Received: 6 September 2025 | Accepted: 26 November 2025

DOI: 10.1111/bjop.70047

the british psychological society
promoting excellence in psychology

COMMENTARY

How AI can advance psychological science

Galit Yovel

School of Psychological Sciences, Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel

Correspondence
Galit Yovel, School of Psychological Sciences, Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel.
Email: gality@tauex.tau.ac.il

Funding information
Israel Science Foundation

Abstract

Artificial intelligence (AI) has transformed scientific inquiry across disciplines, including the psychological sciences. In psychology, AI serves not only as an analytic tool but also as a computational model of the very processes the field seeks to explain. In this commentary, I highlight several ways in which AI can advance fundamental questions in psychological science beyond traditional approaches, thanks to its unprecedented ability to generate high-level perceptual and cognitive human-like representations. These developments provide psychologists with powerful new tools that, if embraced, can significantly advance our understanding of the human mind and behaviour.

KEYWORDS

computational modelling, language, methodology, perception

The rapid emergence of artificial intelligence (AI) has transformed scientific inquiry across disciplines, from engineering and the exact sciences to biology, medicine and, more recently, the social sciences and humanities. This transformation is driven by AI's unparalleled capacity to handle vast and complex datasets, uncover hidden patterns and produce predictive models with accuracy that exceeds traditional approaches. These abilities of AI algorithms also benefit the psychological sciences. Yet AI offers psychology an additional unique advantage, serving not only as an analytic tool, but also as a computational model of the very processes that psychology seeks to explain. This dual role opens new avenues for understanding human minds and behaviours, with significant implications for theory, experimentation and application. Below I outline several ways in which the impact of AI on experimental psychology is already emerging and is likely to shape future research.

AI ALGORITHMS AS MODELS OF THE HUMAN BRAIN AND MIND

The first AI models to achieve human-level performance – astonishing both computer and cognitive scientists – were object and face recognition systems (Krizhevsky et al., 2012, Taigman et al., 2014).

These were followed by reinforcement learning algorithms that went on to defeat world champions in chess and Go (Silver et al., 2016), and, more recently, by large language models capable of generating text and communicating in a human-like manner (Brown et al., 2020). These advances raise a central question: Do such systems perform cognitive tasks in ways similar to humans?

Early studies of the visual system have shown strong correspondences between the hierarchical processing of low- to high-level features in deep convolutional neural networks (DCNNs) and the primate visual cortex (Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016). Similar parallels have since been found for behavioural vision effects (e.g. other race effect; Weber's law) (Dobs et al., 2023; Jacob et al., 2021; Phillips & White, 2025, this volume), and studies with large language models also revealed high similarity with human behavioural and neural responses in language tasks (Caucheteux & King, 2022; Goldstein et al., 2022). While these and many other similar findings revealed unprecedented convergence between humans and computational models, clear cases remain where AI diverges from human representations (e.g. Bowers et al., 2023). These discrepancies may reflect key gaps between current algorithms and humans, such as differences in learning processes, architecture and training data, which future research is likely to address (see challenges, limitations and future directions). Importantly, such divergences are not only shortcomings but may highlight unique aspects of human cognition that future models may or may not converge on (e.g. Butlin et al., 2023).

THE ROLE OF EXPERIENCE IN SHAPING HUMAN COGNITION

One of psychology's most fundamental questions is how experience shapes human cognition. However, it is impossible to systematically study and manipulate humans' naturalistic experience in experimental settings to test its causal effect on human behaviour. Here, deep learning algorithms offer a unique opportunity: experimenters can fully control their training data and optimization objectives, and test how systematically changing them affects the emergence of human-like behaviour (Kanwisher, Khosla, et al., 2023). For example, studies show that DCNNs produce human-like face-specific effects when trained to classify faces, but not when trained to classify objects, (Abudarham et al., 2021; Kanwisher, Gupta, et al., 2023; Yovel et al., 2023) supporting the hypothesis that these effects do not simply emerge from domain-general visual computations but instead require machinery specialized for faces. It remains an open question whether the experience provided to deep learning algorithms corresponds only to developmental exposure or also reflects evolutionary factors that manifest as innate human predispositions.

DECOMPOSING THE MIND INTO ITS COMPONENTS

The ability to train deep learning algorithms on specific modalities in isolation can be used to disentangle components of cognition that are hard to separate in biological minds. For example, when humans see the face of a famous person two processes are intertwined: recognizing its visual features and retrieving conceptual knowledge. In AI, these visual and semantic representations can be separated by training models exclusively on images (purely visual representations) or language (purely semantic representations). By calculating the unique variance each model explains in human neural and behavioural data, researchers can determine their relative unique contributions in different cognitive and brain systems (Shoham, Broday-Dvir, et al., 2024; Shoham, Grosbard, et al., 2024). The same logic can be extended to other modalities and systems (e.g. music, voice, speech), offering a framework for dissecting the building blocks of cognition, which are intermixed in human mental representations.

