FEATURE REVIEW

Understanding by Building Agents That Learn Like the Human Mind and Brain

Uri Hasson · Ariel Goldstein · Hadas Raviv

Princeton University | University of Cambridge

Abstract

Cognitive science has worked mostly in the lab, trading real life for precision, at the cost of findings that struggle to replicate or generalize. We propose a unified research program to study the mind and brain as they operate in the world, organized around three mutually reinforcing objectives. First, real-life recording: capture human experience in dense detail as it unfolds in the wild. Second, behavioral alignment: train cognitively feasible agents that learn from that experience as humans do and come to behave like humans. Third, neural alignment: test whether these agents compute as the brain does, and rebuild the model wherever the two diverge. An agent aligned with humans in both its natural behavior and its underlying neural processes becomes a working mechanistic explanation of how our minds and brains work: understanding by building.

Keywords: ecological validity; naturalistic paradigms; language acquisition; deep neural networks; neural alignment; behavioral alignment; developmental modeling

Highlights

- Most of what cognitive science knows comes from controlled experiments, a form of control now linked to problems with replication and with predicting real-world behavior.
- Dense, continuous recording of real human behavior across all relevant timescales, not brief lab sessions, is now technically and ethically feasible at scale.
- Unlike standard AI trained on aggregate corpora, agents can now learn from one person's real, ordered experience, achieving behavioral alignment with that individual's learning trajectory.
- Recording brain activity during natural behavior adds neural alignment: an agent's computations can be compared against the brain's, revealing shared principles and informative divergences.
- Together, these define a unified paradigm built for cognitive and neural feasibility, matching how people learn and brains compute, not engineering performance.

Introduction

A paradigm shift toward minds in the world

This paper proposes a new paradigm for studying the mind and brain as they operate in real life, shifting the focus away from controlled laboratory conditions that have shaped experimental psychology since the 1800s [1–3] and cognitive science since the 1950s [4,5]. Artificial intelligence underwent a comparable shift toward building models that act in the natural world a generation ago. Systems once built from hand-specified rules and curated representations were gradually replaced by systems that learn directly from large volumes of real-world experience [6]. We argue that psychology and neuroscience are due for an analogous shift — toward building computational agents that learn from the same kinds of input human minds actually receive. They can then be held to the same standard AI systems now face: they must act, predict, and generalize in the world, not merely fit a narrow dataset collected to test a hypothesis. The argument is general, applying to perception, memory, decision making, social cognition and motor control, as much as to language. Language, however, is the domain in which the relevant tools, datasets, and neural evidence are most mature, and so it recurs throughout as our clearest working example in Boxes 1-3.

The mind in the wild

Consider what a mind/brain does on an ordinary day. It perceives: you spot a friend in a crowd and know their face at once, though you last saw them twenty years ago. It remembers: you follow an audiobook for hours, hearing tens of thousands of words without your working memory ever filling up, and afterward, you easily recount the story to a friend. It understands and thinks: you recover a friend's meaning from a half-spoken, ungrammatical sentence, then follow a page of Nabokov, or analyze a formal argument with the same machinery. And it acts: you swiftly bike or fall into step with a dance partner, producing the right movement in the moment, faster than you could put your actions into words. None of this strikes the person doing it as remarkable. But all of it is extraordinary!

Psychology and neuroscience have never shied away from that richness; most books, articles, and talks open with exactly this kind of vivid demonstration. But the path from marvel to explanation has run almost entirely through one strategy: divide and conquer, as has been done with great success in physics and chemistry. To work out how the mind and brain do all this, we are trained to isolate one faculty, one task, one variable at a time. The approach is principled: treat the mind as a set of separable components, each with its own dedicated machinery [7,8]. Its virtues are real. A controlled experiment can pin down a causal variable with a precision no observation of natural behavior could match, and a century of its predominant use has taught us how to carve the mind into perception, memory, language, thought, and action. We have theories of face perception built on isolated faces on a screen [8,9], models of working memory calibrated on lists of random words [4,10], accounts of language acquisition that isolate single phonemes, and models of motor control fit to a single arm reaching for a single object [11,12]. Each studied dimension captures a sliver; none can model how we perceive the friend in the crowd, acquire language from a flood of input (Box 1), tell a story, or dance with a partner. For all we have learned about the pieces, we are no closer to explaining how the pieces combine in real time, and in a real body, to produce behavior in the wild.

This gap between explaining isolated phenomena and explaining behavior in the wild underlies two challenges the field now faces: the replication crisis and the ecological validity predicament (see Glossary). Both are the same failure measured at different scales. The replication crisis exposes the difficulty of generalizing across nearly identical contexts: another laboratory, another sample [13], a slightly altered task. No two studies are ever truly identical, so every replication tests whether a finding survives beyond the conditions that produced it. When it does not, the failure is usually read as sloppiness or a false positive. Part of the reason, however, is deeper: a result obtained under one carefully chosen set of conditions was never designed to hold when those conditions change [14,15]. Failures to replicate are, in part, expected by design.

The ecological validity predicament is the same failure to generalize over wider contexts: from controlled conditions to the natural settings the mind evolved to navigate [16,17]. If findings struggle to travel between two nearly identical lab tasks, they have little hope of traveling from the lab to real life [18]. Divide-and-conquer buys precision by narrowing the conditions under which its findings hold, but a science built by isolating variables has no principled way to know in advance which findings will cross its self-imposed boundaries, and no principled way to recombine the ones that do. That is the deep reason its carefully characterized pieces so rarely add up to a mind that works in the world.

The constructive response to this predicament has already begun. Naturalistic paradigms now study cognition with rich, dynamic, lifelike stimuli rather than impoverished ones [19,20], and a growing methodological literature argues for rethinking the primacy of tight experimental control altogether [18]. We take these moves as a starting point and ask what a fully ecological science of the mind and brain would require. Our answer is: building models of cognition as robust as cognition itself, without retreating into the lab or abandoning its rigor. These models learn from the world as a person actually encounters it, and are judged, as a person's brain is, by what they can do in it.

Box 1. Language acquisition: many theories, no working model.

A child learns language while immersed in a continuous flow of speech, gestures, shared attention, actions, and routines. Explaining how this unfolds requires an account that connects these signals to the emergence of linguistic knowledge. No framework yet provides a unified, empirically validated account of how learner capacities, environmental structure, bodily experience, and social interaction work together under naturalistic conditions. Theories differ in how they answer two questions: what the child brings to learning, and what the environment provides.

What the child brings. Generative and nativist accounts propose that children are born with domain-specific core linguistic knowledge, without which the input would be too impoverished to learn from [85]. The statistical-learning view replaces innate core knowledge with an innate general-purpose learning mechanism that extracts structure from rich input [40]. Embodiment shifts the focus again: learning is constrained and enabled by the child's developing body, sensorimotor capacities, and ongoing interactions with the physical and social world [50]. What the environment provides. Distributional accounts examine the statistical regularities in children's input that can support learning. Usage-based accounts emphasize the structure available in communication, including recurrent constructions, item-based patterns, and relationships among form, meaning, and function [86,87].

Social-interactionist accounts emphasize how responsive partners, joint attention, and contingent interaction organize children's experience, guide attention, and facilitate (or, in some circumstances, gate) early language learning [88].

These perspectives rest on substantial experimental evidence, and many are complementary rather than competing. By design, lab studies isolate a limited set of variables, which is essential for identifying the contribution of individual factors but leaves two questions unresolved: how to adjudicate between theories when they truly conflict, such as whether structure originates in the learner or in the input, and how complementary factors combine to support behavior in the natural world. Neither can be settled when the candidate explanation is never tested under real-life, multidimensional conditions. The accumulated knowledge resembles a collage: we can interpret many pieces individually, but we understand little of how they interact nonlinearly in the real world. We therefore propose a different kind of model, defined by two conditions. First, it learns from the messy, rich input of everyday life, as one child does. Second, its

development can be compared directly with that child's. Requiring both sharply constrains the space of possible explanations.

