A VR and AI-Powered Game-Based Approach to Minimally-Invasive Social Cognitive Assessment

Krijgsman, Caspar; Poos, Jackie; Bidarra, Rafael

DOI
10.1109/CoG64752.2025.11114400

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the IEEE 2025 Conference on Games, CoG 2025

A VR and AI-powered game-based approach to minimally-invasive social cognitive assessment

Caspar Krijgsman
Delft University of Technology
Delft, The Netherlands
Email: caspar@bykrijgsman.com

Jackie Poos
Department of Neurology
Erasmus MC University Medical Center
Rotterdam, The Netherlands
Email: j.m.poos@erasmusmc.nl

Rafael Bidarra
Computer Graphics and Visualization Group
Delft University of Technology
Delft, The Netherlands
Email: r.bidarra@tudelft.nl

***Abstract*—Frontotemporal dementia (FTD) is the second most common cause of young-onset dementia and is typically caused by frontal and/or temporal lobe degeneration [1]. This degeneration impacts personality, judgment, and the ability to recognize emotion. Early diagnosis of FTD is critical for effective treatment, but current diagnostic tools often lack the realism necessary to assess social cognitive abilities like emotion recognition and theory of mind. We propose the customization of an immersive virtual reality (VR) experience to improve early FTD diagnosis by creating realistic social interactions in immersive environments. Our approach is two-fold: (i) an authoring tool that gives neuropsychologists generative AI support for writing a suitable dialog involving faux pas, picking characters and voicing them; (ii) a VR game in which patients observe the enactment of those scripted social scenarios, guided by a virtual companion. We discuss the main features of our current prototype, and identify some development and evaluation challenges ahead.**

***Index Terms*—Frontotemporal dementia, AI-assisted authoring, Virtual reality, Theory of mind, Serious games**

Video Link: https://youtu.be/np0CLYwuhlo

I. Introduction

Frontotemporal dementia (FTD) is a leading cause of young-onset dementia, characterized by social skills impairment due to frontal and/or temporal lobe degeneration. Social cognition, encompassing emotion recognition, theory of mind, and moral reasoning, is commonly affected [1]. Objective tests show deficits in these areas for advanced FTD and Alzheimer's disease (AD) [2]–[4]. Early diagnosis is crucial for effective treatment, ideally in early stages with minimal neuronal loss.

From the variety of social cognitive tests, this research focuses on the Theory of Mind (ToM) tasks of such tests and, more specifically, the Faux Pas test. This test assesses the participant's ability to detect and identify a *faux pas*. Faux pas is a French term denoting unintentional words or behavior that are a social mistake or impolite. In this test, participants are shown and read stories that may contain a faux pas. It is the participant's task to identify this faux pas and then explain why it is classified as such. Passing this task would indicate a social maturity of 9 to 11 years old, as this is the age at which children tend to develop an understanding of social faux pas [5]. The questionnaire in these tests is standardized and consists of roughly 10 questions, some of which serve as control questions without faux pas [6], [7].

So far, faux pas tests have been taken on paper, relying on the participant's ability to 'dive into' the scene or dialogue described. Virtual Reality (VR) technology potentially offers immersive, controlled simulations for early diagnosis, enhancing engagement and replicating real-world interactions [8]. VR is more immersive than traditional methods and has been applied in autism and schizophrenia [9], [10], but little for dementia diagnostics [11]. Recent studies explored VR in AD and FTD to assess memory and navigation [12] but, to the best of our knowledge, no study has explored the use of VR for assessing social cognition deficits in FTD. We suggest that VR can improve diagnosis, making it more reliable, accessible, and engaging, provided that neuropsychologists are in control of tests' customization, which ensures both their efficacy and patient-tailored experiences.

II. Related work

Related work for this research can be classified into three areas: the *intersection of AI with the notion of faux pas*, support *methods for story authoring*, and *VR in diagnosis and treatments*. Here, we briefly survey some of the most relevant work in each of these areas.

