Towards a Latent Space Cartography of Individual Differences in Subjective Experience Using Large Language Models

Shawn Manuel (shawn.manuel@umontreal.ca)
Department of Psychiatry and Addictology, Université de Montréal
2900, boul. Édouard-Montpetit, Montréal, QC H3T 1J4

Frédéric Gosselin (frederic.gosselin@umontreal.ca)
Department of Psychology, Université de Montréal
2900, boul. Édouard-Montpetit, Montréal, QC H3T 1J4

Jean Gagnon (jean.gagnon@umontreal.ca)
Department of Psychology, Université de Montréal
2900, boul. Édouard-Montpetit, Montréal, QC H3T 1J4

Vincent Taschereau-Dumouchel (vincent.taschereau-dumouchel@umontreal.ca)
Department of Psychiatry and Addictology, Université de Montréal
2900, boul. Édouard-Montpetit, Montréal, QC H3T 1J4

Abstract:

Many theoretical perspectives posit that cognitive and affective functioning is greatly determined by how

individuals subjectively experience the world. However, characterizing the breadth and depth of human experience remains a considerable challenge. One persistent problem is the lack of objective tools for quantifying and comparing narrative reports of subjective experiences. Here, we develop a new approach to map and compare reports of experience using modern large language models (LLMs). Using a series of 20 image prompts, we quantified how the verbal reports of experience provided by participants (n=210) deviate from one another and how these variations are linked to subjective experience and cognitive-affective profiles. We found that latent space embeddings of experience can accurately predict subjective valence and arousal judgments in a series of emotional pictures, as well as cognitive-affective profiles determined using computational factor modeling. As such, latent space cartography of experience could offer a promising avenue for objectively quantifying distortions of subjective experiences and ultimately linking them to patterns of neural activity.

Keywords: latent space cartography, individual differences, subjective experience, large language models

Introduction

Pre-trained artificial neural networks offer a unique opportunity for modeling perceptual representations in the human brain. However, these networks are currently incapable of modeling fine-grained inter-individual variability. This is problematic because human observers often report vastly different experiences when observing the same visual stimulus (see Fig. 1). As such, the activation of internal features of ANNs pre-trained with fixed semantic labels will likely fail to capture all the nuances in experience and to model brain activity accordingly.

Despite this, recent work shows that pre-trained embedding models can leverage the semantic annotation of images to explain brain activity associated with complex conceptual representations beyond mere edges, textures, and categories (Doerig et al., 2022). Indeed, through efficient representation learning, such models develop rich latent representational spaces which have often been compared to the "maps of experience" discussed by neuroscientists, psychologists, and philosophers alike, sometimes as quality spaces (Rosenthal, 2005; Silva, 2020), state spaces or cognitive maps (Whittington et al., 2022; Behrens et al., 2018).

In support of this comparison, previous investigations indicate that latent spaces of ANNs are mostly aligned with the similarity spaces generated by humans (Doerig et al., 2022) and that it is even possible to improve their alignment using human ratings (Muttenthaler et al., 2023). However, the former approach was based on the average *semantic scene description* of an image and not the *individual subjective experience.* By asking participants to directly report what they experience when observing visual stimuli, it should be possible to greatly improve our capacity to model individual brain activity.

This offers the untapped possibility of using the latent spaces of LLMs as a normative space to precisely quantify subjective experience and determine how their distortions along some specific semantic dimensions (e.g., social injustice, see Fig. 1) are related to individual differences in cognition and brain activity. Before we