Mechanistic models that work in the natural world

Judged by doing, not by fitting: The under-specification diagnosed in Box 1 does not disappear when a laboratory result is turned into a computational model; it merely takes a more concrete form. A model of memory, perception, or language acquisition is typically developed and validated on a limited set of data points collected under tightly controlled conditions, and judged by how closely it conforms to that dataset, by explained variance, likelihood, or the significance of a parameter of interest. Such a model may explain the data it was designed to fit while providing little evidence that it captures the broader mechanisms of the mind that generated those data. Nothing guarantees it will hold up against a new stimulus set, a new population, or a task one step removed from the one it was built for.

Artificial intelligence has made substantial progress on this problem: a system's competence is established behaviorally, through continuous testing. A self-driving car is judged on roads it has never seen, a transcription system on speech it has never heard, a language model on a question absent from its training set. Generalization across contexts is the central objective, and Sutton [6] argues it has driven seven decades of progress: approaches built on hand-specified features have been repeatedly overtaken by ones that learn from real data how to act in complex environments. Requiring a cognitive model to act in the real world is the central methodological move of our framework. A good model of human cognition should be judged not by explaining a result after the fact, but by whether it can operate under the conditions people actually face, produce the behavior people produce, and reach comparable competence.

Acting is necessary; yet it is not sufficient. The clearest evidence comes from outside psychology. Unlike our lab-bound models, large language models (LLMs) handle the full complexity of natural language: they hold conversations, answer questions, write working code, and generalize to genuinely novel input, meeting the behavioral standard described above. Yet they are not models of human

language, for one central reason: they learn from a volume of input no human could ever digest. A system can pass the behavioral test while learning in a way no human ever could. The gap is not defined by any particular task at which these models fail; tomorrow's model may pass any test we name. It is more fundamental. AI typically learns by gradient descent on a curated, decontextualized, internet-scale dataset, offline, in a static training environment. It lacks essential dimensions of human experience: a body, ongoing interaction with caregivers and others, and a developmental trajectory through which sensorimotor, cognitive, emotional, and higher-order capacities emerge together and shape one another. Closing this cognitive-feasibility gap, alongside the behavioral gap, is what our framework is intended to accomplish.

One unified arc

Acting successfully and learning as a human does are, therefore, complementary goals. A model built both to achieve ecological validity and to explain mechanistically how the mind and brain operate in the real world must be cognitively and neurally feasible along the following dimensions at once, with each measured against a developing human rather than against an engineering benchmark:

● Feasible input - trained on the kind of experience an actual human receives: bounded in volume, multimodal, embedded in a real developmental timeline, in contrast to a curated internet-scale corpus that no person would ever encounter.

● Feasible learning procedure - built on mechanisms that constitute plausible hypotheses about how the mind and brain extract structure from ongoing experience (e.g., incremental daily exposure, consolidation, replay), in contrast to mechanisms chosen solely to improve performance.

● Feasible behavioral target - evaluated against the trajectory and the limits of real human performance, not against a superhuman ceiling. A model that learns language faster, or recalls more perfectly than any human ever has, is not a better model of acquisition; it is evidence that it learned something other than what humans learn.

● Feasible neural computation — A model should be judged not by behavior alone but by direct comparison with measured neural activity. Behavioral fit cannot rule out that a model reaches the right answer by a non-human route: networks trained to classify images can match human accuracy while exploiting cues, such as diagnostic pixels or surface texture, that humans do not rely on [21,22].

These four dimensions map onto three research objectives (Figure 1). Objective 1 is capturing human development in the real world, which supplies feasible input. Building an agent that learns and acts as humans do, by combining a feasible learning procedure with a feasible behavioral target, defines Objective 2. We keep those two dimensions distinct because how an agent learns and how we measure that learning are separate methodological questions. Aligning artificial and biological computation defines Objective 3, which remains distinct because feasible input, learning, and behavior do not by themselves guarantee human-like internal computation. A model can receive the right input, learn through a plausible procedure, and perform as humans do while still computing in ways the brain does not.

human mind and brain compute, opening questions that are currently intractable. The rest of this paper proposes how to build and test that kind of model.

The three-objective framework

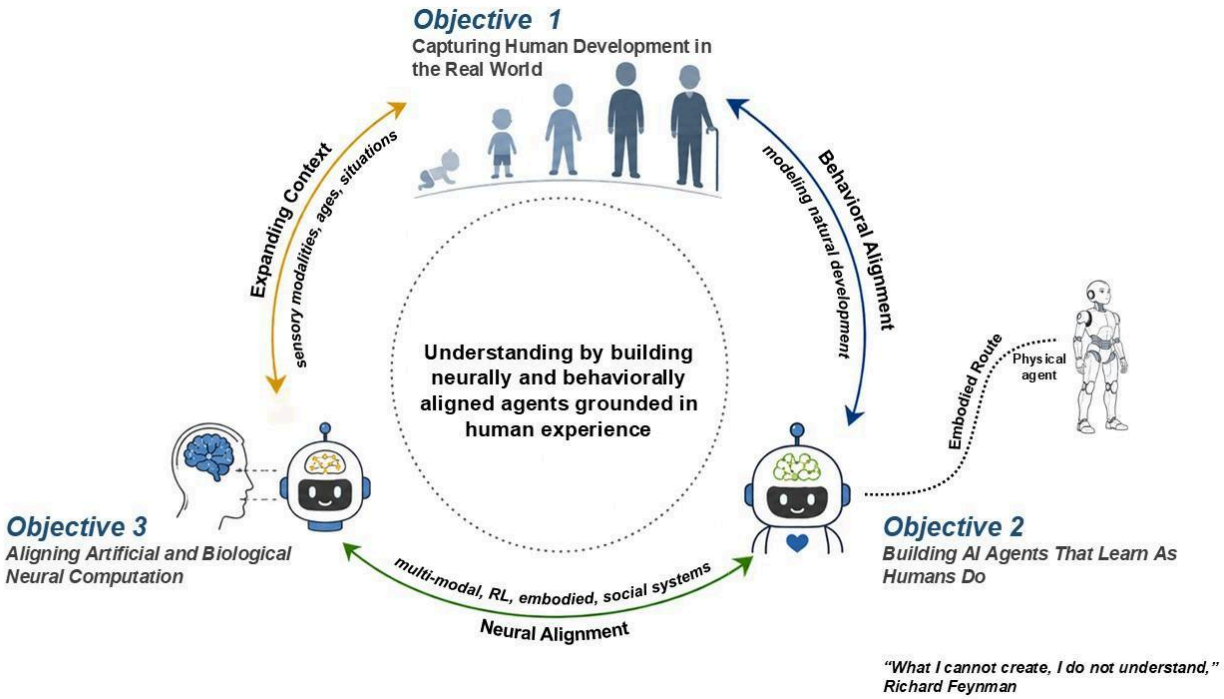


Figure 1. A unified framework for modeling the human brain and mind in real-world contexts. Dense observations of human development inform cognitively feasible artificial agents; behavioral and neural alignment provide complementary tests, creating an iterative cycle of model refinement.

Objective 1: Capturing human development in real life

Feasible input is the dimension on which the rest of the framework rests: an agent's statistical learning is limited to the data it encounters, its perception to the scenes it experiences, its actions to the opportunities available to it. Feasible input means experience as a human actually lives it, situated in a real developmental timeline, rather than reconstructed from a corpus no individual could have encountered. This requirement is easy to state and extraordinarily difficult to satisfy.

