotion and emotional disorders. *Clin. Psychol. Rev.* **30**, 349–362 (2010).
15. Kanstrup, M. *et al.* A guided single session intervention to reduce intrusive memories of work-related trauma: a randomised controlled trial with healthcare workers in the COVID-19 pandemic. *BMC Med.* **22**, 403 (2024).
16. Iyadurai, L. *et al.* Reducing intrusive memories after trauma via an imagery-competing task intervention in COVID-19 intensive care staff: a randomised controlled trial. *Transl. Psychiatry* **13**, 290 (2023).
17. Ramineni, V. *et al.* Treating intrusive memories after trauma in healthcare workers: a Bayesian adaptive randomised trial developing an imagery-competing task intervention. *Mol. Psychiatry* **28**, 2985–2994 (2023).
18. Visser, R. M., Lau-Zhu, A., Henson, R. N. & Holmes, E. A. Multiple memory systems, multiple time points: how science can inform treatment to control the expression of unwanted emotional memories. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, (2018).

19. Sederberg, P. B., Gershman, S. J., Polyn, S. M. & Norman, K. A. Human memory reconsolidation can be explained using the temporal context model. *Psychon. Bull. Rev.* **18**, 455–468 (2011).
20. Ritvo, V. J. H., Turk-Browne, N. B. & Norman, K. A. Nonmonotonic plasticity: How memory retrieval drives learning. *Trends Cogn. Sci.* **23**, 726–742 (2019).
21. Baddeley, A. D. & Andrade, J. Working memory and the vividness of imagery. *J. Exp. Psychol. Gen.* **129**, 126–145 (2000).
22. Holmes, E. A., James, E. L., Coode-Bate, T. & Deeprose, C. Can playing the computer game 'Tetris' reduce the build-up of flashbacks for trauma? A proposal from cognitive science. *PLoS One* **4**, e4153 (2009).
23. Holmes, E. A., James, E. L., Kilford, E. J. & Deeprose, C. Key steps in developing a cognitive vaccine against traumatic flashbacks: visuospatial Tetris versus verbal Pub Quiz. *PLoS One* **5**, e13706 (2010).
24. James, E. L. *et al.* Computer game play reduces intrusive memories of experimental trauma via reconsolidation-update mechanisms. *Psychol. Sci.* **26**, 1201–1215 (2015).
25. Lau-Zhu, A., Henson, R. N. & Holmes, E. A. Intrusive memories and voluntary memory of a trauma film: Differential effects of a cognitive interference task after encoding. *J. Exp. Psychol. Gen.* **148**, 2154–2180 (2019).
26. Varma, M. M. *et al.* A systematic review and meta-analysis of experimental methods for modulating intrusive memories following lab-analogue trauma exposure in non-clinical populations. *Nat. Hum. Behav.* **8**, 1968–1987 (2024).
27. Asselbergs, J. *et al.* A systematic review and meta-analysis of the effect of cognitive interventions to prevent intrusive memories using the trauma film paradigm. *J. Psychiatr. Res.* **159**, 116–129 (2023).
28. Iyadurai, L. *et al.* Preventing intrusive memories after trauma via a brief intervention involving Tetris computer game play in the emergency department: a proof-of-concept randomized controlled trial. *Mol. Psychiatry* **23**, 674–682 (2018).
29. Kanstrup, M. *et al.* Reducing intrusive memories after trauma via a brief cognitive task intervention in the hospital emergency department: an exploratory pilot randomised controlled trial. *Transl. Psychiatry* **11**, 30 (2021).
30. Thorarinsdottir, K. *et al.* Using a brief mental imagery competing task to reduce the number of intrusive memories: Exploratory case series with trauma-exposed women. *JMIR Form. Res.* **6**, e37382 (2022).
31. Kessler, H. *et al.* Reducing intrusive memories of trauma using a visuospatial interference intervention with inpatients with posttraumatic stress disorder (PTSD). *J. Consult. Clin. Psychol.* **86**, 1076–1090 (2018).
32. Kehyayan, A. *et al.* The effect of a visuospatial interference intervention on posttraumatic intrusions: a cross-over randomized controlled trial. *Eur. J. Psychotraumatol.* **15**, 2331402 (2024).
33. Maida, M. *et al.* The role of generative language systems in increasing patient awareness of colon cancer screening. *Endoscopy* **57**, 262–268 (2025).
34. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
35. van der Wel, P. & van Steenbergen, H. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychon. Bull. Rev.* **25**, 2005–2015 (2018).
36. Kahneman, D. & Beatty, J. Pupil diameter and load on memory. *Science* **154**, 1583–1585 (1966).

37. Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* **91**, 276–292 (1982).
38. Picanço, C. R. & Tonneau, F. A low-cost platform for eye-tracking research: Using Pupil© in behavior analysis. *J. Exp. Anal. Behav.* **110**, 157–170 (2018).
39. Wei, W. *et al.* Assessing the cognitive load arising from in-vehicle infotainment systems using pupil diameter. in *Lecture Notes in Computer Science* 440–450 (Springer Nature Switzerland, Cham, 2023).
40. Keene, P. A., deBettencourt, M. T., Awh, E. & Vogel, E. K. Pupillometry signatures of sustained attention and working memory. *Atten. Percept. Psychophys.* **84**, 2472–2482 (2022).
41. Clewett, D., Gasser, C. & Davachi, L. Pupil-linked arousal signals track the temporal organization of events in memory. *Nat. Commun.* **11**, 4007 (2020).
42. Joshi, S. & Gold, J. I. Pupil size as a window on neural substrates of cognition. *Trends Cogn. Sci.* **24**, 466–480 (2020).
43. Konishi, M., Brown, K., Battaglini, L. & Smallwood, J. When attention wanders: Pupillometric signatures of fluctuations in external attention. *Cognition* **168**, 16–26 (2017).
44. Madsen, J. & Parra, L. C. Narratives engage brain and body: bidirectional interactions during natural story listening. *Neuroscience* (2023).
45. James, E. L. *et al.* The trauma film paradigm as an experimental psychopathology model of psychological trauma: intrusive memories and beyond. *Clin. Psychol. Rev.* **47**, 106–142 (2016).
46. Sharma, A., Rushton, K., Lin, I. W., Nguyen, T. & Althoff, T. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. in *Proceedings of the CHI Conference on Human Factors in Computing Systems* vol. 21 1–29 (ACM, New York, NY, USA, 2024).
47. Inan, H. *et al.* Llama Guard: LLM-based input-output safeguard for Human-AI conversations. *arXiv [cs.CL]* (2023).
48. Rebedea, T., Dinu, R., Sreedhar, M., Parisien, C. & Cohen, J. NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails. *arXiv [cs.CL]* (2023).
49. Ayyamperumal, S. G. & Ge, L. Current state of LLM Risks and AI Guardrails. *arXiv [cs.CR]* (2024).
50. Clewett, D. & Murty, V. P. Echoes of emotions past: How neuromodulators determine what we recollect. *eNeuro* **6**, ENEURO.0108–18.2019 (2019).
51. Areias, A. C. *et al.* Transforming veteran rehabilitation care: Learnings from a remote digital approach for musculoskeletal pain. *Healthcare (Basel)* **12**, 1518 (2024).
52. Detre, G. J., Natarajan, A., Gershman, S. J. & Norman, K. A. Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia* **51**, 2371–2388 (2013).
53. Bonsall, M. B. & Holmes, E. A. Temporal dynamics of trauma memory persistence. *J. R. Soc. Interface* **20**, 20230108 (2023).
54. Agren, T., Hoppe, J. M., Singh, L., Holmes, E. A. & Rosén, J. The neural basis of Tetris gameplay: implicating the role of visuospatial processing. *Curr. Psychol.* **42**, 8156–8163 (2023).
55. Kay, L., Keogh, R., Andrillon, T. & Pearson, J. The pupillary light response as a physiological index of aphantasia, sensory and phenomenological imagery strength. *Elife* **11**, (2022).
56. Yeung, R. C., Sokolowski, H. M., Fan, C. L., Fernandes, M. A. & Levine, B. The curse of

imagery: Trait object and spatial imagery differentially relate to symptoms of posttraumatic stress disorder. *Clin. Psychol. Sci.* (2025) doi:10.1177/21677026251315118.

