The I in AI Is More Than a Benchmark

Vojtěch Formánek, 12.5.2025

Faculty of Informatics, Masaryk University

Recent developments in Large Language Models would lead us to believe that we've reached Artificial General Intelligence (AGI), with OpenAi o3 beating a well-regarded ARC-AGI benchmark, which was inspired by the psychological tests of Intelligence (Chollet, 2019). However, this enforces the current AI paradigm, that bases intelligence on emulation of biological neural architectures and falls short of replicating the subjective mental and emotional states that are core to human cognition. What benchmarks such as ARC-AGI don't consider is the many problems with the construct of intelligence in Psychology, especially its measurement. IQ Tests such as WAIS might not reflect the construct (Van der Maas, Kan, & Borsboom, 2014). "*What and IQ test measures is an IQ test*" is a frequently used phrase, that light-heartedly emphasizes the real issue of mapping the results of Intelligence Tests to the construct of Intelligence.

This creates a large gap between the measurement of Intelligence (the I) in AI and the construct as well. Zhao et al. (2022) argue that to bridge it, AI must integrate principles from cognitive psychology. Especially its cognitive and emotional branches, should be foundational to future AI systems, enabling machines to interpret and react to the world in more human-like ways. The prevailing model of AI tends to focus on simulating neural processes without accounting for the complexities of psychological experience. For example, unlike humans, whose memory operates passively and is shaped by emotion and attention, machine memory is based on active deletion and storage mechanisms, which diverge significantly from human cognitive patterns (Zador, 2019). Cognitive psychology offers a means of reconciling these differences by introducing behavioural models that can structure machine learning in more human-compatible ways. The authors advocate for AI development that accounts for the way humans internalize knowledge, express emotions, and adapt their mental states through interaction and learning (Kriegeskorte & Douglas, 2018).

Related Work

Van der Maas, Kan, and Borsboom (2014) revisited the long-dismissed notion (Boring, 1923) that "*what an IQ test measures is an IQ test*" . Traditionally the statement has been viewed as simplistic, however they ultimately reach the conclusion that it does, to an extent, reflect the current paradigm of Intelligence testing.

Central to their argument is the distinction between two models of intelligence measurement: reflective models (such as those used in factor analysis) and formative models (such as those found` in principal components analysis). Reflective models make use of a latent variable: g, or general intelligence. It is assumed to causally influence test scores across a range of cognitive abilities (Jensen, 1999). By contrast, formative models treat g not as a causal factor but as a statistical index, but as a weighted sum of test scores that does not imply the existence of an underlying latent trait (Markus & Borsboom, 2013).

The mutualism model, originally proposed by van der Maas et al. (2006), challenges the reflective interpretation of g. Rather than assuming a single underlying cause, it posits that cognitive abilities like memory, language, and spatial reasoning develop through dynamic, reciprocal interactions. These interactions themselves produce the consistent pattern of positive correlations across subtests, without invoking an underlying g-factor.

This shift has major implications for both theory and practice. If the mutualism model is correct, then g does not reflect an underlying biological or genetic cause (Chabris et al., 2012). Instead, intelligence test scores should be viewed as index variables that summarize observed abilities (e.g. school performance) without assuming a latent causal source, much like indexes of health or economic performance (Howell, Breivik, & Wilcox, 2007).

The mutualism framework also impacts how intelligence tests are constructed and interpreted. Under a reflective model, test items are interchangeable indicators of g, and measurement improves by increasing the number of items with high factor loadings. In a formative model, however, the selection of indicators fundamentally defines the construct. Thus, there is no uniquely correct way to measure intelligence, only pragmatically useful ways, depending on the predictive goals, e.g. school performance (Edwards & Bagozzi, 2000).

Foundations and Evolution of Cognitive-Inspired AI

The incorporation of cognitive psychology into AI development is not new. In the 1980s, Japanese researchers introduced “Kansei Engineering,” which sought to quantify human sensibility and psychological perception for use in engineering applications (Ali et al., 2020). Wang Zhiliang later proposed the concept of “artificial psychology,” extending the scope of psychological modelling in machines to include both low-level and high-level mental functions. This conceptual framework allowed researchers to design systems capable of simulating human responses to environmental stimuli (Zhao et al., 2022).