Cognitive science has historically studied a vanishingly small slice of the experience it seeks to explain. Child development makes the arithmetic concrete. A developmental study might observe a child for an hour in the lab, or at best for a handful of scattered days across years. The first 1,000 days of a child's life contain roughly 12,000 waking hours, so a single visit samples about one ten-thousandth of the period it is meant to characterize, and samples it at one point rather than across a trajectory. Development also runs at several timescales at once: second by second within an exchange, across daily situations that recur thousands of times, and over months and years of accumulated change. These timescales are not independent. A moment of shared attention takes its shape from the situation around it, and it is those situations, repeated day after day, that accumulate.

The scientific ideal is worth stating plainly: continuously monitor a large and diverse sample of people from birth through old age, recording everything each person hears, sees, says, and does as it happens, in the real-life contexts in which it happens. Such a dataset would need to meet two requirements at once: density, enough detail about each individual life to reveal how that person actually learns and behaves, and diversity, enough variation across people, circumstances, and contexts to capture how development differs among them.

The combination of density and diversity may seem excessive; one might argue that sparse sampling across many people, or dense sampling of only a few, would suffice. Either can support certain aggregate summaries, but neither alone reveals how learning actually unfolds. Density recovers the order in which developmental changes accumulate, the trajectory of an individual learner, and the rare events that shape it [23]. Diversity determines which trajectories and mechanisms are shared across people and which depend on particular cultures, environments, or histories. Where dense records exist, they reveal structure that thin sampling obscures [24–26]. Yet we do not currently know how much density or diversity is enough, nor how the two trade off against one another. If artificial learning has taught us one thing, it is that the amount and structure of the available data matter. We make the same bet and state it as a hypothesis: understanding human intelligence requires sampling the experience from which it develops at a scale sufficient to capture both the structure of individual lives and the variation across them.

What stands in the way

No current effort can capture the full density and diversity of human experience envisioned here, for four compounding reasons. The first is infrastructure: maintaining recordings at this volume securely, for

years, across many participants would strain even the most advanced storage and annotation systems in academia. The second is hardware: cameras, microphones, eye-trackers, and sensors for heart rate, skin conductance, and movement are now small enough for extended recording, yet no single device combines an overnight-rechargeable battery, reliable secure transmission, and durability for years in the wild. The third obstacle is ethical, and engineering alone cannot solve it. Sustained recording of a developing person raises questions about consent a young child cannot give, about how individuals can maintain control over an archive of their own private life, and about the others who pass briefly through a recorded space without agreeing to be in it. A continuously running recorder cannot tell in advance the difference between an ordinary afternoon and a moment of vulnerability no one would choose to document. Density, then, is an ethical variable here, not merely a scientific one. The fourth obstacle is institutional and temporal. No

individual researcher, and arguably no single field, can carry out a century-long study; it must be designed to outlive everyone who starts it, surviving retirements, funding cycles, and changes in leadership. We call this the cathedral problem: those who laid the foundations knew they would not live to see the spire. Naming it is a precondition for solving it. We name it in the hope that a future institute will take it on.

Density and diversity: two roles, two compromises

Because both density and diversity are costly, every real research program has to choose which one to prioritize, and that choice determines what the resulting data can reveal.

The first compromise preserves density and sacrifices diversity: record very few individuals, over a limited window of life, but with enough continuity that an individual's actual learning trajectory, rather than a population average standing in for it, becomes visible. Deb Roy's Human Speechome Project recorded his child's home environment nearly continuously for three years, showing that such density was achievable long before it became common practice [27]. Box 2 presents a contemporary example from our lab, covering the first 1,000 days of children's lives [26].

The second compromise runs the other way: sacrifice density to preserve diversity, recording many individuals briefly or in a single narrow context and pooling the results. Developmental science already has a foothold here, with daylong audio recordings of thousands of families [28,29], egocentric video corpora spanning many lives and several countries [30,31], and classroom observation at the scale of an entire school system [32]. Wearable cameras now exist in glasses-sized form [33], making it possible to sample natural visual experience across many people and environments and thereby capture a range of visual statistics that no single recorded life could provide. Yet scale alone does not guarantee diversity: a large dataset may still represent a narrow range of contexts, communities, and ways of living.

Diversity does two different things for a dataset: it can substitute for density, and it can compound it. Substitution works by aggregation — sample many children briefly, pool the records, and the result approximates a densely sampled average child. Whether that approximation holds is an empirical question, and even within language the answer is not uniform. Phonological statistics are widely shared across households and may be recoverable by pooling sparse observations. The lexicon is not: pooling thin slices recovers the high-frequency core but averages away the long tail of household-specific

words tied to each family's people, objects, places, and routines [26]. Because those tails overlap only weakly, aggregation produces a distribution no individual child ever encountered, whereas dense recording preserves the distribution from which a given child actually learns (Box 2). Compounding is a different design. A thousand children recorded for a day yields one average child; a thousand children each recorded for a thousand days yields a thousand children. Only the second supports both targets at once: a model of each individual trajectory, and a model of what makes trajectories differ. Individual differences cannot be recovered from an aggregate, because averaging removes those differences.

Density is required wherever structure is person-specific; diversity can compensate wherever experience is shared; together they make individual variation modelable. This is where Objective 1 has the most room to grow, along the ages, cultures, and circumstances marked by the Expanding Context arrow in Figure 1, and where it constrains Objective 2, whose cognitive feasibility can only be as strong as the experience Objective 1 provides.

Box 2. What a thin slice misses

The First 1,000 Days (1kD) Project set out to record individual children's lives densely at home, from birth through their first 1,000 days. The study included 17 infants in 15 racially and economically diverse U.S. households who were recorded 12 to 14 hours a day for a median of 944 days, using audio and video

devices placed throughout each home. The result is roughly 1.18 million hours of recordings, not a sample of childhood but a near-continuous record of it [26].

Analysis of the recordings revealed two kinds of structure in children's language environments: one that sparse sampling across many homes can recover, and one that only dense sampling within a home can reach. Word frequency declines smoothly from the handful of words a child hears thousands of times to the many thousands of words heard only a few times, with no natural break between common and rare, the Zipfian form familiar from every natural language corpus [92]. That curve is nearly identical from home to home, and because it is shared, a brief visit to many homes recovers it. What differs is where each word sits on the curve. In roughly half the cases, a word occupies a position in one home that it occupies in almost no other, and the divergence concentrates in the tail, where most of a child's vocabulary lives. Resolving those positions is what density buys: a word heard once a week cannot be placed on the curve from a weekend of recording. Nouns vary more than verbs, consistent with the differences between each family's particular people, objects, and routines. Sparse sampling misses that home-specific structure in two opposite ways. Record one home briefly and you capture the common core but not the long tail, where words recur too rarely to surface. Pool short recordings from many homes and the reverse happens: the vocabulary grows large, but each home's distinctive words are diluted, producing a flat tail in which nothing is as frequent as it was for any actual child.

Density and diversity are competing demands on finite recording. The 1kD chose density: 15 homes rather than a thousand, three years rather than a few days. These recordings make it possible to model how a single child learns from the distribution that child actually heard. Box 3 asks what a learning agent can do with input of this kind.

Objective 2: Agents that learn as humans do

The same machinery, across cognitive domains: Objective 2 is defined by two of the four dimensions, a feasible learning procedure and a feasible behavioral target. The agent should learn under the same kind of constraints as a human and be evaluated against what humans actually achieve (Figure 1). We illustrate the objective through language (Box 3), for which dense developmental data already exist (Box 2), but neither the standard nor the commitment behind it is limited to words. If artificial and biological networks belong to the same family of computational solutions, the principles that enable a system to learn from real-world experience and act within it should extend across domains, helping to explain how a child learns to recognize a face, anticipate a thrown ball, or infer another person's emotions from a glance. The same logic extends beyond our own species: an animal reared on its own real experience should be modelable in the same way outlined in Figure 1, and its brain tested against the resulting predictions (Box 4).

