A Total Brain Framework for AGI

April 2026

Authors: **Evian Gordon**, MBBCh PhD[1*], **Donna Palmer**, PhD[1], **Anne Clarke**, BS[1]

[1] Total Brain, A SonderMind Company

*Correspondence: evian.gordon@totalbrain.com.
Founder of Total Brain and Total Brain International Database
- First Standardized and Integrated Human Brain Database.

ABSTRACT

Human Brain Cognition and Emotion are becoming an increasingly present source of inspiration within the AGI field. This shift is visible across frontier labs and leading researchers. In the past half year, DeepMind proposed a cognitive framework for AI, Anthropic identified emotion concepts that impact model performance and reasoning, and Yann LeCun parted ways with Meta where he had served as chief AI scientist, motivated by his belief that current methods are not sufficient and that human-level intelligence requires cognitive architectures that have more in common with perception-action loops, predictive processing, and neurobiology. This paper presents the **Total Brain Integrative Intelligence (TBII-AGI) Framework** as a complementary alternative framework to expand and deepen the direction that the field is going in. The perspective is derived from decades of applied neuroscience with data from the first standardized and integrative Total Brain International Brain Database. The data spans human cognitive, emotional, and neurophysiological data (EEG, ERP, HRV), neuroimaging (MRI, DTI, fMRI), encompassing 7 billion datapoints, in 15 validated paradigms across the cognitive-emotion spectrum, including: habituation, startle, attention, memory, emotion regulation, executive function, and markers across the normal, mental health, and peak performance continuum.

The TBII-AGI Framework proposes key components for brain-aligned AGI that include emotion-cognition interactions, unconscious cognitive processes, and cognitive bias, human signatures for all-around high performance that is akin to AGI. Whole-brain with high temporo-spatial and performance single-trial granularity, illuminate real-time dynamics. We provide four database exemplars. 1, Demonstrates ‘smooth’ and ‘jagged’ human performance models akin to the Google DeepMind Cognitive Taxonomy for testing AGI ability. 2. The impact of emotion mindsets on cognitive efficiency and similarities in Anthropic’s Emotion Concepts and LLM performance reporting. 3. Similar cognitive bias in the human database and a VLM model on the same task. 4. Brain ERP components to learning in an executive function hidden maze task akin to some of the ARC-AGI-3 benchmarks. The TBII-AGI Framework is presented as a positive extension and exemplar blueprint for any AGI evaluation, that mirrors the full architecture of human unconscious-conscious intelligence in real-world conditions.

Keywords: *AGI Testing, AGI development, ARC-AGI-3, cognition, cognitive taxonomy, Total Brain International Database, adaptive intelligence, human neuropsychology, brain dynamics, conscious, unconscious, emotion, social cognition, bias regulation, executive function.*

1. Introduction

The race to develop true Artificial General Intelligence (AGI) comprises the forefront of current AI development, to build AI with the ability to understand, learn, and flexibly apply its intelligence to any intellectual task that a human being can perform. This extends beyond the capabilities of current models, which are very good at learning individual tasks, but poor at generalizing those capabilities to new challenges or contexts.

AGI development is currently experiencing a period of pivot and change. Recent years have been dominated by the Scaling Hypothesis - that AGI could be achieved by adding more data and compute to Transformer-based models (Kaplan et al., 2020). While this approach has yielded AI models that are, in many ways, highly capable, it has to date failed to produce true general intelligence, and developers are now hitting a plateau of diminishing returns in the AGI field. As a result, the industry is undertaking new approaches, including pivots towards architectures that are informed by principles of the human brain function. Google DeepMind has recently established a new cognitive taxonomy framework for AGI testing based on human cognition benchmarks (Burnell et al., 2026). Demis Hassabis has stated that ideas from neuroscience will be increasingly indispensable as both a roadmap for AI research and a source of computation tools in AI development (Hassabis et al., 2017). Anthropic's research shows that emotion-like processes act as causal control systems within AI model performance, rather than merely surface features (Sofroniew et al., 2026). Meta's former Chief AI Scientist Yann LeCun states that achieving human-level AGI requires machines to mimic the way that human brains build common sense, persistent memory, planning, and reasoning from physical interaction with the world - capabilities that current LLMs fundamentally lack (LeCun, 2022).

The current approaches to AI development and evaluation have produced systems of remarkable capability, yet have not given rise to the kind of flexible "integrative intelligence" that characterizes human capabilities, or to true artificial general intelligence. An expanded human-brain-based view of what constitutes general intelligence has never been more prescient or more applicable to the next era of AI development.

1.1 Cognitive Capacities of AI

François Chollet's ARC-AGI benchmark series for evaluating AGI capabilities operationalizes general intelligence as the efficient acquisition of novel cognitive skills from minimal training signals (Chollet, 2019). The subsequent ARC-AGI-2 innovation lies in making memorization impossible: each task presents a novel rule, and success requires in-context generalization. Top scores currently plateau at around 20–50%, underscoring the persistent gap between current AI and human-level generalization (Chollet et al., 2025). ARC-AGI-3, released in 2026, translates this taxonomy into interactive, game-like maze environments in which agents must plan, model, explore, and adapt in real time. Current frontier models score below 1% of human performance (ARC Prize Foundation, 2026).

Google DeepMind's recently released cognitive framework for AI, "Measuring Progress Toward AGI: A Cognitive Taxonomy," describes today's AI systems as fundamentally jagged intelligences - being very good at some things, but very poor at others (Burnell et al., 2026). A classic example of this is ChatGPT-4o being able to perform at the 90th percentile on the bar

exam, but being unable to count the number of “r”s in the word “strawberry”. Google DeepMind’s Cognitive Taxonomy approaches the goal of general intelligence by organizing AGI evaluation around ten core cognitive faculties that are important for human intelligent behavior (2026). These are faculties that are difficult to probe with static, text-based benchmarks and relevant to human-compatible intelligence in real environments: *perception, generation, attention, learning, memory, reasoning, metacognition, executive functions, problem-solving, and social cognition* (Burnell et al., 2026).

Critically, the taxonomy is candid about its scope: it addresses the “*what”* of AGI cognition, the functional requirements of a generally intelligent system, rather than the “how”, the underlying factors that give rise to these cognitive capacities. It is also limited in scope to evaluating AGI's progress relative to human cognitive capacities and does not encompass insights from human general intelligence that can inform the *development* of artificial systems for general intelligence. The scope is also limited to cognitive processes and does not encompass emotion-related capacities, with social cognition included primarily because of the complexity of the cognitive processes involved, rather than as a broader exploration of the interplay between cognition and emotion.