IMPROVING AND ENRICHING PSYCHOLOGICAL MEASURES

AI, and in particular large language models (LLMs), also expands the tools available for measuring psychological constructs. A long-standing challenge in psychological sciences is converting complex phenomena – such as fear or extraversion – into quantitative measures. A common, justified, criticism of experimental psychology is that this reductionism fails to capture the richness and idiosyncrasies of the constructs it seeks to measure: fear, for example, is more than a galvanic skin response (GSR) or a score on an anxiety questionnaire. One of AI's most significant contributions is the ability to transform free-text descriptions into high-dimensional feature vectors that represent their semantic content. These representations enable quantitative analysis of subjective experiences without constraining them to fixed questionnaire items, making it possible to capture nuanced, context-rich and individualized accounts of experience, with particular promise for clinical psychology (Laricheva et al., 2024, this volume), personality research (Jones et al., 2024, this volume), emotion science, creativity (Kern et al., 2024, this volume) and other fields where traditional simplified measures often fail to convey the full complexity of the phenomena being studied.

CHALLENGES, LIMITATIONS AND FUTURE DIRECTIONS

While these examples illustrate the many ways AI can advance psychological science, there are still important challenges and unanswered questions. A common criticism is that using AI to understand the mind is like replacing one black box with another. Although we can train these systems to perform tasks at human-level performance, we do not fully understand the representations they form or the mechanisms they use to solve the task. Yet, psychologists' expertise in studying the black box of the human mind for the past century can be valuable for uncovering the black box of AI algorithms. Visualization techniques, analysis of representational geometry and controlled manipulations (e.g. ablations, training data) are some of the ways that can shed light on the nature of AI representations and their correspondence to human cognition. Moreover, AI explainability is a rapidly evolving field of research, and insights from it can directly inform the study of the brain and mind (Qi et al., 2024; Sano et al., 2024; Soydaner & Wagemans, 2024, this volume). Many of the current gaps between AI and humans are likely to narrow as advances in training, architectures, multimodality, embodiment and inter-agent communication (He et al., 2024, this volume) make AI more human-like. AI holds the potential to fundamentally advance psychological science by providing rigorous, scalable tools and computational models capable of elucidating the mechanisms underlying high-level human cognition.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID


Galit Yovel https://orcid.org/0000-0003-0971-2357

REFERENCES

Abudarham, N., Grosbard, I., & Yovel, G. (2021). Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition. *Cognitive Science*, *45*(9), e13031. https://doi.org/10.1111/cogs.13031

Bowers, J. S., Malhotra, G., Dujmović, M., Llera Montero, M., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385. https://doi.org/10.1017/S0140525X22002813

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., & Henighan, T. (2020). Language models are few-shot learners. *arXiv*.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*. https://doi.org/10.48550/arXiv.2308.08708

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2023). Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(32), e2220642120. https://doi.org/10.1073/pnas.2220642120

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

He, J. K., Wallis, F. P. S., Gvirtz, A., & Rathje, S. (2024). Artificial intelligence chatbots mimic human collective behaviour. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12764

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, *12*(1), 1872. https://doi.org/10.1038/s41467-021-22078-3

Jones, A. L., Shiramizu, V., & Jones, B. C. (2024). Decoding the language of first impressions: Comparing models of first impressions of faces derived from free-text descriptions and trait ratings. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12717

Kanwisher, N., Gupta, P., & Dobs, K. (2023). CNNs reveal the computational implausibility of the expertise hypothesis. *iScience*, *26*(2), 105976. https://doi.org/10.1016/j.isci.2023.105976

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254. https://doi.org/10.1016/j.tins.2022.12.008

Kern, F. B., Wu, C., & Chao, Z. C. (2024). Assessing novelty, feasibility and value of creative ideas with an unsupervised approach using GPT-4. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12720

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Laricheva, M., Liu, Y., Shi, E., & Wu, A. (2024). Scoping review on natural language processing applications in counselling and psychotherapy. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12721

Phillips, P. J., & White, D. (2025). The state of modelling face processing in humans with deep learning. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12794

Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2024). Explanation strategies in humans versus current explainable artificial intelligence: Insights from image classification. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12714

Sano, T., Shi, J., & Kawabata, H. (2024). The differences in essential facial areas for impressions between humans and deep learning models: An eye-tracking and explainable AI approach. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12744

Shoham, A., Broday-Dvir, R., Yaron, I., Yovel, G., & Malach, R. (2024). Text-related functionality of visual human pre-frontal activations revealed through neural network convergence. *Communications Biology*, *8*(1), 1129. https://doi.org/10.1101/2024.04.02.587774

Shoham, A., Grosbard, I. D., Patashnik, O., Cohen-Or, D., & Yovel, G. (2024). Using deep neural networks to disentangle visual and semantic information in human perception and memory. *Nature Human Behaviour*, *8*(4), 702–717.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Soydaner, D., & Wagemans, J. (2024). Unveiling the factors of aesthetic preferences with explainable AI. *British Journal of Psychology*. https://doi.org/10.1111/bjop.12707

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701–1708).

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yovel, G., Grosbard, I., & Abudarham, N. (2023). Deep learning models challenge the prevailing assumption that face-like effects for objects of expertise support domain-general mechanisms. *Proceedings of the Royal Society B: Biological Sciences*, *290*, 20230093. https://doi.org/10.1098/rspb.2023.0093

How to cite this article: Yovel, G. (2025). How AI can advance psychological science. *British Journal of Psychology*, *00*, 1–4. https://doi.org/10.1111/bjop.70047