When a cognitive scientist meets a large model, the tendency is to search it for the constructs we already hold and to lose interest when those constructs are absent [34]. We propose treating these systems as a source of new ideas rather than a mirror of old ones. Despite the engineering complexity surrounding them, their core computational principles can be stated in remarkably compact mathematical terms. Next-token prediction minimizes the cross-entropy between predicted and observed input [35]; attention assigns weight according to the normalized similarity between a query and a set of keys [36]; reinforcement learning modifies predictions and actions in proportion to a scalar reward [37]. Each is a precise hypothesis about a computation the mind may need to perform on natural input, and each has a counterpart in a major branch of psychology: prediction in predictive coding and statistical learning, attention in the study of selective attention, and reward-driven reweighting in associative learning. These three primitives, predict, attend, and reweight by reward, recur across the five stages of a single learning agent discussed below: statistical learning, perception, feedback and instruction, embodiment, and world modeling. Each is a plausible

hypothesis about how a human learns. The input it runs on in industry, vast and shuffled and decontextualized, is not. Objective 2 is therefore to reconstruct these computational principles within a cognitively grounded learning system trained on the feasible experience Objective 1 provides.

Statistical learning. Before it does anything resembling reasoning, a large language model spends most of its existence predicting what comes next in a stream, trained with no predefined rules, no innate knowledge of language structure, and no explicit instructions [38]. Stripped of its industrial scale, this pretraining objective provides a mechanistic implementation of predictive coding, a principle central to theories in psychology and neuroscience [39]. Eight-month-olds, for example, can segment words from continuous speech using the statistical regularities between sounds, after only minutes of listening [40]. The computational principle transfers naturally between infants and large language models; the experience on which it operates does not. A child's statistical learning is bounded, sequential, and embedded in one continuous timeline, exactly the kind of experience Objective 1 captures, and what an agent trained on it would receive (Box 3).

Box 3. From real speech to a growing lexicon

In the 1kD project, we operationalized Objective 2 by pairing ultra-dense recordings of children's first 1,000 days (Box 2) with a cognitively feasible learning model designed to learn language as children do in their natural habitat [55]. The model received speech derived from the recordings of a single child in a cognitively plausible order: day by day, chronologically. It began with random weights and no built-in linguistic knowledge. An encoder learns to classify the continuous speech stream into discrete units roughly corresponding to subphonemes, and a decoder learns which units reliably follow which. Once a sequence becomes predictable enough, it counts as a learned word. Critically, the training protocol implemented a replay mechanism with no analog in standard machine learning: at the end of each day the agent replayed a sample of recently heard speech before training resumed. Without replay, the model learned almost no words from raw exposure.

Eight agents were trained, one per child, each on the speech recorded around that child alone. Architecture and training procedure were identical across all eight; what differed was the child's own input and three interpretable learner-internal parameters: error accumulation window, replay amount, and learning rate. Every agent discovered phoneme-like categories, acquiring an average of 37.5 phonemes within a similar window, and every agent acquired thousands of words along a trajectory close to the corresponding child's, with a mean absolute timing error of 22 days across developmental milestones. The agents also recovered the age at which each child reached half of their measured vocabulary, which ranged across children from roughly 12 to 22 months; differences in the steepness of learning were captured only partially. Because the learner-internal parameters vary while the architecture does not, the model separates variation in the learner from variation in experience, nature interacting with nurture, rather than averaging children together.

The result is explicitly a starting point. The implemented model takes a single input stream, clean isolated speech, and produces a narrow outcome, a growing word-form lexicon. The next steps, already flagged in the original work, are the remaining stages of Objective 2: adding the visual scene a child sees while hearing speech, adding social context, and giving the model a body that acts, perceives, and adjusts. None of these requires a new theory, only the same approach applied one step at a time.

Perception: grounded experience rather than labeled data. Artificial systems can learn to associate what objects look like with the words used to describe them by encountering images and language together, without explicit category labels, extracting the correspondence from recurring structure in the input [41]. This

bears on one of psychology's oldest questions: how much of what infants perceive is built in, and how much can be extracted from grounded experience by a general-purpose learner [42]. Vong et al. [43] trained an agent to associate images and words using roughly 60 hours of recordings from a single child's perspective, and it acquired word-referent mappings from that experience alone. The debate remains open, in part because no learner has yet been trained on the dense, continuous, multimodal experience an infant receives across development [31].

Feedback and instruction: learning from another mind. Pretraining lets a model extract statistical structure from its input, but behavior is further shaped in post-training, when the model produces an output, receives an evaluation, and adjusts [44]. This is reinforcement learning from feedback, with a

clear human counterpart: a child does not passively absorb language but produces it and is corrected, praised, or ignored across many interactions with familiar people built on shared attention [45]. Training can also be ordered as a curriculum, with experiences introduced so as to support progressively more complex learning [46], a principle shared by formal schooling and by the naturally changing experience captured in recordings such as 1kD. The mismatch with industry lies again in the feedback: reinforcement learning from human feedback aggregates millions of anonymous, often one-time judgments into a single reward model, whereas a child receives repeated, context-sensitive feedback from a few familiar people across years.

Embodiment: the domain AI cannot circumvent. Advanced models extend the artificial agent's architecture into the physical world: systems that integrate vision, hearing, and language to generate motor commands [47]. This is embodied cognition made literal: the body structures action, action structures perception, and perception structures the next action [48–50]. Embodiment also closes a loop that perception alone leaves open. An agent that only observes can never learn whether its predictions were useful; acting and observing the consequence is itself a learning signal, long studied in motor control as a problem of prediction and correction rather than instruction [11].

World models: imagining what happens next. A further stage in the development of artificial systems moves beyond predicting the next word or pixel by constructing an internal model of how situations unfold, letting the system simulate outcomes before acting [51,52]. This resembles an important feature of human thought: mentally projecting how events may unfold, how an action may change a situation, or how another person may respond, before committing to a course of action [53,54]. World models therefore fold reasoning, planning, and social prediction into a single continuum, one that began with the statistical learning of sounds and lights.

Holding the ambition to a feasible standard

Industrial AI can already implement all five stages in some form. What it cannot do is implement them under the conditions a human encounters: the input, feedback, learning opportunities, and behavioral targets that a behavioral test, on its own, never constrains. Objective 2's contribution is to rebuild this developmental arc from feasible input and cognitively feasible learning protocols, which should be, at minimum, plausible hypotheses about how a human learner acquires the capacity in question, with the result judged against an individual developmental trajectory rather than a general leaderboard. Box 3 reports the first working link in that chain [55]: initial evidence that real input, a cognitively plausible mechanism, and a human behavioral target can together yield a mechanistic account of how individual children develop and of why they differ.

Behavioral alignment is more than an end in itself. If artificial and biological networks belong to the same family of solutions, an agent that learns from the same experience and arrives at the same behavior is a candidate for having reached it through shared computational principles [34,54]. Objective 2 therefore focuses on learning rather than performance: shared principles would be principles of how competence is acquired, not descriptions of competence once acquired. That learning is not a passive intake of data. It proceeds through continuous interaction between the brain, the body, and the

environment, so the trajectory an agent must reproduce is one of acting in a world and being changed by what comes back.

Box 4. Running the same cycle in another species

Building the brain of an adult animal means modeling the learning that shapes it across two nested timescales — the phylogenetic history that sets a species' priors and the ontogenetic learning from experiences and developmental processes that make them an individual [93]. Behavior lives in a body and a world, across both.