57. Piaggio, D. *et al.* Pupillometry via smartphone for low-resource settings. *Biocybern. Biomed. Eng.* **41**, 891–902 (2021).
58. Barry, C. *et al.* At-home pupillometry using smartphone facial identification cameras. *Proc. SIGCHI Conf. Hum. Factor. Comput. Syst.* **2022**, (2022).
59. deBettencourt, M. T., Cohen, J. D., Lee, R. F., Norman, K. A. & Turk-Browne, N. B. Closed-loop training of attention with real-time brain imaging. *Nat. Neurosci.* **18**, 470–475 (2015).
60. Corriveau, A., Rosenberg, M. D. & deBettencourt, M. T. Cognitive neuroscience of attention and memory dynamics. *PsyArXiv* (2025) doi:10.31234/osf.io/n7tma_v1.
61. Saproo, S., Shih, V., Jangraw, D. C. & Sajda, P. Neural mechanisms underlying catastrophic failure in human–machine interaction during aerial navigation. *J. Neural Eng.* **13**, 066005 (2016).
62. Lau-Zhu, A., Henson, R. N. & Holmes, E. A. Selectively interfering with intrusive but not voluntary memories of a trauma film: Accounting for the role of associative memory. *Clin. Psychol. Sci.* **9**, 1128–1143 (2021).
63. Torous, J. *et al.* The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry* **24**, 156–174 (2025).
64. Bufano, P., Laurino, M., Said, S., Tognetti, A. & Menicucci, D. Digital phenotyping for monitoring mental disorders: Systematic Review. *J. Med. Internet Res.* **25**, e46778 (2023).
65. Coravos, A., Khozin, S. & Mandl, K. D. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ Digit. Med.* **2**, 1–5 (2019).
66. Fernandez Rojas, R., Brown, N., Waddington, G. & Goecke, R. A systematic review of neurophysiological sensing for the assessment of acute pain. *NPJ Digit. Med.* **6**, 76 (2023).
67. Galatzer-Levy, I. R. & Onnela, J.-P. Machine learning and the digital measurement of psychological health. *Annu. Rev. Clin. Psychol.* **19**, 133–154 (2023).
68. Holmes, E. A., Grey, N. & Young, K. A. D. Intrusive images and 'hotspots' of trauma memories in Posttraumatic Stress Disorder: an exploratory investigation of emotions and cognitive themes. *J. Behav. Ther. Exp. Psychiatry* **36**, 3–17 (2005).
69. Grey, N. & Holmes, E. A. 'Hotspots' in trauma memories in the treatment of post-traumatic stress disorder: a replication. *Memory* **16**, 788–796 (2008).
70. National Public Radio. *Classical Pianist Jeremy Denk*. (National Public Radio, 2022).
71. Blackburn, I.-M. *et al.* The revised cognitive therapy scale (cts-r): Psychometric properties. *Behav. Cogn. Psychother.* **29**, 431–446 (2001).
72. Dreyfus, H. L. *The Dreyfus Model of Skill Acquisition*. 181–183 (Falmer Press, London, 1989).