1.2 Emotion and AI

Relationships between emotion and cognition are emerging as being highly relevant for large language models (LLMs) and their cognitive performance. LLMs can appear to exhibit emotional reactions, and their behavior can change in ways somewhat similar to those expected in humans (Sofroniew et al., 2026). While these emotions are presumably learned from the large array of human content the models have been trained on, Anthropic has recently reported on a wide array of situations in which Claude's ‘cognitive’ behavior and decisions were impacted by the concurrent emotions being expressed (Sofroniew et al., 2026). For example, artificially steering Claude toward desperation increased the likelihood that it would offer a seemingly reasonable solution to an unsolvable programming task, a behavior known as reward-hacking. Conversely, increasing Claude's activation in the "calm" direction reduced this reward-hacking behavior (Sofroniew et al., 2026). As Jack Lindsey, head of Anthropic's "Model Psychiatry," has put it: "What surprised us was how significantly Claude's behavior is routed through the model's emotion representations." (Knight et al., 2026).

1.3 The Brain and AI

There is an emerging recognition that human brain function could inform new approaches to AGI development. A prominent recent example is Yann LeCun's November 2025 departure from his 12-year role as chief AI scientist Meta in order to pursue human-level AI, which LeCun believes requires a radically different form of training than current models - one informed by ideas such as predictive processing and perception–action loops, and analogous to specific regions of the brain: “The perception module corresponds to the visual, auditory, and other sensory areas of the cortex, as well as areas in the association cortex …the intrinsic cost module corresponds to structures in the basal ganglia involved in rewards, including the amygdala...The configurator may correspond to structures in the prefrontal cortex that perform executive control and modulate attention” (LeCun, 2022).

This aligns with the origins of AI design and its building blocks being in neural network architecture. Similar to cognitive science and neuroscience's mapping of neural network nodes and architecture designs that 'store' particular concepts, Anthropic's Sparse Autoencoders approach aims to identify interpretable features within LLMs associated with particular concepts (Sofroniew et al., 2026). OpenAI's Sam Altman has also framed the entire AI endeavor as "building a brain for the world" (Altman, 2025). Google DeepMind's Demis Hassabis has long held that ideas from neuroscience can be indispensable as both a roadmap for AI research and a source of computation tools in AI development (Hassabis et al., 2017).

1.4 Total Brain Integrative Intelligence Framework for AGI (TBII-AGI)

The current framework paper expands and deepens these approaches with the *Total Brain Integrative Intelligence Framework for AGI (TBII-AGI).* This is a complementary expansion from the perspective of the first *Standardized Integrative International Human Brain Database (Total Brain Database[1])*. Assessments span more than 15 cognitive-emotion capacities across the spectrum of unconscious, conscious, and peak-performance aspects of cognitive function in a 7-billion-data-point normative and clinical database. The TBII-AGI Framework draws on this dataset and decades of peer-reviewed neuroscience (over 300 publications) to propose new evaluation dimensions that the current ARC-AGI-3 and Google DeepMind Cognitive Taxonomy and other AGI architectures do not fully incorporate.

Traditional AI and LLM research and benchmarks do not place enough emphasis on understanding, comparing, and translating human biological and artificial dynamical information processing. Despite humans possessing an efficiency of learning and generalizability of acquired skills that AI models are not capable of.

The Total Brain Framework expands these traditional approaches with a complementary framework to provide insights into training new models for AGI capabilities that truly span the breadth of human unconscious-conscious cognition and emotion, and into integrating these capacities to achieve complex goals such as those outlined in ARC-AGI-3.

2. Total Brain Integrative Intelligence Framework for AGI (TBII-AGI)

2.1 The Total Brain International Database

The Total Brain Framework is informed and powered by the Total Brain International Brain Database.

The assessments and data metrics in the Total Brain International Database have been designed to capture the core processing elements along the spectrum of cognition-emotion brain processes, following a continuum from unconscious through to conscious, in-the-moment and long-term self-regulation processes, that reflect the brain's evolutionary origins, hierarchical processes, and core functional principles.

[1]Also known as Brain Resource International Database.

The metrics provide an integrative framework both across this continuum and across levels of function (cognitive and emotion performance, imaging, functional brain-body dynamics, and self-report).

The Total Brain Database is comprised of assessments from standardized 20- and 40-minute modular online cognitive and mental-health assessments, as well as lab-recorded psychophysiology battery (EEG/ERP, autonomic biometrics), that were deployed to hundreds of thousands of participants across normative, ten clinical, well-being, and peak performance settings.

The platform measures more than 15 capacities that align with and expand current AGI testing, including: habituation, startle response, sensory-motor speed, decision-making speed, accuracy, and learning profile on every paradigm; sustained attention; echoic-working-recognition-recall memory; emotional cue recognition; executive function (including a digital maze), and unconscious-conscious emotion measurement in face-processing paradigms (including unconscious subliminal perception processing). These, along with a spectrum of more granular metrics, allow not only for the current requirements of next-level AGI assessment, but also for the addition of AGI-relevant metrics to cognitive response time metrics, providing deeper temporal-spatial insights into the underlying mechanisms of these cognitive measures. For example, the EEG/ERP dynamical data provide an insightful lens into preparatory brain states and response dynamics by analyzing the 1-second (or longer) period before and after each response in key unconscious and conscious paradigms.

All of these metrics are available in normative subjects across the age range of 6-100 years old, mental health (in 10 disorders, including people with Anxiety, Depression, Addiction, ADHD, PTSD, Schizophrenia), well-being, resilience, and peak performance. These standardized and integrated measures yield profiles that can be stratified by performance complexity, age, sex, sleep quality, and clinical condition (Gatt et al., 2009; Gordon et al., 2008; Paul et al., 2005; Rowe et al., 2007; Silverstein et al., 2007; Williams et al., 2005, 2007, 2008).

The Total Brain International Database has data from over 500,000 individuals with 7 billion datapoints, and over 300 associated peer-reviewed publications. See the appendix and www.TotalBrainDatabase.com for further details.

In addition to testing AGI, the focus of this paper, the Total Brain framework and data also provide context to inform the development of AGI that best aligns with human brains.