Nothing in the loop of Figure 1 is specific to humans. Zador et al. [91] made a kindred argument for AI at large, proposing an embodied Turing test that judges an artificial system against the sensorimotor competence of a living animal — an artificial beaver on dam-building rather than against a language or game benchmark. Transforming our loop into a different species turns the proposal into a research program and expands it, as we do for humans, by recording the natural behavior of animals (Objective 1), using this data to build virtual embodied agents that align behaviorally with the animals (Objective 2), and ensuring internal neural alignment (Objective 3) between the processes in the animals' brains and those in the agents' artificial brains.

The first two-thirds of the loop already exist for one species, in a form more rigorous than Objective 1 can reach in humans. Justin Wood's lab raises newly hatched chicks inside fully controlled visual environments from the moment of hatching, before a single instant of uncontrolled visual experience, removing by construction the consent and privacy limits that make comparable density impossible in a human infant. The same lab has taken the Objective 2 step: deep networks raised on the identical visual input the chicks received are being used to test whether a visual system is shaped by the statistics of its own developmental experience rather than built from innate structure [94]. Markerless tracking [95] now makes a dense behavioral recording routine in freely moving animals across species. What is missing is Objective 3. No one has yet recorded neural activity from these controlled-reared animals, and no one has asked whether an agent built from their exact developmental diet computes as their brains do. The behavioral and modeling infrastructure exists; the expertise to close the loop, neural recording in behaving animals and encoding models for a non-human cortex, mostly sits in different labs from those running the rearing studies. This is an invitation to extend the program into animal cognition.

Objective 3: Aligning artificial and biological neural computation

Feasible neural computation: the dimension behavior cannot reach. This dimension addresses what the other three cannot establish on their own. An agent can receive feasible input, learn through a feasible procedure, and behave within the trajectory and limits of human performance, and still arrive at that behavior by computing something the brain does not. Determining whether artificial and biological systems follow the same computational route requires direct comparison with measured neural activity.

Beyond correlation: what alignment is actually for: The standard tool is the encoding model: record neural activity during natural behavior and test whether an agent's internal activations robustly predict it. Networks trained on real-world statistics already capture real cortical structure and function.

Goal-driven vision networks predict primate visual cortex [56,57], and language models explain much of the variance in language cortex during natural listening [58,59], to the point that model-derived stimuli can drive or suppress that cortex on demand [60]. Our own electrocorticography (ECoG) work extends this line in three steps: the brain and an autoregressive model share the same predict-then-register-surprise objective [61];

the geometry of word meaning in key language areas matches a model's contextual embeddings closely enough to predict held-out words the model never saw in training [62]; and a model's layer hierarchy maps onto the brain's own processing-timescale hierarchy, with deeper layers tracking later activity in Broca's area [63].

A predictive match, however close, is a measurement rather than an explanation: it locates where and how strongly two systems align, without saying why, or how to build a model that aligns more closely. Alignment must therefore move from comparing outputs to comparing the internal machinery that produces them. Richards et al. [64] define an artificial network by three design choices: the objective it minimizes, the architecture that performs the computation, and the learning rule that updates its connections. We add a fourth that the standard three treat as given: the data the system learns from. This quartet describes a model's internals and is not the same as the four feasibility dimensions above, which describe what a model must match in a developing human; the two meet at one point, since the experience a network learns from is exactly what feasible input supplies. In the sense proposed by Doerig et al. [65], the approach treats artificial networks as a language for stating falsifiable theories of brain computation.

What a brain-like architecture would require: If artificial and biological networks belong to the same family of computational solutions, artificial models must eventually take on the features that define the biological ones: spiking rather than continuous units, recurrent rather than purely feedforward dynamics [34], biologically realizable learning rules [66], circuit motifs that mirror cortical and subcortical anatomy, and something approaching the brain's extraordinary energy efficiency [67]. The payoff of building even one such feature is already visible. When Margalit and colleagues added a single biological requirement to a task-driven vision network- that nearby units respond similarly, as neighboring neurons do across the cortical sheet- the model spontaneously reproduced the topography of the visual system, from orientation pinwheels in early areas to face- and place-selective patches downstream, at reduced wiring cost [68]. One constraint borrowed from anatomy turned a predictive model into an explanation of why the visual cortex is laid out as it is. But one constraint is a start, not an architecture. Today's mainstream models incorporate almost none of the brain's defining features, and the gap is starkest for the learning rule. Standard deep learning minimizes a single global loss by backpropagation, broadcast across the whole network at once, a procedure with no known biological implementation, since nothing in the cortex delivers a precise network-wide error signal to a real synapse [69]. The most comprehensive alternative is local and error-driven: each synapse adjusts on the difference between two nearby activity states it can actually observe, timed by cortical rhythms and routed through corticothalamic circuits. This rule was recently shown to approximate backpropagation closely enough to learn at scale while remaining implementable in real circuitry [70]. It delivers three of the four design choices in biologically realizable form: an objective computed locally, an architecture that gives the thalamus a real role rather than treating it as a relay, and a learning rule carried by sparse, event-driven

spikes. The fourth, the data, is what this paper has argued for since Objective 1: experience delivered as a nervous system receives it, bounded, ordered, and in context.

Closing the loop from within: A large language model can hold an enormous context in view at once, now hundreds of thousands of words, weighing every word against every other in parallel. The brain does neither. Across most of the cortical hierarchy, the window over which a region integrates incoming information is short, often shorter than a single paragraph [71]. That yields a direct prediction: if cortical integration windows are that short, a model given only a 30-word context should predict neural activity during natural listening as well as one given a thousand words or more. The extra context should add nothing, which is what we found [72]: measuring cortical responses while people listened to a 30-minute story, widening the model's contextual window did not improve the fit. Yet the brain's capacity to integrate over time is not short. We bind each paragraph to the one before it, to the idea that opened the story, and to knowledge acquired over years. The brain achieves that by a different mechanism than a flat context window: it compresses the past into an episodic summary, plausibly through the hippocampal system, and retrieves it to combine with the short

online window. This is the full Objective 3 cycle. A divergence between model and brain reveals a missing computational principle, and the more plausible model must integrate over long timescales as the brain does, by summarizing and retrieving rather than holding everything in view.

How understanding emerges from building: Building the agent provided us with a cognitive insight that neither the recordings nor existing theories provided. A system learning continuously from a non-repeating stream cannot revise itself continuously without destabilizing; it must separate an acting mode that accumulates error from a learning mode in which weights are updated. The brain does this by deferring the adjustment of synaptic weights and concentrating it during sleep. Similarly, our agent that lacks neurotransmitters and does not require sleep still faces the same pressure, which forces it to dissociate action from learning, which makes this a claim about computation rather than biology (Box 5).

Box 5. What building revealed about deferred learning

Why should a learner ever stop learning to consolidate what it has learned? Memory research established that updating the synaptic weights takes time. Building an agent revealed why they might have to. The 1kD agent learns from one child's recorded experience, day after day, in order. To align and track the agent's development against the child's, training was paused each week and the agent tested on the following week's input before training resumed [55]. That training schedule, imposed by the need to assess the agent’s ability to act, turned out to be illuminating.

Children are always learning, and our agent also cannot stop learning. Industrial systems avoid the problem of continuous learning by training once and then freezing the model, so learning and use never have to happen at the same time. Our agent, like a child, must keep learning for as long as it runs while also processing incoming speech. What building revealed is that deferring learning and dissociating it from online processing solves this. Deferral appeared across timescales. Within minutes, the agent accumulates prediction error rather than acting on it, committing only after sufficient information was gathered. Across the day, those commitments stay provisional. Overnight, the day's events are revisited

by replay, and most weight updates are made. Without that nightly replay, the agent learns almost nothing.