Key milestones in this trajectory include Marvin Minsky’s seminal “Society of Mind” theory (Auxier, 2006), which modelled intelligence as the collective output of simple cognitive agents. More recently, DeepMind introduced the “Theory of Mind” concept in machine learning, emphasizing the ability of AI agents to predict and interpret the mental states of others (Rabinowitz et al., 2018). Tools like PsychLab (Leibo et al., 2018) and developments in explainable AI (Taylor & Taylor, 2021) have extended these ideas by using experimental psychological paradigms to evaluate AI performance and interpretability.

They reflect an important trend: rather than relying solely on data and pattern recognition, modern AI research is increasingly oriented toward understanding the *processes* by which intelligent agents perceive, infer, and react – an area where cognitive psychology provides indispensable insight (Miller, 2019). Zhao et al. (2022) outline three major applications that demonstrate how cognitive psychology enhances AI’s ability to simulate

human intelligence: facial attractiveness prediction, affective computing, and musical emotion modelling.

Facial Attractiveness

Facial attractiveness is a psychologically rich domain often believed to be entirely subjective. However, studies reveal high consistency in attractiveness judgments across cultures, suggesting a degree of universality in human aesthetic perception (Han et al., 2020). Based on this insight, researchers developed datasets such as SCUT-FBP5500, which includes facial images annotated with attractiveness ratings and detailed feature points (Liang et al., 2018). Deep learning models like ResNet-18 and ResNeXt-50 trained on this data achieved high correlation with human ratings, with Pearson coefficients exceeding 0.85.

Subsequent research employed multi-task learning frameworks to refine predictions by combining facial beauty assessment with gender and race classification tasks, resulting in increased accuracy (Vahdati & Suen, 2021). Meta-learning approaches further enabled models to learn individualized preferences from limited samples, demonstrating effective adaptation across demographic groups (Lebedeva et al., 2022). The authors also describe their own contributions, including the construction of large-scale facial databases segmented by age, gender, and ethnicity, and the development of geometric models that integrate face structure and skin texture for more holistic attractiveness evaluations (Zhao et al., 2019a; Zhao et al., 2020). These findings affirm that machine learning, when informed by psychological principles, can approximate subjective human judgments with surprising precision.

Affective Computing

Affective computing focuses on equipping machines with the ability to recognize and simulate emotions. This area is inspired by research indicating that decision-making and intelligence are deeply intertwined with emotion. Patients with impairments in the limbic system, for instance, often demonstrate poor judgment despite intact logical reasoning (Bechara et al., 2000).

Pioneered by Picard (1995), affective computing integrates inputs from facial expression (FACS; Ekman, 1972), speech patterns, body posture, and neurophysiological data such as EEG to model emotional states. Using DRML (deep region and multi-label learning) and bidirectional LSTM with directed self-attention, researchers have developed models that achieve over 70% accuracy in speech and facial emotion recognition across benchmark datasets like IEMOCAP and EMO-DB (Li et al., 2021).

Electroencephalography (EEG)-based affective computing also shows promise. The DEAP dataset, for example, links EEG readings with emotional responses to music videos, achieving recognition rates as high as 98% for certain emotions (Luo et al., 2020). Zhao et al. (2019d) developed a multimodal system combining speech and facial features to assist in diagnosing depression. Enhanced speech signals, when fused with facial expression data, improved diagnostic accuracy to over 82%. These methods not only enhance human–machine interaction but also open new avenues in mental health care, allowing for automated and continuous emotional monitoring with clinical implications.

Music Emotion

Music has a well-documented impact on human emotions. Cognitive and neuroscientific studies reveal that music activates brain regions linked to reward and emotional processing, including the amygdala and nucleus accumbens (Rahman et al., 2021). Zhao et al. (2022) describe the development of an affective brain–computer music interface (aBCMI) that detects a user's emotional state through EEG and physiological signals, then selects or generates music to shift the individual toward a target emotional state.

Experimental results showed the system could achieve meaningful classification of user arousal and valence levels and successfully induce desired mood states (Daly et al., 2016). Additional applications include using AI-based music instruction to aid in therapy, particularly for patients with psychological or neurological impairments (Li & Liu, 2022).

This research highlights how music, when analyzed through the lens of cognitive psychology and implemented with AI tools, can regulate affective states and facilitate personalized mental health interventions.

While cognitive psychology offers profound insights into human intelligence, applying it to artificial systems is not without challenges. One key difficulty lies in the subjectivity and variability of psychological traits. Emotional and aesthetic judgments are shaped by cultural, social, and individual experiences, complicating efforts to model them universally (Zhao et al., 2022). Furthermore, mental states are inherently dynamic and often ambiguous, defying the precise quantification required by computational models.