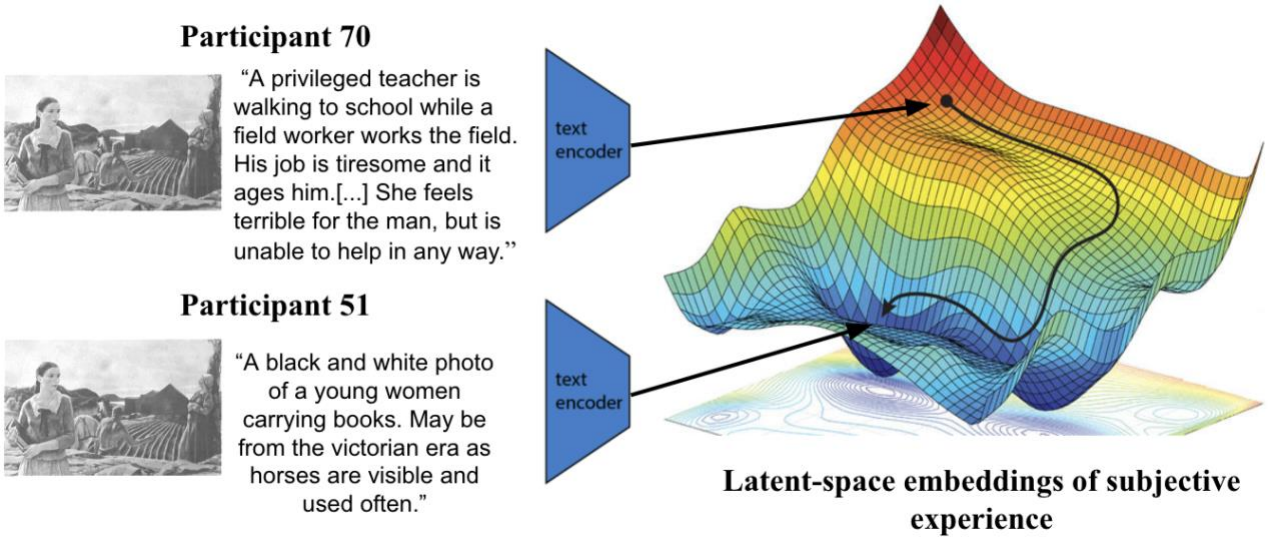

Figure 1. On the left, two subjective reports from participants observing the same visual stimulus. On the right, a conceptual representation of the latent semantic space obtained by extracting word embeddings from verbal reports to an ambiguous visual stimulus.

can use this method for the purpose of brain modeling, we will first validate that the method can capture inter-individual variations in higher-level correlates of subjective experience such as valence and arousal, as well as cognitive-affective functioning.

Methods

210 participants were recruited to write a report of their subjective experience when faced with 20 stimuli, including 10 images from the International Affective Picture System (IAPS)(ref) and 10 cards from the Thematic Apperception Test (TAT)(ref). They were later asked to provide valence and arousal ratings for IAPS images. Participants also completed a series of psychometric inventories shown to capture distinct cognitive and affective trait and symptom profiles related to foundational cognitive abilities such as goal-directedness, decision-making and metacognition (Gillan et al., 2016; Rouault et al., 2018). Text embeddings of verbally reported subjective experience are obtained from OpenAI's GPT-class embedding model '3-large'. Classification models (SVMs) are implemented to predict specific cognitive and affective profiles determined by computational factor modeling. Nested-cross validation, including one thousand random permutations are implemented to assess significance.

Results

Exploratory analyses revealed that classification models trained using subjective report embeddings from IAPS images provided fair performance in predicting subjectively perceived valence (AUC = .87, p = .009) and marginally better than random performance for arousal (AUC = .625, p =.04), though not significant when corrected for multiple comparisons. In line with previous work, factor analysis was applied to the psychometric inventories to reduce shared variance at the item level across questionnaires into three well-replicated latent factors widely used in computational psychiatry and metacognitive computational neuroscience (Rouault et al., 2018; Wise et al., 2023). TAT images provide the best classification performance for higher or lower scores on the 'Mood and impulsivity' factor (AUC = .67; p = .013), which has been associated with deficiencies in metacognitive planning; and the 'Anxiety' factor (AUC = .71; p = .008) associated with increased sensitivity to loss (Fig. 3A, top and bottom respectively), while the IAPS images allow the prediction of the 'Compulsive behaviors and intrusive thoughts' factor (AUC = .65; p = .027), linked to an overly active reward sensitivity mechanism (Fig. 3B).