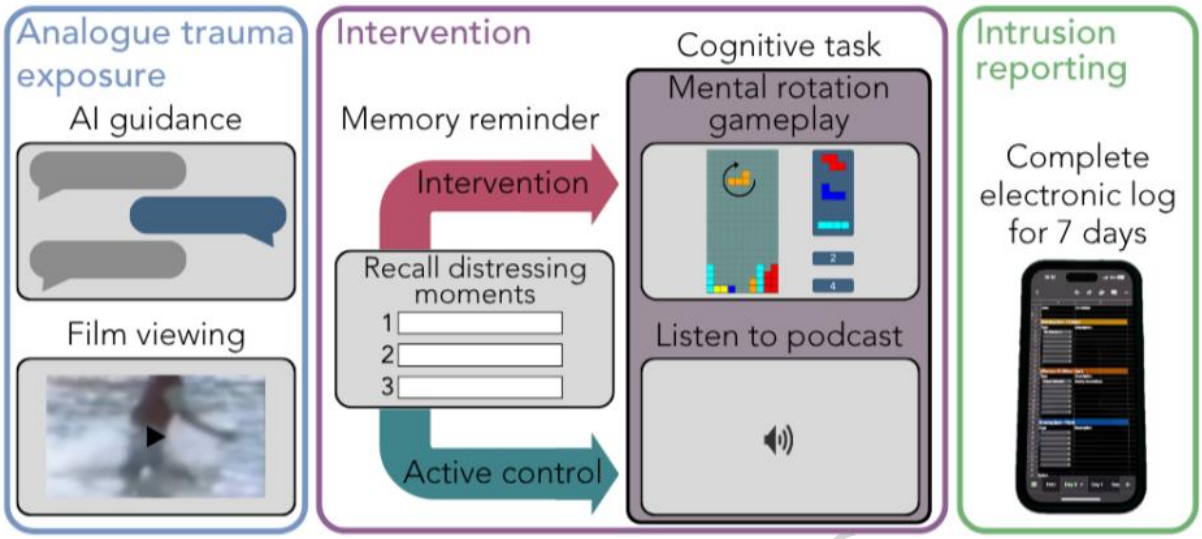

Figure 1 ANTIDOTE experimental protocol overview. Participants completed ANTIDOTE, an AI-guided intelligent neurotech prototype to reduce the number of intrusive memories in an experimental model of trauma. Participants were continuously monitored throughout the protocol with neurophysiological measures. All participants received instructions during the protocol from an AI guide and were exposed to analogue trauma (an 11.5-minute film composed of traumatic video clips). A blurred still from a video is shown for illustrative purposes. After a brief rest period (10 minutes, not depicted), all participants were given a memory reminder to recall and briefly describe their most distressing moments from the film. Next, participants completed a cognitive task (15 min) according to their random condition assignment. The intervention group (red) played a visuospatial block puzzle game that emphasized mental rotation and mental imagery during computer gameplay, while the active control group (blue) listened to a podcast discussing classical music. Finally, participants reported intrusions, electronically logging and briefly describing any intrusive memories from the film for the following 7 days. ANTIDOTE was derived from a digital treatment for intrusive memories, the Imagery Competing Task Intervention (ICTI) previously developed from our group and originally delivered with a human guide.