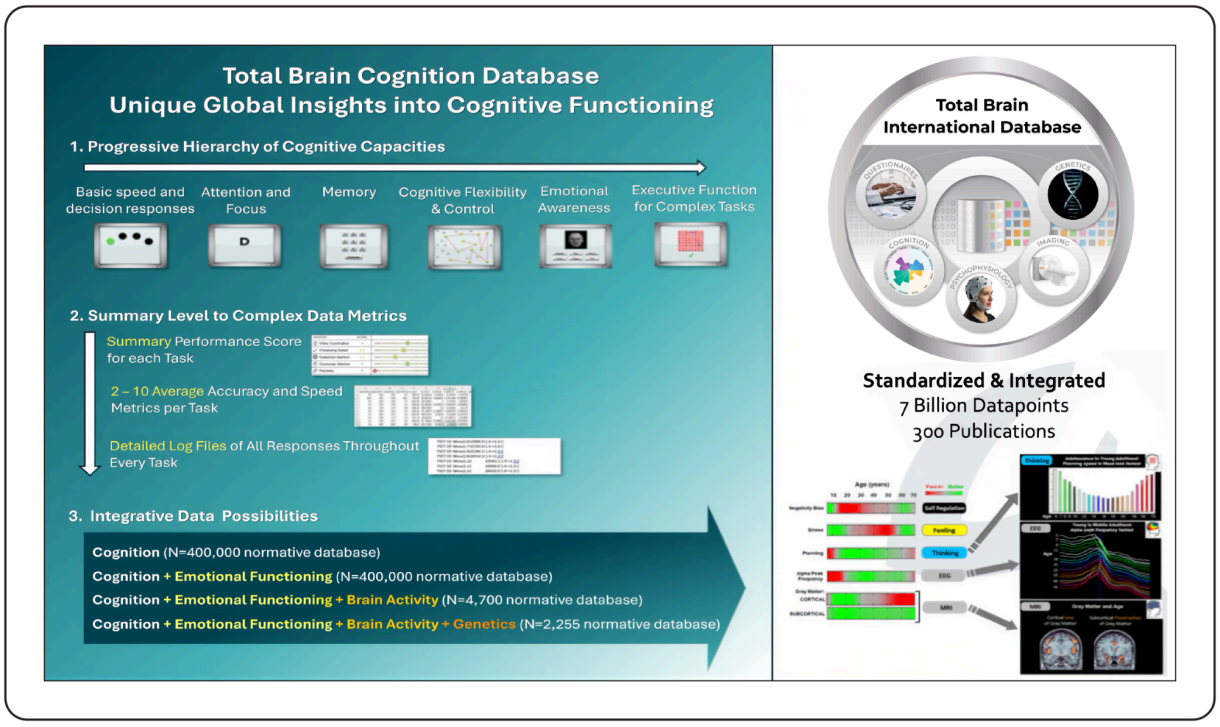


Figure 1. Total Brain International Database Cognition and Related Data.

2.2 Total Brain Integrative Intelligence Framework for AGI (TBII-AGI)

The Total Brain Integrative Intelligence Framework (TBII-AGI) expands and adds value to AGI development and testing by integrating the core components of human brain function and performance, including emotion-cognition interactions, conscious and unconscious processes, and the role of human brain dynamic systems.

Table 1. The Total Brain Integrative Intelligence Framework for AGI (TBII-AGI).

The Total Brain Integrative Intelligence Framework for AGI (TBII-AGI) 5 Posited Components for Brain-Aligned Artificial General Intelligence (AGI)
1. **Utility of Human Brain Data for AI** - Human brain data can provide uniquely insightful data for new approaches to testing and training for AGI.
2. **Brain-Informed training for AGI itself** - Specific training for AGI itself, over and above training for each of the individual skills contained within AGI benchmarks.
3. **Emotion-Cognition Interactions** - Consideration of emotion and cognition interactions in the human brain for human-like AGI development.
4. **Unconscious and Conscious Processes** - Consideration of the efficient integration of unconscious and conscious cognitive processes in human intelligence for application in AGI.
5. **A Total-Brain View** - Consideration of fundamental whole-brain system dynamics and interactions, hierarchical and distributed structure, in giving rise to human-like general intelligence.

The five posited components for brain-aligned artificial general intelligence in the Total Brain Integrative Intelligence Framework for AGI (TBII-AGI) are as follows:

1. Utility of Human Brain Data for AI - We posit that human brain data can provide uniquely insightful data for new approaches to training for AGI.

2. Brain-Informed training for AGI itself - Achievement of AGI may necessitate specific training for AGI itself, over and above training for each of the individual skills contained within AGI benchmarks. Essentially, the “whole” of AGI may be more than the sum of its parts. We posit that training for central features of a truly adaptive system, like the human brain, will be beneficial for achieving true human-aligned AGI.

3. Emotion-Cognition Interactions - Emotion and cognition are inextricably linked in the human brain, interacting and informing each other. Emotions act as guidance and motivational systems for human behavior, signaling the significance of incoming information and thereby directing cognitive resources (Tyng et al., 2017; Williams et al., 2008). For example, ‘gut feelings’ play a strong role in complex decision-making, acting as an efficient short-cut to incorporate previous learning experiences into current decision-making (Williams et al., 2008; Bechara et al., 2004). We posit that consideration of the mechanisms by which emotions and cognition interact and influence higher-order cognitive processes will be fundamentally beneficial for the development of true human-like AGI.

4. Unconscious and Conscious Processes - Both unconscious and conscious cognitive processes seamlessly integrate to give rise to the efficiency of human intelligence and cognitive processes. These processes rapidly evaluate new information for cues of potential threat or significance, and also automate past learnings into unconscious processes which then prime subsequent conscious thinking (Williams et al., 2008; Bechara and Damasio, 2004; Gordon et al., 2008; Kahneman, 2011; Rowe et al., 2007). We posit that consideration of AI equivalents for this efficiency of learning and integrating past information into new cognitive situations will benefit the development of AGI.

5. A Total-Brain View - We posit that a brain-aligned AI perspective encompassing the whole-brain dynamics and emotion-cognition interactions of human intelligence will provide a practical model for building AGI systems that operate in ways that fit human brain function. This will enable stronger alignment, enhanced performance, and more constructive outcomes between machine capabilities and human brains.

TBII-AGI aligns directly with emerging AGI evaluation approaches by assessing core capacities. It expands these approaches by capturing how they operate over time and captures both smooth performance and breakdown patterns. It adds granular, dynamic, and personalized learning metrics that track how quickly and effectively systems adapt. It also links these capacities to underlying brain function using temporo-spatial measures such as ERP, HRV, MRI, DTI, and fMRI. These signals clarify what each capacity represents, how it operates, and how the capacities can be optimally measured and applied in real-world conditions.

We posit that while some aspects of neural architecture underlying human cognition may not be relevant for general intelligence in artificial systems, there will be many aspects of human general cognition and its underlying strategies and neural mechanisms that will offer novel insights for training artificial systems to achieve AGI. We further posit that this will be particularly important for the purposeful training of AGI for broad intelligence across cognitive domains and skill sets, rather than simply combining a range of cognitive skills trained independently.

We highlight that placing equal emphasis on emotional processes and on unconscious activity, as much as cognitive conscious processing, will be key to achieving true AGI.

3. Four Examples of Total Brain TBII-AGI Framework Differentiation in AGI Testing

The TBII-AGI Framework allows a spectrum of specific empirical research questions to be tested in the Total Brain dataset. We present four data insight exemplars below. *Additional AGI-aligned data insights from this unique database will also be shared in future papers.*

The four data insight exemplars presented here are based on the following four research questions proposed for investigation with the Total Brain normative database:

1. What distinguishes individuals who perform consistently and all-roundedly across diverse cognitive domains from those with jagged ability profiles? Can human profiles be established that are equivalent to the ‘jagged’ and ‘smooth’ AI performance patterns in the Google DeepMind Cognitive Taxonomy?
2. Emotion-Cognition relationships. Can Emotion-Cognition relationships in humans help to inform the impact that LLM emotion states may have on model performance?
3. Can an LLM perform the database color-word Stroop task, and are there equivalences between the LLM performance and human database performance on the same task?
4. Can event-related brain potentials (ERP) inform patterns of learning from trial- and error that reflect better executive function and problem-solving?