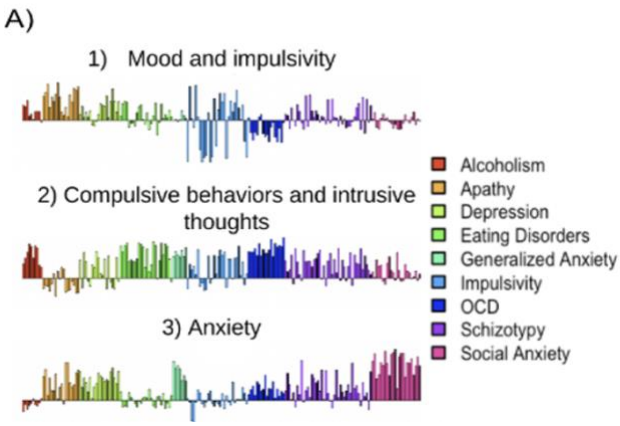

Figure 2. Item loadings for each factor obtained through computational factor modeling, colored by questionnaire of origin.

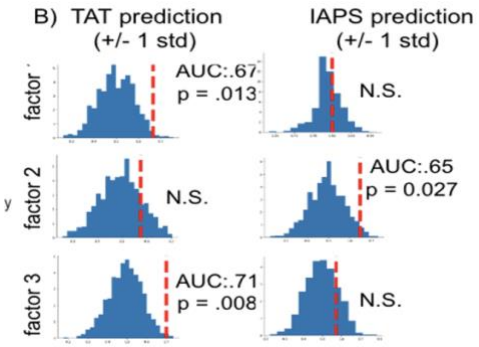

Figure 3. Permutation histograms for classification performance using TAT images (left column) and IAPS images (right column) to predict participants above or below one standard deviation from the mean of each factor.

Conclusion

Objectively quantifying subjective experience still poses a great challenge (Ledoux & Pine, 2016). We propose a new method to map individual differences in subjective experience reports using LLMs. Our methods present the benefit of being both model and construct agnostic, whereby computational factor modeling can be freely applied to any questionnaire or measure and quantitively linked to psycholinguistic markers by an arbitrary word embedding model. Our results support the use of individualized embeddings in controlled experimental settings to represent complex semantic content. Ultimately, we hope to refine brain imaging techniques using this method.

Acknowledgments

S.M. was supported in part by the FRQNT Strategic Clusters Program and UNIQUE (Unifying Neuroscience and Artificial Intelligence – Quebec). V.T-D. was

supported in part by the Canadian Institute of Health Research, and the *Fond de recherche du Québec - Santé* and the *Fondation de l'Institut universitaire en santé mentale de Montréal*.

References

Behrens TEJ, Muller TH, Whittington JCR, Mark S, Baram AB, Stachenfeld KL, et al. What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. Neuron. 2018;100:490–509.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*.

Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. Elife. 2016;5.

LeDoux, J. E., & Pine, D. S. (2016). Using neuroscience to help understand fear and anxiety: a two-system framework. *American journal of psychiatry*, *173*(11), 1083-1093.

Muttenthaler L, Linhardt L, Dippel J, Vandermeulen RA, Hermann K, Lampinen AK, et al. Improving neural network representations using human similarity judgments. arXiv [csCV]. 2023.

Rosenthal D. Consciousness and Mind. Oxford University Press; 2005. 22. Silva L. Towards an Affective Quality Space. Journal of Consciousness Studies. 2023;30:164–195.

Rouault M, Seow T, Gillan CM, Fleming SM. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. Biol Psychiatry. 2018;84:443–451.

Whittington JCR, McCaffary D, Bakermans JJW, Behrens TEJ. How to build a cognitive map. Nat Neurosci. 2022;25:1257–1272.

Wise T, Robinson OJ, Gillan CM. Identifying Transdiagnostic Mechanisms in Mental Health Using Computational Factor Modeling. Biol Psychiatry. 2023;93:690–703.