Over the past few years, some preliminary studies have been conducted on LLMs' abilities to succeed in ToM tasks. In 2024, Strachan et al. studied the theory of mind in various LLMs, including ChatGPT 3.5 and ChatGPT 4. [13] They found that most models performed well on the first-order and second-order false-believe tests, with ChatGPT 3.5 and ChatGPT 4 outperforming human test subjects. Moving up to the faux pas task, they found these models lacking. While both ChatGPT 3.5 and 4 correctly identified the mental state of both characters involved, their answers to the questions were insufficient. An earlier study by Shapira et al. confirms these findings, showing that they can identify the faux pas and comprehend the story but fail to correctly answer the detection question and the reasoning [14]. Beyond taking the test, Shapira et al. also show it is hard for LLMs to generate these stories, resulting in only simplistic ones. They mostly fail to create implicit scenarios where one of the characters

Short Paper

2025 IEEE Conference on Games (CoG) | 979-8-3315-8904-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/COG64752.2025.11114400

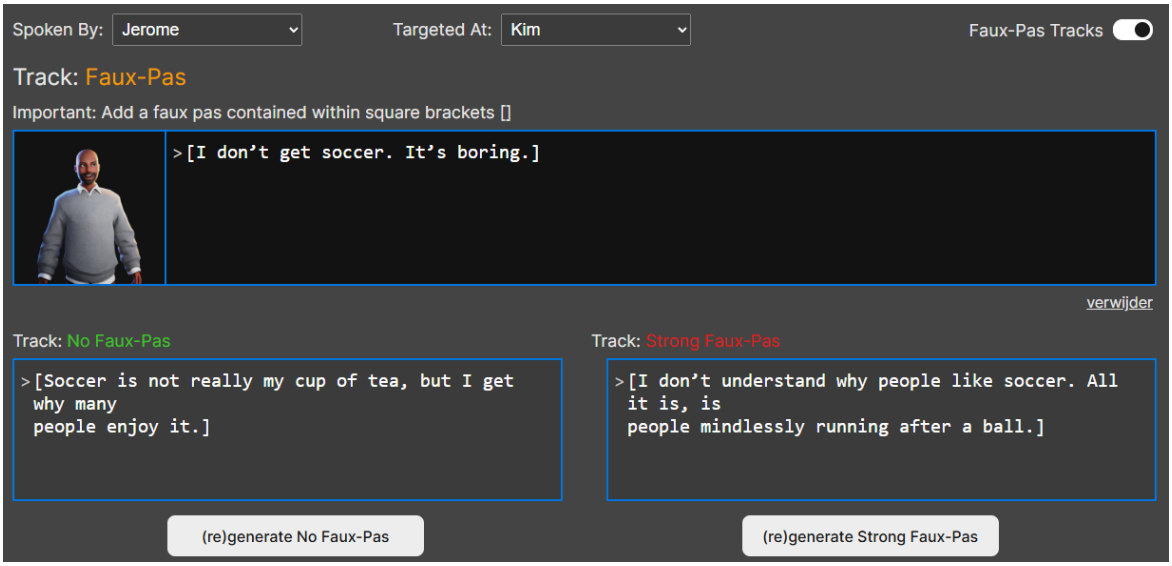

Fig. 1: A screenshot of the authoring tool in Faux-Pas mode, showing the main editable entry in the top center, with the AI-generated versions no faux pas (bottom-left) and strong faux pas (bottom-right).

lacks information (e.g., explicitly mentioning "not realizing that the woman was one of the guests at the dinner party").

Research into digital story authoring has a long history [15]. Back in 1997, Umaschi and Cassell proposed SAGE (Storytelling Agent Generation Environment), a construction kit that supports children's creation of their own wise storytellers to interact with [16]. Work like this has inspired tools for so-called mixed-initiative story authoring, particularly focused on assisting rather than replacing authors in their creative endeavor [17]. This has also been applied within the healthcare sector, like research by Bickmore and Ring [18].