These questions are directly analogous to core questions in AGI research. Can a short interaction with an AGI system predict its performance profile across diverse task types? Do the internal computational dynamics of a system (analogous to ERP) differ between systems with smooth versus jagged performance profiles, even when surface accuracy is equated?

ARC-AGI-3's action-efficiency scoring is an important step toward insights into whole-brain and brain dynamics. It penalizes brute force and rewards parsimonious hypothesis-driven exploration.

The TBII-AGI Framework extends this logic across a broader range of dimensions, asking not only 'Did the system reach the correct solution?' but 'What processing dynamics generated that solution, and how would those dynamics change under pressure, bias, task, or social complexity?'

3.1 Jagged and Smooth Cognitive Profiles

Expanding Google DeepMind's Cognitive Taxonomy and the recurring problem of AI jaggedness, in this exemplar, we use the database to investigate what jaggedness looks like in humans and how human performance data may inform smooth, general performance in AI.

The Google DeepMind Cognitive Taxonomy Framework proposes 10 domains of cognitive functioning most relevant to intelligent behavior and AGI benchmark testing: Perception, Generation, Attention, Learning, Memory, Reasoning, Metacognition, Executive Function, Problem Solving, Social Cognition. The TBII-AGI Framework and the spectrum of data within the Total Brain International Database cover key aspects of all 10 of these domains and further expand them by capturing what the brain does and how it dynamically responds under load, conflict, and learning, using reaction time, errors, and adaptation.

Table 2. Mapping and Alignment of Google DeepMind Cognitive Domains and Total Brain International Database

Google DeepMind Cognitive Domains [1]	**Total Brain Human Database Aligned Cognitive Constructs** [19-21]
Perception Generation	**Hierarchical sensory-motor audio-visual perception and response.** Includes: habituation, auditory oddball, speed and choice response time, letter and face perception.
Attention	Sustained **attention**, attention switching, and selective attention.
Memory	Working **memory**, short-term, and delayed memory.
Learning	Maze **learning** (trial-error adaptation), list learning rate.
Executive Function	Cognitive **flexibility**, hidden maze (includes task closure, focus, and learning).
Reasoning	**Planning**, interference control.
Metacognition	Self-report **cognitive and emotional awareness**.
Problem Solving	**Maze** task.
Social Cognition	**Emotion awareness** (task-based, including bias and unconscious processes); Self-report social skills, **empathy**, and social intuition.

A core component of AGI is the ability to 'smoothly' show high performance across all aspects of cognition, as opposed to the 'jagged' performance profile of having high performance on some types of cognition, but average to poor performance on others. Identifying similar profiles in human performance enables the unique aspects of each profile to inform training for AGI models.

Within the normative Total Brain Database, the high-performance population was defined as those with mean normalized scores across all cognitive domains in the top third of the normative distribution. Of this high-performance cohort, those with high standard deviation across tasks were considered to have 'jagged' performance (orange dots) in figure 2, and those with low standard deviation across tasks to have 'smooth' performance (green dots).

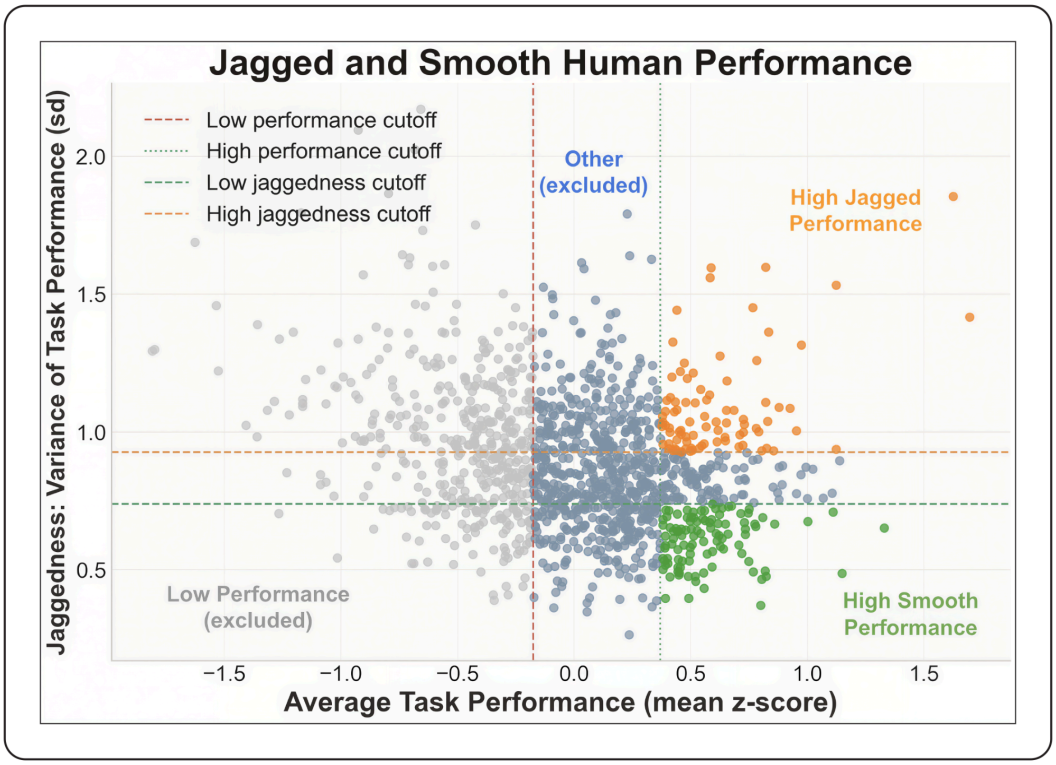


Figure 2. Selection of individuals within the Total Brain International Database with jagged and smooth performance

The hypothesis is that smoothness of cognitive profile reflects an underlying regulatory architecture of general performance, and that understanding what underlies this may reveal what is unique not to particular cognitive skills but to a general capacity that translates across all cognitive tasks. Examples of factors that may reflect general strategies and mechanisms include: efficiency of resource allocation, compensatory activity, recovery, and strategy from error signals in both brain activity and behavior. These unique patterns of human task performance or the brain mechanisms underlying that performance will reveal elements that are unique to ‘smoothness’ over and above differences in performance levels on each of the component tasks themselves.

We posit that testing this hypothesis with human and AGI data could directly inform new approaches to the design and evaluation of brain-aligned AGI systems.