To mitigate these issues, the authors advocate for expanding interdisciplinary research in areas such as big data medicine, brain–computer interfaces, and general artificial intelligence. Future systems should emphasize multimodal data fusion and high-dimensional feature extraction to better capture the complexity of human cognition. As AI and psychology continue to evolve together, their integration promises to yield more intelligent, empathetic, and socially adaptive machines.

Projective tests

IQ tests constitute a famous example of the benchmarking paradigm in Psychometrics. The Thematic Apperception Test (TAT) and the Rorschach Inkblot Test (Meyer, 2004) are two of the most well-known examples of another paradigm – projective assessments. They are designed to uncover underlying thoughts, feelings, and personality traits by eliciting interpretive responses to ambiguous stimuli. In the TAT, individuals are shown a series of evocative, often emotionally charged images and asked to create stories about what is happening, what led up to the scene, and what might happen next. These narratives reveal internal motivations, conflicts, and relational dynamics. Similarly, the Rorschach test presents subjects with abstract inkblot images and asks them what they see, encouraging spontaneous interpretation. The idea behind both tests is that when faced with ambiguity, people project their unconscious thoughts and emotions onto the stimulus, thereby externalizing inner psychological structures. Though controversial in terms of reliability and validity, these tests remain valuable for their unique ability to access subjective and often inaccessible aspects of human cognition and affect.

Methods

To meaningfully assess whether artificial intelligence systems demonstrate human-like intelligence, we propose a new approach, rooted in the projective methodologies. Our approach emerges from a fundamental limitation in both traditional psychometrics and contemporary AI evaluation: the inability of structured, quantitative tests to access or characterize internal, subjective mental states. This limitation is particularly critical given the centrality of such states—motivation, intention, affect, self-reflection—to any robust conception of human intelligence. A consequence of this is that we need to acknowledge them as unique forms of something akin to intelligence and instead focus on their internal workings.

In psychology, projective tests such as the Rorschach Inkblot Test or the Thematic Apperception Test have historically been employed not because of their psychometric elegance, but because of the unique epistemological stance they embody. Projective methods do not aim to quantify traits directly. Instead, they seek to elicit patterns of response that reveal underlying cognitive and affective structures. This indirect probing of mental content—by allowing the subject to structure and make sense of ambiguous stimuli—offers a model of assessment where interpretation itself becomes the data. XAI are not projective tests, they are too low-level, they embody neurophysiological measurements such as EEG or ECG instead. The insights they provide likely cannot give us the insights we need.

The methodological value of projective testing lies in the assumption that individuals cannot help themselves to impose meaning on ambiguity. In doing so, they externalize cognitive templates, affective predispositions, and perceptual biases. Applied to AI models, this principle holds promise precisely because it bypasses surface-level performance and probes for internal structures. The critical question is not whether an AI system can generate a correct or expected output, but whether it can generate a coherent, context-sensitive, and semantically rich interpretation when faced with ambiguity. An interpretation that can reveal representational depth.

Current AI benchmarks, such as those measuring language comprehension or visual reasoning, overwhelmingly rely on closed-ended tasks or narrowly defined goals. These are insufficient for evaluating whether AI systems possess any capabilities such as internal modelling, perspective-taking, or spontaneous conceptual elaboration. Projective-style methods can be challenging to AI systems, because they require generating and sustaining internally coherent responses under uncertainty and with minimal structure – conditions that resemble real-world cognition more closely.

We propose assessing that AI systems through tasks structurally modelled on projective tests: narrative completions, ambiguous image descriptions, ethical dilemmas, or hypothetical social scenarios in which multiple interpretations are plausible, and no definitive answer exists. The system's response would be evaluated not by correctness but by dimensions such as thematic integration, emotional coherence, perspective consistency, and generalization from prior knowledge. These dimensions mirror constructs long understood in psychology as markers of cognitive and emotional complexity.

Critically, such evaluations must be analysed qualitatively and interpreted in relation to other concepts and contexts. Just as responses on projective tests are never read in isolation but understood as part of a pattern across multiple domains. Moreover, this approach accounts for the inherent immeasurability of internal states in AI systems. Just as we cannot directly observe

the mental content of a human subject but infer it through expressive behaviour under ambiguity, we can analogously infer the structural properties of AI cognition through its behaviour when constructing meaning without external help. A model that consistently generates emotionally congruent narratives, adopts coherent points of view, or displays of preference-like behaviour under ambiguity suggests a level of internalization beyond pattern matching.