The brain's cascade has a similar dynamic. Initial potentiation is electrical and volatile, decaying within minutes unless stabilized; stabilization requires protein synthesis and takes hours; structural remodeling extends over hours to days; and systems-level consolidation, redistributing memories from hippocampus to cortex, unfolds over weeks, driven by reactivation during sleep [96,97]. This literature establishes that consolidation is deferred, not why a learning system would be built that way. The agent offers an answer at the level of computation: a system learning in real time from the continuous, non-repeating stream of an ever-changing life cannot reweight itself continuously without destabilizing, so it must accumulate error while acting and defer its updates to a separate phase. Complementary learning systems theory reached a kindred separation from a different direction, the logic of interference rather than the processing demands of a continuous input stream [98,99]. The convergence of two independent routes is itself informative.

Why controlled experiments remain indispensable

A reasonable objection runs through everything proposed here: prediction is not explanation. An agent whose internal activity predicts a cortical response has demonstrated, by itself, only a correlation. A system can produce the right output, or even the right pattern of neural activity, for reasons unrelated to how the brain produces it. Controlled experiments exist precisely to break the confounds that a predictive fit silently absorbs. Abandoning them for naturalistic modeling, the objection concludes, would be to mistake a curve that fits for a mechanism that explains.

We accept the premise but not the conclusion, because the goal of this framework is not predictive fit. The agents of Objective 2 are not statistical summaries of experimental or neural data; they are mechanistic models that learn from real input to produce natural behavior under cognitively feasible constraints. Once built, such an agent can be brought into the lab and subjected to the same controlled experiments as a human participant: deprived of a source of input, presented with a minimal contrast, or tested on a transfer task it never trained on. Unlike a human participant, it can also be reset, rerun, lesioned, and inspected weight by weight. A controlled experiment can show that a variable matters; a working agent can supply a candidate mechanism by which that variable interacts with others to support behavior in the wild, and show where it breaks down. It gives the experimentalist a second kind of subject that is built rather than observed.

None of this makes the resulting models self-explanatory. Even a system we build ourselves can produce emergent behavior that surprises us, and a complete wiring diagram is not an account of what the system computes. Building is not an alternative to theory or experiment but a way of making a theory precise enough to test and falsify. Predictive coding is a useful example. The standard question, does the brain predict, has produced many elegant experiments, but each is tied to a particular task, stimulus, and population, so it is hard to know which results generalize. The constructive alternative is to build an agent whose core computation is prediction, train it on the ordered natural input humans actually receive, and ask whether its behavior and internal dynamics align with a person's and a brain's. When alignment holds, prediction is no longer a hypothesis tested one experiment at a time; it is an

implemented mechanism. When it fails, the failure reveals what the hypothesis lacks and forces a rebuild.

Brunswik's insistence on representative sampling and Fisher's insistence on systematic manipulation [73] were treated for a century as rival philosophies; in this framework they become consecutive operations on the same problem. Naturalistic modeling identifies which computational principles matter and supplies a mechanism for how the mind and brain learn to handle the messiness of real life. A model trained on real experience then generates hypotheses precise enough to justify the cost of a designed experiment, and that experiment, run on a human, an agent, or both, reveals whether the proposed mechanism is the right one. The limitation was never in controlled experiments but in treating them as the core goal rather than as one step toward models that act in the real world.

The three objectives do not demand the same kind of data. Objective 2 demands density: an agent that develops must be trained on a substantial fraction of one individual's real experience, which is why Objective 1 pays the steep price of recording a life. Objective 3 demands something else. To test whether a trained model computes as the brain does, breadth matters more than duration: record from as many neurons and areas as possible, but an hour of natural experience at a time may suffice, provided we sample across as many contexts as we can reach. The target is the model's competence measured against human generalization: how well a system built from one developmental history holds up across situations a person handles effortlessly. Seen this way, bringing a model into the lab is one more context, stripped down, among the many the brain takes in stride. Density builds the model; diversity tests it.

Toward mechanistic models of atypical minds

The same logic that lets us model a typical mind and brain also lets us probe the mechanisms behind atypical development [74,75]. If a cognitively and neurally feasible model reproduces a typical trajectory from an individual's real experience, it becomes a reference against which atypical cases can be interrogated, not by cataloging correlated risk factors or population averages [76], but by asking a sharper, individual-level question: which perturbation is sufficient to reproduce the atypical outcome [77]? That depends on knowing the typical range [78]. Behavioral diversity among typical subjects along the dimensions of Objective 1 sets the baseline against which any divergence is measured, as a trajectory that reads as atypical in one population may be unremarkable in another [79].

The four dimensions of feasibility supply the axes along which an artificial learner can diverge from a typical individual: the input it receives, the procedure by which it learns, the behavior it is aimed to produce, and the neural computation it implements. Each perturbation is a candidate mechanistic account in the sense used throughout this article: a model capable of producing the phenomenon rather than merely describing it. But generating a candidate mechanism is half the task. Each perturbation is a hypothesis, and it is valid only if it accounts for a person's atypical behavior and their atypical brain activity [80,81]. This alignment of behavioral and neural patterns, which we require in typical cases too, is what makes it possible to understand how a specific clinical condition develops, from developmental disorders and learning difficulties [82] to addiction [83] and a broad range of psychiatric and neurological conditions.

Concluding remarks and future perspectives

Scientific paradigms rarely shift because someone proves the old one wrong [84]. They shift when new methods make new questions answerable. Only then does it become clear that the old framework, for all its power, had answered a narrower question than anyone intended. The controlled experiment exemplifies this limitation. It remains the appropriate method for demonstrating that a variable has an effect, but it cannot fully explain how such an effect is processed by a real mind, embodied in a real person, learning from the continuous structure of real-life experience.

For the first time, the three capabilities required to attempt such an explanation exist at once. We can record a human life in dense detail as it unfolds. We can build an agent that learns from that experience under cognitively feasible constraints. And we can test whether the agent computes as the brain does. None of these is entirely new on its own; their convergence is, and it changes what counts as a complete theory of the mind. A theory of language acquisition should now explain how individual children, developing in particular, one-of-a-kind homes, come to speak as they do, not only how children acquire language on average [23]. A theory of neural computation should predict the activity of a specific brain region at a particular moment during natural behavior, and then account for that activity mechanistically. By building learning agents trained on naturalistic recordings of single children, the 1kD Project provides an initial feasibility test of this framework. Whether such agents also converge with biological neural computation is the next critical test (see Outstanding Questions).

No single lab or scientific field can build this framework alone. There comes a stage in the life of a field when understanding can no longer be assembled by a lone investigator and must instead be built at scale. As particle physics matured, thousands of scientists pooled their efforts to build accelerators capable of testing its theories. Once neural network theory began to work, thousands more built the systems that turned it into modern AI. We therefore invite researchers across fields to join forces: to record real lives, to build agents that learn from them, and to test those agents against human behavior and the brain. Feynman held that we understand only what we can create. For psychologists and neuroscientists, creation means building models that act as we do. For a century, cognitive science could observe the mind and record the brain but could not build a system that developed under comparable conditions, and so it held itself to a lesser standard: to describe, to correlate, and to fit a dataset after the fact. That constraint is beginning to lift. Whether an agent that learns and behaves as we do also computes as we do remains the falsifiable claim this framework is designed to test. But it is now a question we are equipped to address. We will understand the mind to the extent we can build it.

Outstanding Questions

Expanding human recording (Objective 1).