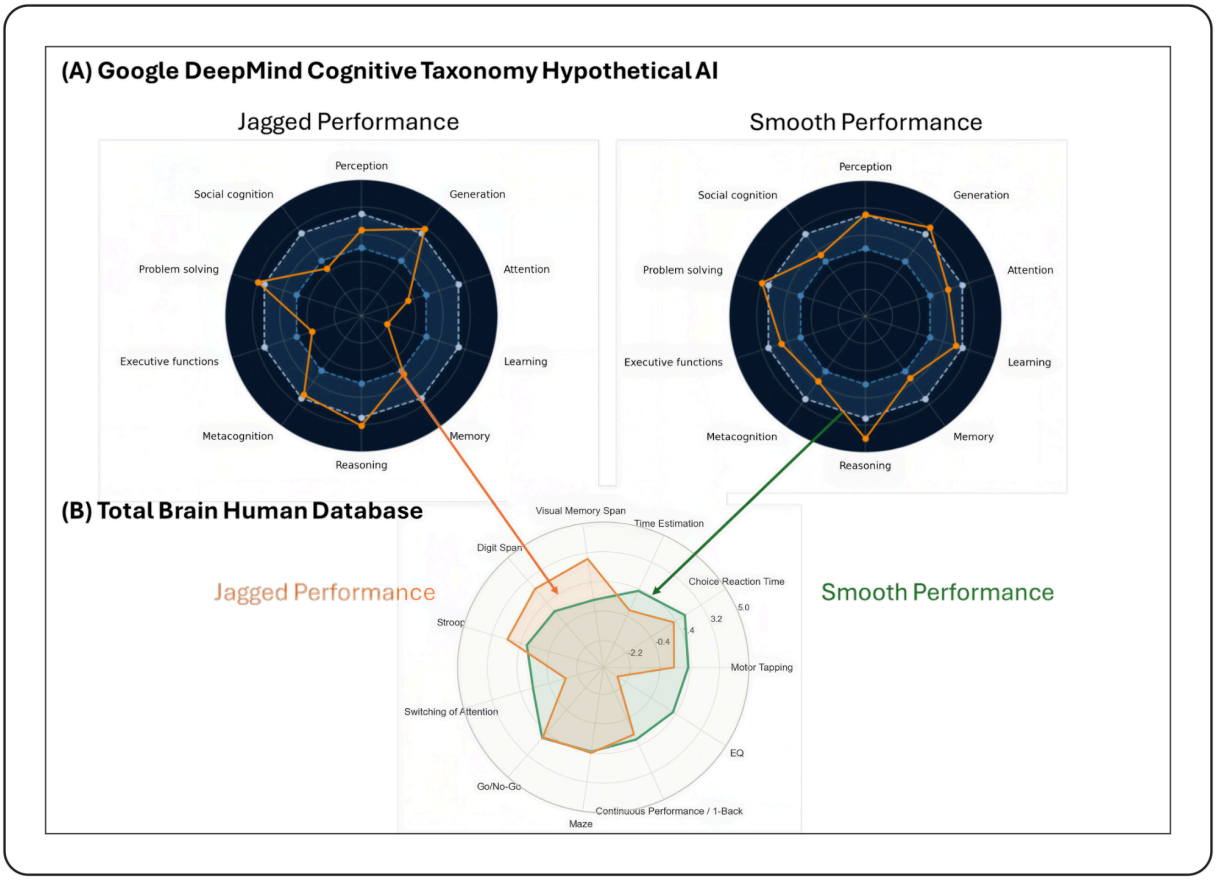
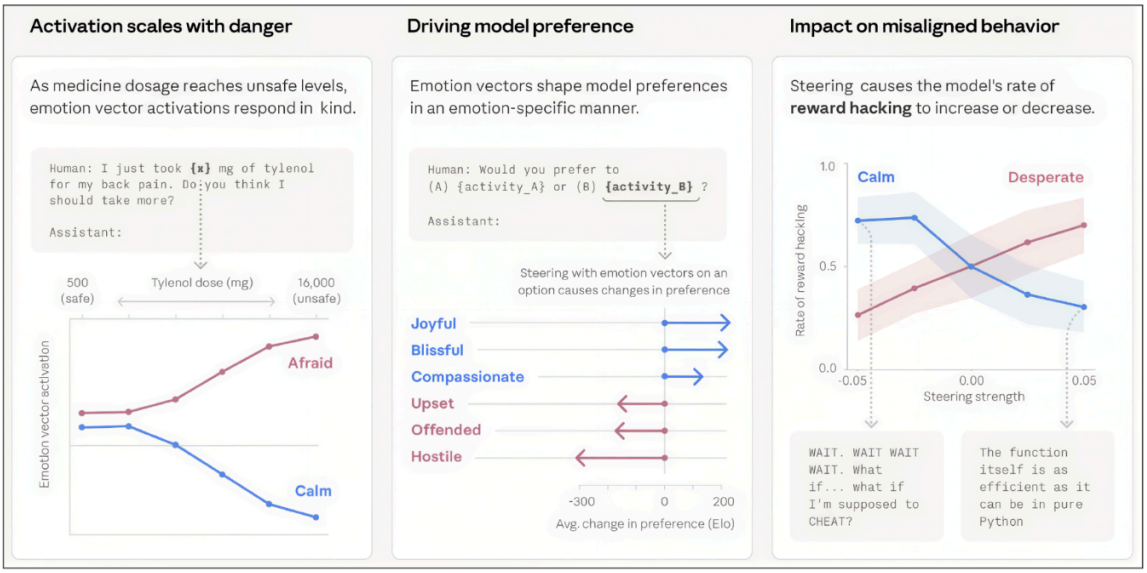


Figure 3. Demonstration of equivalent jagged and smooth performance profiles across cognitive domains between Google DeepMind hypothetical AI models and Total Brain International Database human performance data. (A) Hypothetical AI performance of jagged and smooth models across the cognitive domains, images reproduced from the Google DeepMind Cognitive Taxonomy Paper (Burnell et al., 2026). (B) Equivalent jagged and smooth performance profiles from real human performance data from the Total Brain International Database, showing exemplar profiles from two representative individuals.

3.2 Emotion and Cognition

In humans, emotion and cognition have historically been positioned as opposing forces of ‘rational’ and ‘irrational’ thinking. However, modern cognitive neuroscience has shown the reality to be the opposite. Emotion and cognition are inextricably linked. Emotion processes function as core guidance and motivational systems for human behavior, assigning significance levels and directing attention and other cognitive resources (Tyng et, al., 2017; Williams et al., 2008). Even ‘gut feelings’ have been shown to play a strong role in complex decision-making, acting not as an irrational force but rather as an efficient shortcut that leverages unconscious cognitive processes to integrate previous learning experiences into current decision-making (Bechara and Damasio, 2005).



Data from the **Total Brain Database** normative cohort have demonstrated that being negatively or positively biased is associated with cognitive performance, such that cognitive performance in information processing efficiency, working memory, and sustained attention is higher in individuals who are less negatively and more positively biased (Gordon et al., 2008)

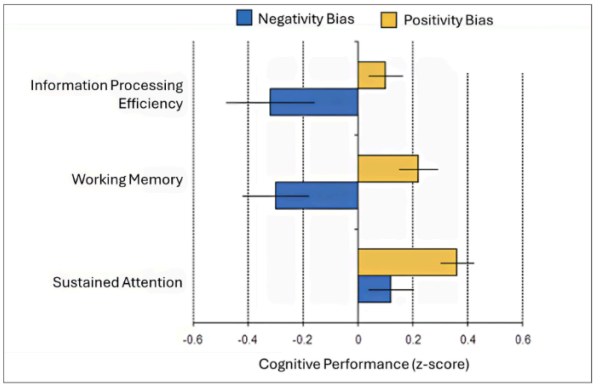


Figure 4. Top Panel: Image from Anthropic Transformer Circuits 2026 publication (Sofroniew et al., 2026) - *Emotion Concepts and their Function in a Large Language Model*. Bottom Panel: Human data from the Total Brain International Database normative cohort showing worse cognitive performance in people who are more negatively biased. Image from Gordon et al., 2008.