With the adoption of more simulated approaches, VR has acquired an increasingly attractive role in healthcare. Existing studies in schizophrenia and autism have successfully deployed VR environments in treatment plans [19], [20]. Besides for treatment, also diagnosis is finding promising developments [21]. Tan et al. presented a screening tool that tests the elderly across six cognitive domains [22], and a large-scale citizen science project by Spiers et al. provided contributions to developing personalized diagnostics, particularly for early Alzheimer's detection [23]. These are not entirely new research topics; in the late 1990s, Elkind already identified the usefulness of virtual reality to diagnose people with neurological dysfunctions [24].

III. Approach

Our approach is twofold. First, we developed an authoring tool that enables neuropsychologists, not programmers, to design the desired social interactions carefully. The challenge here is to make this content creation process accessible across all its steps. Second, we developed a complete pipeline for the multimedia enactment of that authored social interaction. Here, we deploy all the data of the authored social interactions, to create its immersive enactment for the participant.

A. Authoring

We approach authoring by first dividing it into two aspects, the story and the metadata. The story consists of narration and dialogue and encompasses all that could be read or spoken in a scenario. The metadata consists of the additional information needed to create a multimedia experience from the story.

For the story aspect, we can rely primarily on conventional authoring methods. Text editors can help authors create both the dialogue and the narration. Within the dialogue, one can insert a faux pas, when and where desired. For this, authoring extends beyond simple text editing through AI co-creation. The author is asked to provide a faux pas of a subtle variety while generative AI writes along, providing both a version without a faux pas and one with a strong faux pas, as can be seen in figure 1. In our prototype, we use Chat-GPT4 [25] as the LLM backend facilitating the co-created versions. Naturally, the human author can always manually tweak those versions, if desired.

The authoring metadata regards, among other aspects, scene appearance and character features. For example, it contains information on the scene location, characters' look and voices, as well as where the characters are looking during the conversation and their emotions. The metadata is authored alongside the dialog text boxes, for example selecting options for emotions and targets. Other settings, like location and character appearances, are configurable in the pre-writing configuration window. In our prototype, character creation is facilitated by Ready Player Me, offering an out-of-the-box video game avatar creator compatible with Unity [26].

B. Immersive Enactment

The second element of the approach is a multimedia platform capable of enacting the previously authored social interaction content. Here, we identify four main elements:

scenery, *sound*, *realism*, and *interaction*. As mentioned in Section II, VR has promising applications in the mental health sector. Therefore, we chose VR as the preferred platform for enacting the authored scenes. However, the authored content can also be enacted on a normal computer display, if desired.

Our approach to scenery is to have several pre-built 3D environments where all our social interactions can be enacted. After discussions with experts, we realized one can tell hundreds of different stories and maintain sufficient diversity within several generic locations. These locations include a living room, an office, and a store. However, in the future, these could as well be AI-generated on demand, if required.

The second aspect of story enactment is sound, which encompasses both ambient sounds and the voices of the characters. Since the focus should be entirely on the conversation, ambient sound is kept at a minimum. Each character should have a unique human-sounding voice. Recent improvements in generative AI for text-to-speech overcome the need for voice actors to pre-record a dialogue [27]. Offering a list of voice models with samples allows neuropsychologists to assign best-fitting voices for the chosen actors. The prototype utilizes voices from ElevenLabs [28] and consists of a list of 3 Dutch male and 3 Dutch female voices.

The third aspect is realism. For the purposes of our approach, no photorealism is needed; however, all characters and objects are kept to realistic sizes, and the appearance of characters and objects is mundane. So the most relevant realism here is related to social interactions: gaze direction, mouth movements, gestures and emotional expressions.

Finally, interaction is the touchstone of the enactment. In Section 1 we mentioned that faux pas tests require neuropsychologists to ask participants questions about the social interaction. Doing so with the participants inside a virtual reality experience will draw them out of the experience. To avoid that break of immersion, we introduce a virtual assistant, controlled by the neuropsychologists, which will guide the participant and ask questions.