This methodology deliberately resists overquantification. Quantitative evaluation, while indispensable in many domains, is ill-suited for phenomena where the primary data are patterns of organization, not scalar outcomes. Instead, a projective framework embraces ambiguity as a methodological strength, recognizing that how a system handles uncertainty is often more revealing than how it performs under constraint.

Finally, drawing from psychometrics, we suggest this approach be domain-specific rather than global. Just as intelligence tests are most valid when used to support diagnostic or educational interpretation within a bounded context, projective paradigms for AI should be applied to assess capabilities within well-defined cognitive or emotional domains (e.g., narrative reasoning, moral judgment, aesthetic preference), rather than as sweeping claims about AGI.

Adopting projective-style methods to evaluate AI shifts the focus from performance metrics to the structural and interpretive characteristics. It aligns evaluation with the theoretical assumption that intelligence involves the capacity to structure the world meaningfully under uncertainty – something best assessed not through task success alone, but through patterns in interpretation. This approach reflects the complexity of cognition.

Discussion

We've outlined the case to use projective-method-inspired frameworks to assess artificial intelligence, rather than relying on traditional, quantitative evaluations. To focus on subjective interpretations to effectively assess whether artificial intelligence systems exhibit human-like intelligence. Current AI benchmarks inadequately capture internal mental states such as motivation, affect, or self-reflection, which are essential to a full understanding of intelligence. Drawing from psychology, projective tests like Rorschach, fill this role. They are not used for their quantifiability, but for their ability to elicit subjective, interpretive responses. We've argued that AI should similarly be evaluated on its capacity to construct coherent, contextsensitive interpretations under ambiguous conditions. This approach focuses the underlying representational structures rather than assessing performance accuracy.

The proposed approach could evaluate AI through open-ended, ambiguous tasks, such as narrative completions or ethical dilemmas, where multiple interpretations are possible and no single answer is correct. Evaluation focuses on dimensions like emotional coherence, thematic integration, and generalization, with responses interpreted qualitatively and relationally across tasks. This method resists overquantification and acknowledges that internal cognitive organization in AI, much like in humans, can only be inferred through behaviour under uncertainty. The projective paradigm offers a more nuanced and domain-specific lens for AI assessment, highlighting the system's ability to structure meaning rather than merely execute programmed outputs.

Conversely, this doesn't exclude the room for benchmarking approaches in AI. They remain invaluable for domain-specific tasks, where the correct answers are clearly defined. Benchmarks such as ARC-AGI should be used to compare artificial agents between each other. However, based on almost a century of research on Intelligence in Psychology, we believe that they are not suited for detecting whether the agents pose Intelligence themselves. Such as their counterparts in Psychology – IQ tests cannot tell us whether an agent, human or AI, is Intelligent – it is not clear what that means. Instead, we propose to focus our efforts in a more immediately available domain – qualitatively evaluating the AI's mental states and subjective reasoning.

Acknowledgements

The author used OpenAI ChatGPT-4.5 for article summarization, stylistic improvement of the text. The author checked the contents and takes full responsibility for the text.

References

Ali, S., Wang, G., & Riaz, S. (2020). Aspect-based sentiment analysis of ride-sharing platform reviews for Kansei engineering. *IEEE Access, 8*, 173186-173196. https://doi.org/10.1109/ACCESS.2020.3025823

Auxier, R. E. (2006). The pluralist: An editorial statement. *The Pluralist*, v–viii.

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex, 10*(3), 295–307. https://doi.org/10.1093/cercor/10.3.295

Boring, E.G. (1923). *Intelligence as the tests test it.* The New Republic, 36, 35–37.

Buhari, A. M., et al. (2020). FACS-based graph features for real-time micro-expression recognition. *Journal of Imaging, 6*(12), 130. https://doi.org/10.3390/jimaging6120130

Chabris, C.F., Hebert, B.M., Benjamin, D.J., Beauchamp, J., Cesarini, D., van der Loos, M., & Laibson, D. (2012). *Most reported genetic associations with general intelligence are probably false positives.* Psychological Science, 23, 1314–1323.

Daly, I., et al. (2016). Affective brain–computer music interfacing. *Journal of Neural Engineering, 13*(4), 046022. https://doi.org/10.1088/1741-2560/13/4/046022

Dickens, W.T., & Flynn, J.R. (2001). *Heritability estimates versus large environmental effects: The IQ paradox resolved.* Psychological Review, 108, 346–369.

Edwards, J.R., & Bagozzi, R.P. (2000). *On the nature and direction of relationships between constructs and measures.* Psychological Methods, 5, 155–174.