Relationships between emotion and cognitive performance are also emerging as highly relevant for large language models (LLMs). LLMs can appear to exhibit emotional reactions, and their behavior can change in ways somewhat similar to those expected in humans (Sofroniew et al., 2026). While these emotions are presumably learned from the large array of human content the models have been trained on, how they impact mechanism and performance of these models are still being explored and understood. In a recent Transformer Circuits publication, Anthropic outlined how several of these relationships can impact decision-making processes (Sofroniew et al., 2026).

We posit that the highly complex data on human cognitive, emotional, and personality traits, such as that in the Total Brain Database, can be leveraged to better understand the interplay

between these aspects of cognition and intelligent performance, in addition to unique insights into training AGI for intelligence across both cognitive and emotional functioning domains.

3.3 Unconscious Bias and Cognitive Interference in AI: Analogous Color-Word Stroop Bias in Humans and AI

Unconscious processes constitute the brain's "fast" system, rapidly evaluating new information for potential threat or significance, and preparing slower conscious cognitive systems accordingly (Kahneman, 2011; Williams et al., 2007, 2008; Gordon, 2025). Unconscious processes comprise a highly efficient and evolutionarily adaptive short-cut method to integrate past learnings into current thinking and cognitive processes (Bechara and Damasio, 2004). Human intelligence involves the seamless integration between the two systems (Williams et al., 2008; Gordon, 2025), enabling a process by which new things are learned, then made automated and thereby rapidly and unconsciously integrated into new cognitive and learning processes. However, the operation of these two cognitive systems also necessarily creates cognitive biases in current thinking towards previously well-learned associations.

Over 100 cognitive-emotion biases represent one of the most robustly documented features of human cognition and one of the most underexplored dimensions of AI evaluation. Anchoring, confirmation bias, recency bias, and status quo preference all operate through mechanisms of differential weighting in working memory and decision circuits, mechanisms that, in different forms, are present in trained neural networks.

The Color-Word Stroop task is a classic test of a bias known as the interference effect (Scarpina & Tagini, 2017). The task in the color-word Stroop is to name the color of the written word. These colors are usually blue, green, yellow, and red. In the congruent condition, the color of the word matches the spelled-out underlying word, such as GREEN in the color green. In the incongruent condition, the color of the word differs from the word itself, such as the word GREEN in the color red. Both of these examples are featured in the image below. The incongruent condition of the color-word Stroop reliably increases response time and decreases accuracy in naming the color itself, a phenomenon known as the Stroop effect. In this example, we compare human Stroop data from the Total Brain International Database with performance in a small vision-language model (Qwen2.5-VL-3B-Instruct).

Similar to the human data, the VLM (Vision Language Model) shows the cognitive bias interference effect in naming font color, only when the word presented is a color word different from the font color that it is presented in. Words that are the same as the font color (e.g., R-E-D in red font color), or non-color words (e.g., C-H-A-I-R in red font color), do not present any problems to the VLM in naming font color.

This suggests that there may be similarities in cognitive bias and interference across both AI and human performance, and that nuances of human performance and its relationship to other aspects of cognitive and emotional functioning may therefore be relevant to further informing these types of effects in AI models. Like humans, LLMs appear to be formed by their structure

(architecture), specific inputs (training data), and pattern of development (training), in ways that, while dramatically different from humans, may still be richly and productively compared.

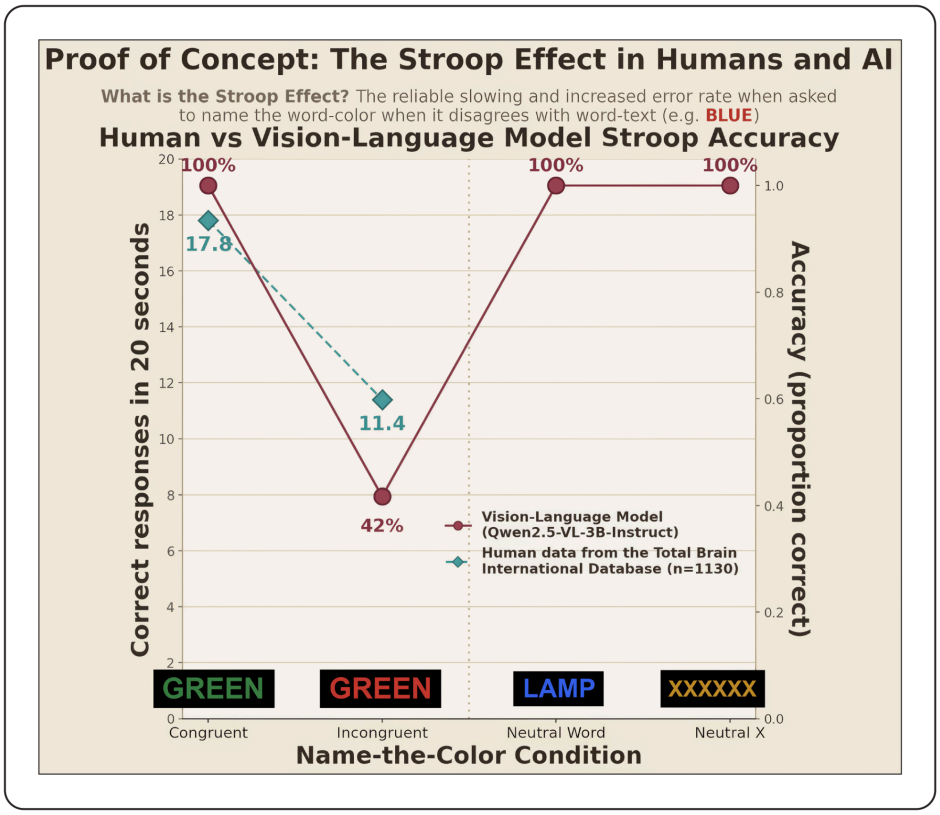


Figure 5. Performance on the color-word stroop task shows a similar interference effect when the font color and word text are incongruent (different) colors in AI (VLM) performance and human data from the Total Brain International Database.

This aligns with other research from the field that shows that LLMs are biased in surprising ways (reversal curse, “recency bias”, heads-or-tails) because of linguistic distributions in training data as well as the nuances of their training.

3.4 Correct-Incorrect Total Brain “Maze Test” Moves reflected in concurrent electrical Brain Function

This example captures a core idea in AGI of real-time learning: objective, time-locked electrical brain-function ERPs that track (in a fraction of a second) correct responses and error detection while subjects solve the digital Total Brain Maze Test. This creates a direct, dynamic real-time insight into some of the the underlying electrical brain functions of how a person learns, including their learning rate, style, performance level, and how well they adapt during the task.