These four aspects are brought together into a prototype. The prototype was created using Unity and uses OpenXR for the VR controls. It is an Asynchronous VR experience, meaning that the participant in VR has a different experience than the person behind the monitor, who can control the experience, e.g. skip or start new interactions. Figure 2 shows a screenshot of the prototype.

IV. Evaluation

Evaluation of this approach focuses on two phases: the evaluation of the authoring experience and the evaluation of the immersive enactment in the VR experience.

The authoring experience was evaluated on its usability, as well as on the quality of the AI Co-authoring of faux pas with severity variations. We had several neuropsychologists use the tool, creating two test stories each. Meanwhile, we monitored how often they generated faux pas tracks and what manual changes they made to the tracks. For the usability, we used the System Usability Scale (SUS) [29]. After creating both

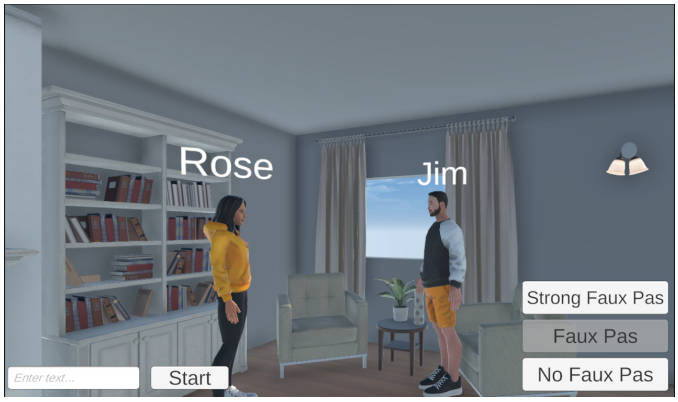

Fig. 2: A screenshot of the immersive enactment from the perspective of the specialist monitoring the test.

stories, their respective spoken versions were played back to them. They were then asked to rate the quality of the voices following the text-to-speech quality test [30]. Results from the AI co-authoring of faux pas show six regenerations over the first eight tracks. These regenerations were equally divided between the no faux pas and strong faux pas tracks. When asked about the reason for the regeneration, the answers were curiosity in all but one case. No manual changes were made to any of the generated tracks. Here, we present one segment of the test stories they authored, showing the three alternative co-authored tracks (green: no faux pas; orange: normal faux pas; red: strong faux pas):

Introduction: *Jerome and Mark are having lunch at Jerome's place. Mark had his first soccer practice that day. The television playing in the background is showing a soccer match.*

Jerome:

Soccer is not really my cup of tea, but I get why many people enjoy it.

I don't get soccer. It's boring.

I don't understand why people like soccer. All it is, is people mindlessly running after a ball.

With the current results on the System Usability Scale [29], we achieve an average usability score of 82 out of 100. From the lowest scored items, we conclude that a simple user manual or quick guide could easily raise their scores.

The first results of the text-to-speech quality test [30] are harder to interpret because they are gathered from various voices and audio samples. However, on average, the naturalness of voices is acceptable, and the comprehension is more than sufficient. All test participants indicated that these voices could nearly always be deployed for an artificial social interaction. Most of them even commented on how impressively natural these voices sound.

The VR experience will be administered to a healthy control group. Here, we want to measure the level of immersion and dissociation of the real world. To measure this, we used a modified version of the IEQ questionnaire [31]. These

modifications omit aspects of immersion through interaction, given that the experience is mostly a passive observation, and they follow those proposed by Rigby et al. [32] for adoption to passive media. As of writing, this evaluation is still underway.