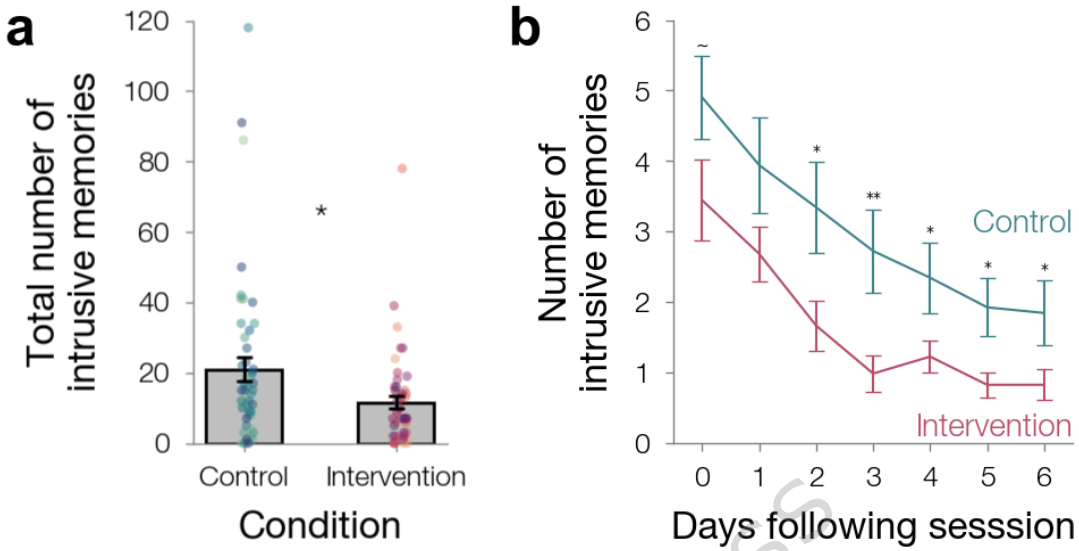

Figure 2 ANTIDOTE reduces the number of intrusive memories. (a) Total number of intrusive memories reported in the electronic log per condition. Participants in the intervention condition reported significantly fewer intrusive memories than those in the control condition (* p=0.01). The total number of intrusive memories was totaled over the 7-day period following the experimental session. Each dot represents one participant. Bars show group means; error bars indicate standard errors of the mean. **(b)** Time course of intrusive memories reported per day during the 7-day period. The in-person study session occurred on Day 0. Lines indicate group means for the intervention (blue) and control (red) conditions; error bars indicate standard errors of the mean. Statistically significant or trending between-group differences on individual days are marked (** p<0.01, * p<0.05, ~ p<0.1).