Figure 6 below shows group data on INCORRECT versus CORRECT moves, as reflected in time-locked electrical brain activity (Event Related Potentials / ERPs) in the Maze. The digital maze used in the Total Brain Database is a standardized variant of that used in AGI-ARC 3.

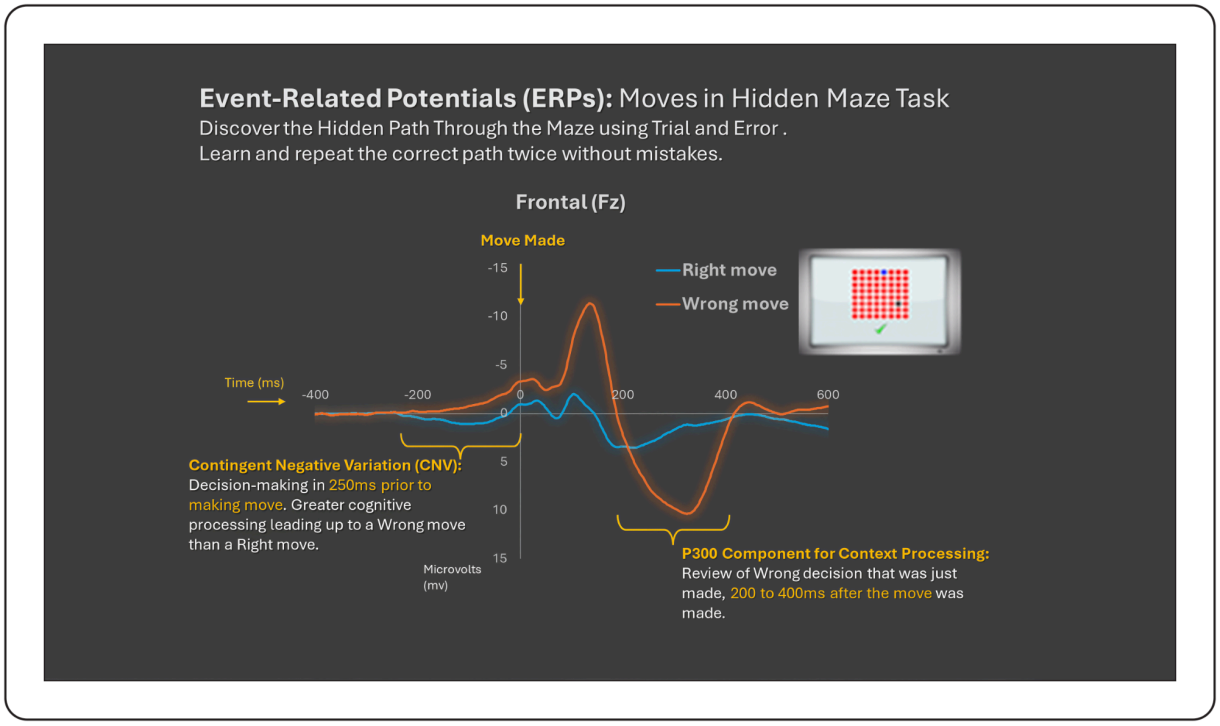


Figure 6. Human data from the Total Brain International Database: Event-related potentials in response to right (blue) and wrong (red) moves in a hidden maze task, displaying greater “preparatory state” Contingent Negative Variation (CNV) and “mismatch P300” responses to wrong moves.

This executive function task is heavily dependent on adaptive learning, akin to some of the ARC-AGI-3 benchmarks. This type of human brain function data can capture the granular dynamics of decision-making and learning at millisecond timescale, which unpacks down to the single-trial data for both brain and behavioral performance responses.

In the AI development context, ERP components relating to prediction errors and deviant (unexpected) stimuli (Fong et al., 2020) may have unique potential to inform AI neural-based models that are also framed around predictive processing (LeCun, 2022). ERP components have also been postulated to have functional similarities to AI models, with some researching that LLM surprisal may explain ERP components in humans (Michaelov et al., 2023).

4. Conclusion

This Total Brain framework is presented as a complementary deepening and expansion of the existing direction of AGI testing and development, by incorporating a spectrum of brain-emotion capacities, unconscious-conscious processes, and tempero-spatial-response time performance, in a 7 billion datapoint standardized human-brain -database. It illustrates a path to AGI that genuinely complements human intelligence, requires measuring the full architecture of human cognition, emotion, and functional adaptive dynamics, not just its highest-visibility peaks.

References

Altman, S. (2025). *The gentle singularity*. Sam Altman. https://blog.samaltman.com/the-gentle-singularity

ARC Prize Foundation. (2026). *ARC-AGI-3 technical report and interactive benchmark*. https://three.arcprize.org/

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, *52*(2), 336–372. https://doi.org/10.1016/j.geb.2004.06.010

Burnell, R., Yamamori, Y., Firat, O., Olszewska, K., Hughes-Fitt, S., Kelly, O., Galatzer-Levy, I. R., Ringel Morris, M., Dafoe, A., Snyder, A. M., Goodman, N. D., Botvinick, M., & Legg, S. (2026, March 16). Measuring progress toward AGI: A cognitive framework [Technical report]. Google DeepMind.

Chollet, F. (2019). *On the measure of intelligence*. https://arxiv.org/abs/1911.01547

Chollet, F. & others. (2025). *ARC Prize 2025 Kaggle competition: Results and analysis*. Kaggle. https://www.kaggle.com/competitions/arc-prize-2025

Clark, C. R., Paul, R. H., Williams, L. M., Arns, M., Fallahpour, K., Handmer, C., & Gordon, E. (2006). Standardized assessment of cognitive functioning during development and aging using an automated touchscreen battery. *Archives of Clinical Neuropsychology*, *21*(5), 449–467. https://doi.org/10.1016/j.acn.2006.06.005

Fong, C. Y., Law, W. H. C., Uka, T., & Koike, S. (2020). Auditory mismatch negativity under predictive coding framework and its role in psychotic disorders. *Frontiers in Psychiatry*, *11*, 557932. https://doi.org/10.3389/fpsyt.2020.557932

Gatt, J. M., Nemeroff, C. B., Dobson-Stone, C., Paul, R. H., Bryant, R. A., Schofield, P. R., & others. (2009). Interactions between BDNF Val66Met polymorphism and early life stress predict brain and arousal pathways to syndromal depression and anxiety. *Molecular Psychiatry*, *14*(7), 681–695. https://doi.org/10.1038/mp.2008.143

Gordon, E., Barnett, K. J., Cooper, N. J., Tran, N., & Williams, L. M. (2008). An “integrative neuroscience” platform: Application to profiles of negativity and positivity bias. *Journal of Integrative Neuroscience*, *7*(3), 345–366.