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Khateeb, M., Anwar, S. M., & Alnowami, M. (2021). Multi-domain feature fusion for emotion classification using DEAP dataset. *IEEE Access, 9*, 12134–12142. https://doi.org/10.1109/ACCESS.2021.3051281

Howell, R.D., Breivik, E., & Wilcox, J.B. (2007). *Is formative measurement really measurement?* Psychological Methods, 12, 238–245.

Jensen, A.R. (1999). *The g Factor: The Science of Mental Ability.* Westport, CT: Praeger.

Kan, K.J., Wicherts, J.M., Dolan, C.V., & van der Maas, H.L.J. (2013). *On the nature and nurture of intelligence and specific cognitive abilities: The more heritable, the more culture dependent.* Psychological Science, 24, 2420–2428.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature*

Neuroscience, 21, 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Lebedeva, I., Ying, F., & Guo, Y. (2022). Personalized facial beauty assessment: A metalearning approach. *The Visual Computer*, 1–13. https://doi.org/10.1007/s00371-021-02387-w

Leibo, J. Z., d'Autume, C. D. M., Zoran, D., Amos, D., Beattie, C., Anderson, K., et al. (2018). Psychlab: A psychology laboratory for deep reinforcement learning agents. arXiv [Preprint].

arXiv:1801.08116,

Li, D., Liu, J., Yang, Z., Sun, L., and Wang,Z.(2021).Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Syst. Appl. 173:114683. doi: 10.1016/j.eswa.2021.114683

Li & Liu. (2022). Design of an incremental music teaching and assisted therapy system based on artificial intelligence attention mechanism. *Occupational Therapy International*, 2022, 7117986. https://doi.org/10.1155/2022/7117986

Liang, L., et al. (2018). SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. *Proceedings of the 2018 International Conference on Pattern Recognition (ICPR)*, 1598–1603. https://doi.org/10.1109/ICPR.2018.8546038

Luo, Y., et al. (2020). EEG-based emotion classification using spiking neural networks. IEEE Access, 8, 46007–46016. https://doi.org/10.1109/ACCESS.2020.2978163

Markus, K., & Borsboom, D. (2013). *Frontiers of Validity Theory: Measurement, Causation, and Meaning.* New York, NY: Routledge.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. (2018). “Machine theory of mind,” in Proceedings of the international conference on machine learning (Orlando, FL: PMLR), 4218–4227.

Meyer, G. J. (2004). The reliability and validity of the Rorschach and Thematic Apperception Test (TAT) compared to other psychological and medical procedures: An analysis of systematically gathered evidence. *Comprehensive handbook of psychological assessment*, *2*, 315-342.

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review, 28*(2), 454–475. https://doi.org/10.3758/s13423-020-01825-5

van der Maas, H.L.J., Dolan, C.V., Grasman, R.P.P.P., Wicherts, J.M., Huizenga, H.M., & Raijmakers, M.E.J. (2006). *A dynamical model of general intelligence: The positive manifold of intelligence by mutualism.* Psychological Review, 113, 842–861.

Vahdati, E., & Suen, C. Y. (2021). Facial beauty prediction from facial parts using multi-task and multi-stream convolutional neural networks. *International Journal of Pattern Recognition and Artificial Intelligence, 35*(2), 2160002. https://doi.org/10.1142/S0218001421600028

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications, 10*, 3770. https://doi.org/10.1038/s41467-01911786-6

Zhao, J., Cao, M., Xie, X., Zhang, M., and Wang, L. (2019a). Data-driven facial attractiveness of Chinese male with epoch characteristics. IEEE Access 7, 10956–10966. doi: 10.1109/ACCESS.2019.2892137

Zhao, J., Su, W., Jia, J., Zhang, C., and Lu, T. (2019c). Research on depression detection algorithm combine acoustic rhythm with sparse face recognition. *Cluster Comput*. 22, 7873–7884. doi: 10.1007/s10586-017-1469-0

Zhao, J., Zhang, M., He, C., Xie, X., and Li, J. (2020). A novel facial attractiveness evaluation system based on face shape, facial structure features and skin. Cogn. Neurodynamics 14, 643–656. doi: 10.1007/s11571-020-09591-9

Zhao, J., Wu, M., Zhou, L., Wang, X., & Jia, J. (2022). Cognitive psychology-based artificial intelligence review. *Frontiers in neuroscience*, *16*, 1024316.

OpenAI. (2025). *ChatGPT* (Feb 27) [Large language model]. https://chat.openai.com/