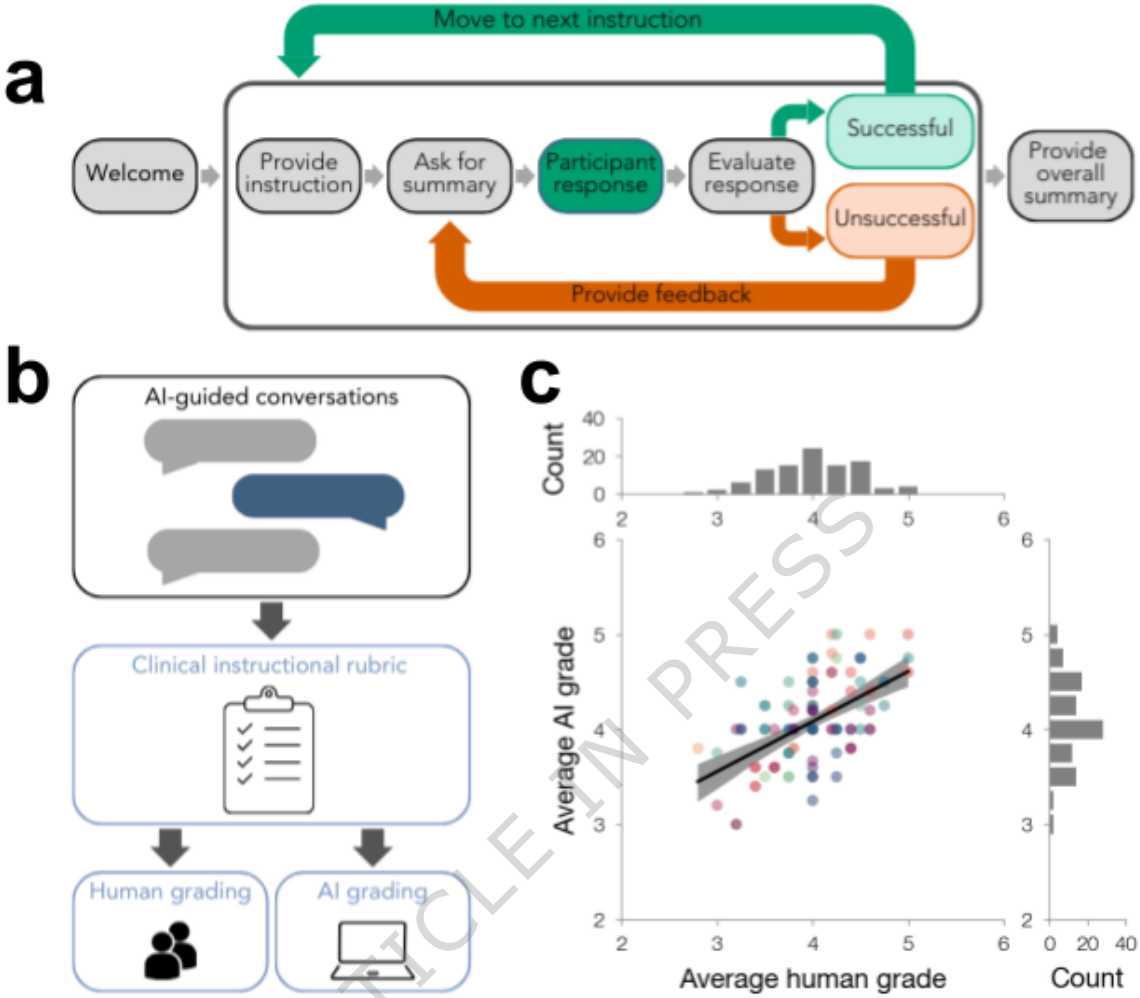

Figure 3 AI guidance and evaluation. (a) ANTIDOTE delivered automated instructions throughout key components of the experimental protocol by a series of five human-AI conversations. The AI guide delivered multiple instruction segments, each as a discrete step: the AI presented the instruction, asked the participant to summarize it, and evaluated the participant's response. If the participant included all key points, the summary for that segment was accepted, and the conversation moved to the next instruction (green). Otherwise, the AI provided corrective feedback and requested a revised summary for that instruction before moving on (red). Upon successful completion of all instruction segments, the AI guide presented a consolidated summary. **(b)** Each AI-guided instruction conversation was evaluated using a rubric previously developed to train human psychology researchers and clinicians to deliver the instructions[15–17]. In this study, we used two types of ratings: (i) two human raters scored each AI-participant conversation, and (ii) an AI grader assigned scores based on the same criteria. **(c)** Human and AI grading alignment. The human grades (*x*-axis) were strongly correlated with the AI grades (*y*-axis; $p<0.001$). Each dot represents a participant's mean score across all conversations. The line depicts the linear fit, and the shaded area reflects 95% CIs. Score distributions for each rater type are projected onto the respective axes as marginal histograms, with bar height indicating the number of participants per bin.