V. Conclusion

So far, early diagnosis of Frontotemporal dementia (FTD) has been using standard paper and pen faux-pas tests to assess theory-of-mind abilities, an approach that lacks in realism, immersion and customizability. We presented a novel approach to improve early FTD diagnosis that (i) strongly facilitates the authoring of customized dialogue narratives containing fine-tuned faux pas elements; and (ii) realistically recreates social interactions in an immersive VR environment. In the narrative authoring phase, neuropsychologists use a dialogue creation online tool that provides AI-assistance for faux pas fine-tuned generation. In the immersive enactment phase, a patient observes the enactment of such an authored narrative, guided by a virtual companion present in that same scenario.

Evaluation of the complete pipeline is presently underway, but preliminary results are very encouraging, with neuropsychologists particularly praising the accessibility of content authoring, as well as its flexibility and seamless integration.

References

[1] L. C. Jiskoot, J. M. Poos, M. E. Vollebergh, S. Franzen, J. van Hemmen, J. M. Papma, J. C. van Swieten, R. P. Kessels, and E. van den Berg, "Emotion recognition of morphed facial expressions in presymptomatic and symptomatic frontotemporal dementia, and Alzheimer's dementia," *Journal of neurology*, vol. 268, pp. 102–113, 2021.

[2] N. Moreau, S. Rauzy, F. Viallet, and M. Champagne-Lavau, "Theory of mind in Alzheimer disease: Evidence of authentic impairment during social interaction." *Neuropsychology*, vol. 30, no. 3, p. 312, 2016.

[3] L. L. Russell, C. V. Greaves, M. Bocchetta, J. Nicholas, R. S. Convery, K. Moore, D. M. Cash, J. van Swieten, L. Jiskoot, F. Moreno *et al.*, "Social cognition impairment in genetic frontotemporal dementia within the genfi cohort," *cortex*, vol. 133, pp. 384–398, 2020.

[4] E. Van den Berg, J. Poos, L. Jiskoot, B. Montagne, R. Kessels, S. Franzen, J. Van Hemmen, W. Eikelboom, E. Heijboer, J. De Kriek *et al.*, "Impaired knowledge of social norms in dementia and psychiatric disorders: Validation of the social norms questionnaire–Dutch version (snq-nl)," *Assessment*, vol. 29, no. 6, pp. 1236–1247, 2022.

[5] M. Adenzato, M. Cavallo, and I. Enrici, "Theory of mind ability in the behavioural variant of frontotemporal dementia: an analysis of the neural, cognitive, and social levels," *Neuropsychologia*, vol. 48, no. 1, pp. 2–12, 2010.

[6] A. A. Spek, E. M. Scholte, and I. A. Van Berckelaer-Onnes, "Theory of mind in adults with HFA and Asperger syndrome," *Journal of autism and developmental disorders*, vol. 40, pp. 280–289, 2010.

[7] A. Spek, "Faux-pas," Mar 2018. [Online]. Available: https://www.anneliesspek.nl/faux-pas/

[8] A. Oker, "Embodied social cognition investigated with virtual agents: The infinite loop between social brain and virtual reality," *Frontiers in Virtual Reality*, vol. 3, p. 962129, 2022.

[9] M. R. Kandalaft, N. Didehbani, D. C. Krawczyk, T. T. Allen, and S. B. Chapman, "Virtual reality social cognition training for young adults with high-functioning autism," *Journal of autism and developmental disorders*, vol. 43, pp. 34–44, 2013.

[10] K. Gainsford, B. Fitzgibbon, P. B. Fitzgerald, and K. E. Hoy, "Transforming treatments for schizophrenia: Virtual reality, brain stimulation and social cognition." *Psychiatry research*, vol. 288, p. 112974, 2020.

[11] M. Bekooy, D. D. Berendsen, M. Dierikx, R. Piepenbrink, J.-W. v. Rhenen, and R. Bidarra, "A theory-of-mind game for the early detection of frontotemporal dementia," in *Interactive Storytelling: International Conference on Interactive Digital Storytelling*, ser. LNCS 14384. Springer International Publishing, Nov 2023, pp. 137–145.