● Can the 1kD cohort grow from seventeen children in fifteen households to a thousand children spanning more cultures, languages, incomes, and clinical backgrounds, holding age and setting fixed so that variation across people can be isolated?

● What further sensing is needed to capture the bodily side of experience: individual vision, gaze, and head motion from wearable glasses, and physiological signals such as heart-rate variability, skin temperature, breathing rate, and skin conductance?

● Which inputs can be pooled across individuals by leveraging diversity, and which require dense sampling of each individual to be modeled accurately?

● Can a synthetic corpus of fictional families, matched statistically to real developmental data, train agents while avoiding the ethical and logistical costs of recording real lives?

Extending the agent (Objective 2).

● Within an actor–critic reinforcement-learning framework, how does closing the loop between perceptual input and motor output reshape cognition?

● Can an agent be trained inside a virtual world model, then given a body in a matched environment, and how much of the explainable variance in human and animal behavior can such a brain–body–environment system capture?

● Can the framework extend to multiple autonomous agents, virtual or robotic, that interact and thereby develop social cognition?

Deepening neural alignment (Objective 3).

● Which brain structures, added to the model, would extend the range of cognition it captures: a cortical hierarchy for integrating across events, a hippocampus for episodic memory, or thalamus–basal ganglia circuits for combining error signals during learning?

● What is the right level of alignment? A replica is not the goal, but which features must match: architecture, biologically plausible objectives, learning rules? Must the network spike, or be recurrent? The answer will depend on the modeling goal.

Glossary

Behavioral alignment: agreement between an agent and a person in what they do, assessed against an individual's developmental trajectory rather than an aggregate benchmark.

Cathedral problem: the difficulty of designing a study that must outlive the careers, funding cycles, and institutions of everyone who begins it.

Cognitively feasible: satisfying all four dimensions along which a model must match a developing human, namely feasible input, feasible learning procedure, feasible behavioral target, and feasible neural computation.

Density: the proportion of an individual's ongoing experience that a dataset actually captures, and the resolution at which it captures it.

Diversity: the range of people, cultures, environments, and circumstances a dataset spans.

Ecological validity predicament: the failure of findings established under controlled conditions to generalize to the natural settings the mind evolved to navigate.

Encoding model: a model that predicts measured neural activity from the internal activations of a computational system processing the same input.

Mechanistic account: a complete, implementable model that produces the phenomenon of interest, as distinct from a description of the conditions under which the phenomenon occurs.

Neural alignment: correspondence between the internal computations of an artificial agent and the measured activity of the brain during natural behavior.

Replication crisis: the difficulty of reproducing published findings across laboratories, samples, and slightly altered tasks.

Representative sampling: Brunswik's requirement that the situations sampled in a study represent those the organism actually encounters.

Acknowledgments

I thank the past and present members of my lab at Princeton University, whose ideas and work have shaped this research program over many years. I am especially grateful to Liat Hasenfratz, Arkadii Tsyhanov, and Casey Lew-Williams for their essential contributions to the 1,000 Days Project, and to Sam Nastase, Itamar Jalon, Zaid Zada, and Ken Wang for their work on neural alignment. I thank Google Research, and in particular Mariano Schain, Daria Lioubashevsky, Avinatan Hassidim, Yossi Matias, Michael Brenner, and Blaise Agüera y Arcas, for supporting this research and for helping us deepen our understanding of deep learning. We thank Jeanne DeVoe for editorial assistance with the manuscript. Finally, I thank Elizabeth L. Paluck, David Poeppel and Justin Newell Wood for their thoughtful comments on the manuscript.

Declaration of generative AI in the manuscript preparation process

During the preparation of this work, the authors used Claude (Anthropic) for editorial assistance, including language refinement and reference formatting. All ideas, arguments, and scientific claims are the authors' own. The authors reviewed and edited the output as needed and took full responsibility for the content of the published article.

References

1. Fechner, G.T. (1860) Elemente der Psychophysik, Breitkopf & Härtel.

2. Wundt, W. (1874) Grundzüge der physiologischen Psychologie, Wilhelm Engelmann.

3. Ebbinghaus, H. (1885) Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie, Duncker & Humblot.

4. Miller, G.A. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63, 81–97.

5. Chomsky, N. (1957) Syntactic Structures, Mouton.

6. Sutton, R. (2019) The bitter lesson. Incomplete Ideas (blog), incompleteideas.net/IncIdeas/BitterLesson.html.

7. Fodor, J.A. (1983) The Modularity of Mind, MIT Press.

8. Kanwisher, N., McDermott, J., and Chun, M.M. (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

9. Kanwisher, N. (2017) The quest for the FFA and where it led. J. Neurosci. 37, 1056–1061.

10. Kahana, M.J. (2020) Computational models of memory search. Annu. Rev. Psychol. 71, 107–138.

11. Wolpert, D.M., Ghahramani, Z., and Jordan, M.I. (1995) An internal model for sensorimotor integration. Science 269, 1880–1882.

12. Chen, Y. and Taylor, J.A. (2025) Spatial dynamics shape the interaction between motor adaptation processes. In Proc. 47th Annu. Meet. Cogn. Sci. Soc. (CogSci).

13. Henrich, J., Heine, S.J., and Norenzayan, A. (2010) The weirdest people in the world? Behav. Brain Sci. 33, 61–83.

14. Pashler, H. and Harris, C.R. (2012) Is the replicability crisis overblown? Three arguments examined. Perspect. Psychol. Sci. 7, 531–536.

15. Shamay-Tsoory, S.G. and Mendelsohn, A. (2019) Real-life neuroscience: an ecological approach to brain and behavior research. Perspect. Psychol. Sci. 14, 841–859.

16. Brunswik, E. (1947) Systematic and Representative Design of Psychological Experiments, University of California Press.

17. Neisser, U. (1976) Cognition and Reality: Principles and Implications of Cognitive Psychology, W.H. Freeman.

18. Nastase, S.A., Goldstein, A., and Hasson, U. (2020) Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. NeuroImage 222, 117254.

19. Sonkusare, S. et al. (2019) Naturalistic stimuli in neuroscience: critically acclaimed. Trends Cogn. Sci. 23, 699–714.

20. Ulanovsky, N. (2025) Natural Neuroscience: Toward a Systems Neuroscience of Natural Behaviors, MIT Press.

21. Geirhos, R. et al. (2019) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In Int. Conf. Learn. Represent. (ICLR).

22. Bowers, J.S. et al. (2023) Deep problems with neural network models of human vision. Behav. Brain Sci. 46, e385.

23. Ghazanfar, A.A. and Gomez-Marin, A. (2024) The central role of the individual in the history of brains. Neurosci. Biobehav. Rev. 163, 105744.

24. Roy, B.C., Frank, M.C., DeCamp, P., Miller, M., and Roy, D. (2015) Predicting the birth of a spoken word. Proc. Natl. Acad. Sci. USA 112, 12663–12668.

25. Adolph, K.E. et al. (2012) How do you learn to walk? Thousands of steps and dozens of falls per day. Psychol. Sci. 23, 1387–1394.

26. Raviv, H. et al. (2026) The First 1,000 Days (1kD) Project: collecting and analyzing an ultra-dense naturalistic dataset of human baby development. bioRxiv. https://doi.org/10.64898/2026.03.19.712982.

27. Roy, D. et al. (2006) The Human Speechome Project. In Proc. 28th Annu. Conf. Cogn. Sci. Soc..

28. Gilkerson, J. and Richards, J.A. (2008) The LENA Natural Language Study (Technical Report LTR-02-2), LENA Foundation.

29. Bergelson, E. et al. (2023) Everyday language input and production in 1,001 children from six continents. Proc. Natl. Acad. Sci. USA 120, e2300671120.