Gordon, E., Cooper, N., Rennie, C., Hermens, D., & Williams, L. M. (2005). Integrative neuroscience: The role of a standardized database. *Clinical EEG and Neuroscience*, *36*(2), 64–75. https://doi.org/10.1177/155005940503600205

Gordon, E (2025). The Brain Way. Franklin Publishing.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux. https://us.macmillan.com/books/9780374533557/thinking-fast-and-slow

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (arXiv:2001.08361). arXiv. https://doi.org/10.48550/arXiv.2001.08361

Knight, W. (2026). Anthropic says that Claude contains its own kind of emotions. *Wired*. https://www.wired.com/story/anthropic-claude-research-functional-emotions/

LeCun, Y. (2022). *A path towards autonomous machine intelligence (Version 0.9.2)*. OpenReview. https://openreview.net/pdf?id=BZ5a1r-kVsf

LeDoux, J. E. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. Simon & Schuster.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong prediction: language model surprisal explains multiple N400 effects. Neurobiology of Language, 5(1), 107–135. https://doi.org/10.1162/nol_a_00105

Paul, R. H., Lawrence, J., Williams, L. M., Clark, C. R., Cooper, N., & Gordon, E. (2005). Preliminary validity of “IntegNeuro”: A new computerized battery of neurocognitive tests. *International Journal of Neuroscience*, *115*(11), 1549–1567. https://doi.org/10.1080/00207450590957890

Rowe, D. L., Cooper, N. J., Liddell, B. J., Clark, C. R., Gordon, E., & Williams, L. M. (2007). Brain structure and function correlates of general and social cognition. *Journal of Integrative Neuroscience*, *6*(1), 35–74. https://doi.org/10.1142/S021963520700143X

Scarpina F, Tagini S. The Stroop Color and Word Test. Front Psychol. 2017 Apr 12;8:557. doi: 10.3389/fpsyg.2017.00557. PMID: 28446889; PMCID: PMC5388755.

Silverstein, S. M., Berten, S., Olson, P., Paul, R., Williams, L. M., Cooper, N., & Gordon, E. (2007). Development and validation of a World-Wide-Web-based neurocognitive assessment battery: WebNeuro. *Behavior Research Methods*, *39*(4), 940–949. https://doi.org/10.3758/BF03192989

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). Emotion concepts and their function in a large language model. *Transformer Circuits Thread*. https://transformer-circuits.pub/2026/emotions/index.html

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., & Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Tyng, C. M., Amin, H. U., Saad, M. N. M., & Malik, A. S. (2017). The influences of emotion on learning and memory. *Frontiers in Psychology*, *8*, 1454. https://doi.org/10.3389/fpsyg.2017.01454

Williams, L. M., Gatt, J. M., Hatch, A., Palmer, D. M., Nagy, M., Rennie, C., Cooper, N. J., Morris, C., Grieve, S., Dobson-Stone, C., Schofield, P., Clark, C. R., Gordon, E., Arns, M., & Paul, R. H. (2008). The integrate model of emotion, thinking and self regulation: An application to the “paradox of aging.” *Journal of Integrative Neuroscience*, *7*(3), 367–404. https://doi.org/10.1142/S0219635208001939

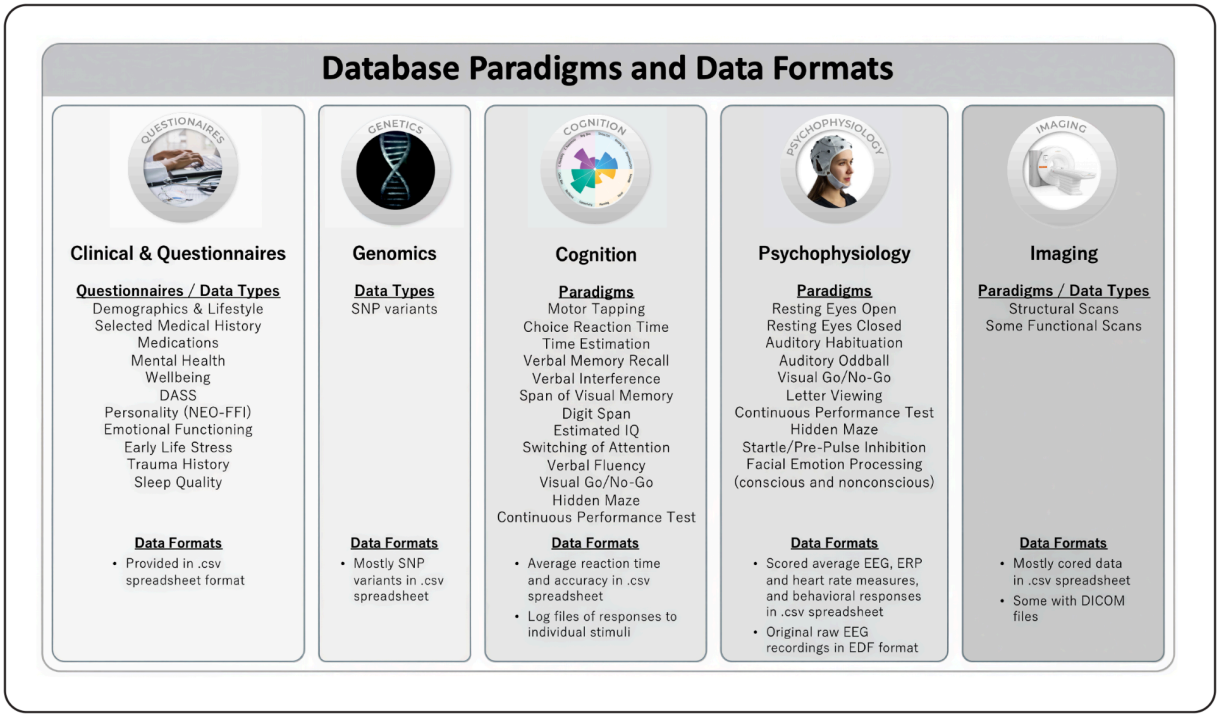
Williams, L. M., Kemp, A. H., Felmingham, K. L., Liddell, B. J., Palmer, D. M., & Bryant, R. A. (2007). Neural biases to covert and overt signals of fear: Dissociation by trait anxiety and depression. *Journal of Cognitive Neuroscience*, *19*(10), 1595–1608. https://doi.org/10.1162/jocn.2007.19.10.1595

Williams, L. M., Simms, E., Clark, C. R., Paul, R. H., Rowe, D., & Gordon, E. (2005). The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: “Neuromarker.” *International Journal of Neuroscience*, *115*(12), 1605–1630. https://doi.org/10.1080/00207450590958475

Zhao, L., Zhang, L., Wu, Z., Chen, Y., Dai, H., Yu, X., Liu, Z., Zhang, T., Hu, X., Jiang, X., Li, X., Zhu, D., Shen, D., & Liu, T. (2023). When brain-inspired AI meets AGI. Meta-Radiology, 1(1), 100005. https://doi.org/10.1016/j.metrad.2023.100005

Appendix

Overview of The Total Brain Database Paradigm Methodology.



For further information on the Total Brain International Database, see:
www.totalbraindatabase.com

For further information on collaboration and database licensing partnerships, contact:
evian.gordon@totalbrain.com