[12] J. Strong, "Immersive virtual reality and persons with dementia: a literature review," *Journal of gerontological social work*, vol. 63, no. 3, pp. 209–226, 2020.

[13] J. W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi *et al.*, "Testing theory of mind in large language models and humans," *Nature Human Behaviour*, vol. 8, no. 7, pp. 1285–1295, 2024.

[14] N. Shapira, G. Zwirn, and Y. Goldberg, "How well do large language models perform on faux pas tests?" in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 438–10 451. [Online]. Available: https://aclanthology.org/2023.findings-acl.663/

[15] B. Kybartas and R. Bidarra, "A survey on story generation techniques for authoring computational narratives," *IEEE Trans Comput Intell AI Games*, vol. 9, pp. 239–253, 2017. [Online]. Available: https://publications.graphics.tudelft.nl/papers/138

[16] M. Umaschi and J. Cassell, "Storytelling systems: constructing the innerface of the interface," in *Proceedings second international conference on cognitive technology humanizing the information age*. IEEE, 1997, pp. 98–108.

[17] M. Ricardo Bueno Perez and R. Bidarra, "Mixed-initiative story co-creation with TaleMaker," in *Proceedings of FDG*, 2022, pp. 33:1–33:13. [Online]. Available: https://publications.graphics.tudelft.nl/papers/46

[18] T. Bickmore and L. Ring, "Making it personal: end-user authoring of health narratives delivered by virtual agents," in *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*. Springer, 2010, pp. 399–405.

[19] K. Gainsford, B. Fitzgibbon, P. B. Fitzgerald, and K. E. Hoy, "Transforming treatments for schizophrenia: Virtual reality, brain stimulation and social cognition." *Psychiatry research*, vol. 288, p. 112974, 2020.

[20] M. R. Kandalaft, N. Didehbani, D. C. Krawczyk, T. T. Allen, and S. B. Chapman, "Virtual reality social cognition training for young adults with high-functioning autism," *Journal of autism and developmental disorders*, vol. 43, pp. 34–44, 2013.

[21] L. Pieri, G. Tosi, and D. Romano, "Virtual reality technology in neuropsychological testing: A systematic review," *Journal of neuropsychology*, vol. 17, no. 2, pp. 382–399, 2023.

[22] N. C. Tan, J. E. Lim, R. Sultana, J. H. M. Quah, and W. T. Wong, "A virtual reality cognitive screening tool based on the six cognitive domains," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 16, no. 4, p. e70030, 2024.

[23] H. J. Spiers, A. Coutrot, and M. Hornberger, "Explaining world-wide variation in navigation ability from millions of people: citizen science project sea hero quest," *Topics in cognitive science*, vol. 15, no. 1, pp. 120–138, 2023.

[24] J. S. Elkind, "Use of virtual reality to diagnose and habilitate people with neurological dysfunctions," *CyberPsychology & Behavior*, vol. 1, no. 3, pp. 263–273, 1998.

[25] OpenAI, "GPT-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[26] R. P. Me. (2024) Integrate an avatar creator into your game in days - ready player me. [Online]. Available: https://readyplayer.me/

[27] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon, "A survey on audio diffusion models: Text to speech synthesis and enhancement in generative AI," *arXiv preprint arXiv:2303.13336*, 2023.

[28] ElevenLabs. (2024) Free text to speech online with lifelike AI voices — elevenlabs. [Online]. Available: https://elevenlabs.io/text-to-speech

[29] J. Brooke *et al.*, "SUS-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[30] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer speech & language*, vol. 19, no. 1, pp. 55–83, 2005.

[31] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton, "Measuring and defining the experience of immersion in games," *International journal of human-computer studies*, vol. 66, no. 9, pp. 641–661, 2008.

[32] J. M. Rigby, D. P. Brumby, S. J. Gould, and A. L. Cox, "Development of a questionnaire to measure immersion in video media: The film IEQ," in *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, 2019, pp. 35–46.