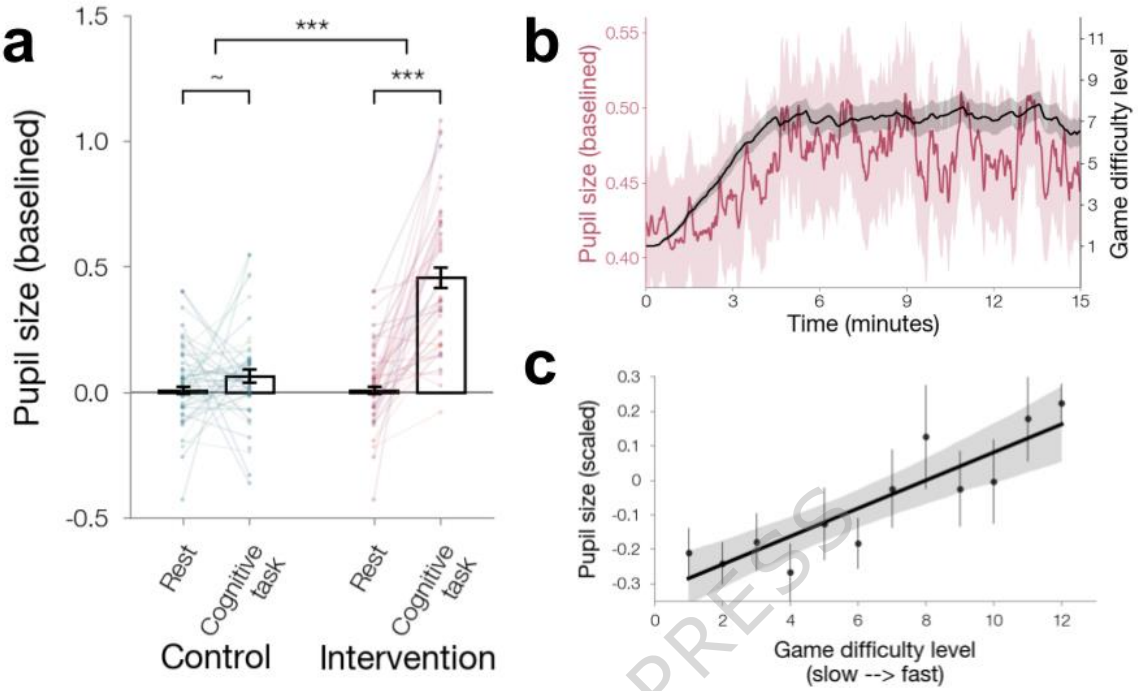

Figure 4 Pupillometry measures during the cognitive task portion of the intervention. (a) The mean pupil size was calculated for both the intervention and control groups, during the 10-minute rest period after watching the videos and during the 15-minute cognitive task (intervention or control). The pupil size during the cognitive task was reliably greater than during rest for the intervention group (mental rotation gameplay, *** p<0.001), and trending but not reliable for the control group (podcast listening, ~ p=0.07). The interaction between the intervention and control groups was reliable (*** p<0.001). Pupil sizes were baselined to a 3-minute period at the beginning of the experimental protocol. The height of the bar is the population mean, and the error bars show the standard errors of the mean. Each participant is depicted as a dot, and data from the same participant are connected by a line. **(b)** Pupil size and game difficulty dynamically fluctuate over time. Across the 15-minute intervention, the pupil size (red) and game difficulty level (black) varied for each participant in the intervention group. The game difficulty level ranged from 1 (the slowest and starting game speed) to 12 (the fastest and hardest game speed). The lines represent the average trajectory over time for all intervention participants, and the shaded areas are the standard errors of the means. **(c)** Pupil size increases with game difficulty level. A multilevel linear regression was used to model the positive relationship between mean pupil size and difficulty level 1–12 (p<0.001), accounting for variation across participants in the intervention group. For visualization purposes, the plot depicts an ordinary least squares regression: each dot shows the mean baselined pupil size data (demeaned and scaled within participants) for a given difficulty level, with error bars representing the standard error of the mean and the shaded area representing 95% CIs.

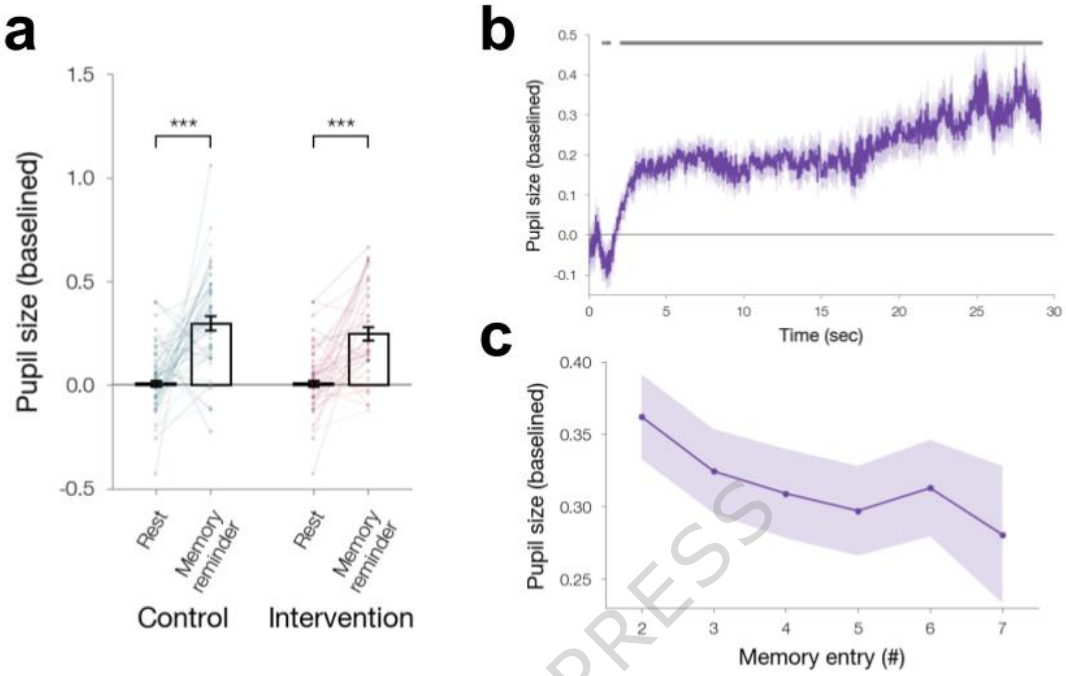

Figure 5 Pupillometry measures during the memory reminder period of the intervention. **(a)** The mean pupil size was calculated for both the intervention and control groups, during the 10-minute rest period after watching the videos, and during the memory reminder period, which was of variable length due to variable numbers of entries. The pupil size was reliably greater than during rest for both the intervention and control groups (*** *p*s<0.001). There was no reliable interaction between groups, consistent with the fact that the memory reminder occurred prior to when the intervention and control groups diverged. Pupil sizes were baselined to a 3-minute period at the beginning of the experimental protocol. The height of the bar is the population mean, and the error bars show the standard errors of the mean. Each participant is depicted as a dot, and data from the same participant are connected by a line. **(b)** Pupil size dynamics for the first entry. We analyzed pupil size from the onset of the memory reminder screen until the first entry, truncating the window to the shortest duration across all participants (*t*=29 sec). Pupil size was initially not reliably different from baseline; significant time points (*p*s<0.01, uncorrected) are marked by gray along the top of the figure. The purple line shows the mean pupil size relative to baseline, and the shaded area represents the standard error of the mean. **(c)** Pupil size per subsequent memory entries. The purple line shows the mean pupil size for each entry relative to the baseline, the shaded area is the standard error of the mean. For visual clarity, only entries up to 7 are shown, as higher numbers of entries were rare. However, all eligible entries were included in the statistical model. Pupil size was largest for the earlier entries and declined with increasing entry number (*p*<0.01).

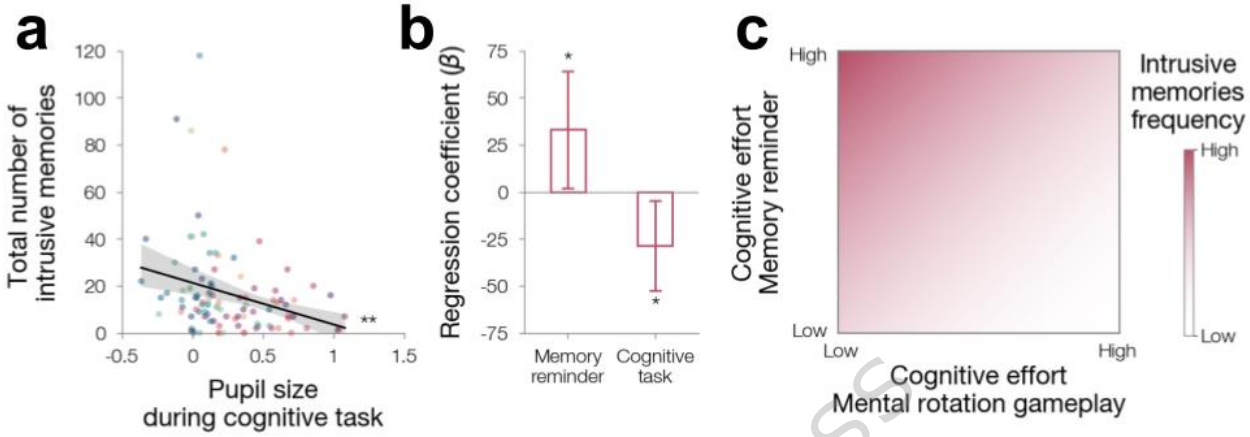

Figure 6 Physiological predictors of intervention success. (a) Pupil size during the cognitive task is associated with intrusive memories. Mean pupil size during the cognitive task (for both intervention and control groups combined) was negatively correlated with the number of intrusive memories reported in the 7-day electronic log (** p=0.002). Pupil size for the cognitive task was baselined against pupil size from the 3-minute rest period collected prior to video viewing. Each participant is depicted as a dot in a unique color; participants in the intervention group are red and the control group are blue. The line depicts the linear fit, and the shaded area is 95% CIs. **(b)** Joint model of memory reminder and cognitive task effort within the intervention group alone. Together pupil size during memory reminder and during the cognitive task, both normalized against the same 3 minute rest period, predicted intrusive memories, but in opposite directions: greater pupil size during the task predicted fewer intrusions, while greater pupil size during the memory reminder predicted more intrusions (* ps<0.05). The height of the bar is the regression coefficient, the error bars represent the confidence intervals. **(c)** Conceptual model of cognitive effort during ANTIDOTE. To reduce the frequency of intrusive memories, the optimal strategy may be to expend low cognitive effort during the memory reminder phase, and high cognitive effort during mental rotation gameplay.