30. Grauman, K. et al. (2022) Ego4D: around the world in 3,000 hours of egocentric video. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 18995–19012.

31. Long, B. et al. (2024) The BabyView dataset: high-resolution egocentric videos of infants' and young children's everyday experiences. arXiv 2406.10447.

32. Kane, T.J. and Staiger, D.O. (2012) Gathering Feedback for Teaching (MET Project Research Paper), Bill & Melinda Gates Foundation.

33. Engel, J. et al. (2023) Project Aria: a new tool for egocentric multi-modal AI research. arXiv 2308.13561.

34. Hasson, U., Nastase, S.A., and Goldstein, A. (2020) Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. Neuron 105, 416–434.

35. Sander, M. and Peyré, G. (2025) Towards understanding the universality of transformers for next-token prediction. In Int. Conf. Learn. Represent. (ICLR).

36. Vaswani, A. et al. (2017) Attention is all you need. Adv. Neural Inf. Process. Syst. 30, 5998–6008.

37. Sutton, R.S. and Barto, A.G. (2018) Reinforcement Learning: An Introduction (2nd edn), MIT Press.

38. Brown, T.B. et al. (2020) Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

39. Millidge, B., Seth, A., and Buckley, C.L. (2021) Predictive coding: a theoretical and experimental review. arXiv 2107.12979.

40. Saffran, J.R., Aslin, R.N., and Newport, E.L. (1996) Statistical learning by 8-month-old infants. Science 274, 1926–1928.

41. Radford, A. et al. (2021) Learning transferable visual models from natural language supervision. In Proc. 38th Int. Conf. Mach. Learn. (ICML), pp. 8748–8763.

42. Spelke, E.S. and Kinzler, K.D. (2007) Core knowledge. Dev. Sci. 10, 89–96.

43. Vong, W.K., Wang, W., Orhan, A.E., and Lake, B.M. (2024) Grounded language acquisition through the eyes and ears of a single child. Science 383, 504–511.

44. Ouyang, L. et al. (2022) Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 35, 27730–27744.

45. Tomasello, M. and Carpenter, M. (2007) Shared intentionality. Dev. Sci. 10, 121–125.

46. Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009) Curriculum learning. In Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML), pp. 41–48.

47. Brohan, A. et al. (2023) RT-2: vision-language-action models transfer web knowledge to robotic control. In Proc. Conf. Robot Learn. (CoRL).

48. Pfeifer, R. and Bongard, J. (2006) How the Body Shapes the Way We Think: A New View of Intelligence, MIT Press.

49. Spencer, J.P. et al. (2009) Short arms and talking eggs: why we should no longer abide the nativist–empiricist debate. Child Dev. Perspect. 3, 79–87.

50. Smith, L.B. and Gasser, M. (2005) The development of embodied cognition: six lessons from babies. Artif. Life 11, 13–29.

51. Ha, D. and Schmidhuber, J. (2018) World models. arXiv 1803.10122.

52. LeCun, Y. (2022) A path towards autonomous machine intelligence. OpenReview (position paper).

53. Goyal, A. and Bengio, Y. (2022) Inductive biases for deep learning of higher-level cognition. Proc. R. Soc. A 478, 20210068.

54. Saxe, A., Nelli, S., and Summerfield, C. (2021) If deep learning is the answer, what is the question? Nat. Rev. Neurosci. 22, 55–67.

55. Raviv, H. et al. (2026) The richness of experience: agent-based learning reveals the mechanisms of early language acquisition. bioRxiv. https://doi.org/10.64898/2026.03.28.715023.

56. Yamins, D.L.K. et al. (2014) Performance-optimized hierarchical models predict neural responses in the higher visual cortex. Proc. Natl. Acad. Sci. USA 111, 8619–8624.

57. Yamins, D.L.K. and DiCarlo, J.J. (2016) Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. 19, 356–365.

58. Schrimpf, M. et al. (2021) The neural architecture of language: integrative modeling converges on predictive processing. Proc. Natl. Acad. Sci. USA 118, e2105646118.

59. Tuckute, G., Kanwisher, N., and Fedorenko, E. (2024) Language in brains, minds, and machines. Annu. Rev. Neurosci. 47, 277–301.

60. Tuckute, G. et al. (2024) Driving and suppressing the human language network using large language models. Nat. Hum. Behav. 8, 544–561.

61. Goldstein, A. et al. (2022) Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. 25, 369–380.

62. Goldstein, A. et al. (2024) Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. Nat. Commun. 15, 2768.

63. Goldstein, A. et al. (2025) Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. Nat. Commun. 16, 10529.

64. Richards, B.A. et al. (2019) A deep learning framework for neuroscience. Nat. Neurosci. 22, 1761–1770.

65. Doerig, A. et al. (2023) The neuroconnectionist research programme. Nat. Rev. Neurosci. 24, 431–450.

66. Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., and Hinton, G. (2020) Backpropagation and the brain. Nat. Rev. Neurosci. 21, 335–346.

67. Roy, K., Jaiswal, A., and Panda, P. (2019) Towards spike-based machine intelligence with neuromorphic computing. Nature 575, 607–617.

68. Margalit, E., Lee, H., Finzi, D., DiCarlo, J.J., Grill-Spector, K., and Yamins, D.L.K. (2024) A unifying framework for functional organization in early and higher ventral visual cortex. Neuron 112, 2435–2451.

69. Whittington, J.C.R. and Bogacz, R. (2019) Theories of error back-propagation in the brain. Trends Cogn. Sci. 23, 235–250.

70. O'Reilly, R.C. (2026) This is how the neocortex learns. arXiv 2606.08720.

71. Hasson, U., Chen, J., and Honey, C.J. (2015) Hierarchical process memory: memory as an integral component of information processing. Trends Cogn. Sci. 19, 304–313.

72. Tikochinski, R., Goldstein, A., Meiri, Y., Hasson, U., and Reichart, R. (2025) Incremental accumulation of linguistic context in artificial and biological neural networks. Nat. Commun. 16, 803.

73. Fisher, R.A. (1935) The Design of Experiments, Oliver and Boyd.

74. Karmiloff-Smith, A. (1998) Development itself is the key to understanding developmental disorders. Trends Cogn. Sci. 2, 389–398.

75. Thomas, M.S.C. and Karmiloff-Smith, A. (2002) Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. Behav. Brain Sci. 25, 727–788.

76. Astle, D.E., Holmes, J., Kievit, R., and Gathercole, S.E. (2022) Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders. J. Child Psychol. Psychiatry 63, 397–417.

77. Harm, M.W. and Seidenberg, M.S. (1999) Phonology, reading acquisition, and dyslexia: insights from connectionist models. Psychol. Rev. 106, 491–528.

78. Marquand, A.F., Rezek, I., Buitelaar, J., and Beckmann, C.F. (2016) Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. Biol. Psychiatry 80, 552–561.

79. Bethlehem, R.A.I. et al. (2022) Brain charts for the human lifespan. Nature 604, 525–533.

80. Montague, P.R., Dolan, R.J., Friston, K.J., and Dayan, P. (2012) Computational psychiatry. Trends Cogn. Sci. 16, 72–80.

81. Huys, Q.J.M., Maia, T.V., and Frank, M.J. (2016) Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. 19, 404–413.

82. Warlaumont, A.S., Richards, J.A., Gilkerson, J., and Oller, D.K. (2014) A social feedback loop for speech development and its reduction in autism. Psychol. Sci. 25, 1314–1324.

83. Redish, A.D. (2004) Addiction as a computational process gone awry. Science 306, 1944–1947.

84. Kuhn, T.S. (1962) The Structure of Scientific Revolutions, University of Chicago Press.

85. Chomsky, N. (1965) Aspects of the Theory of Syntax, MIT Press.

86. Tomasello, M. (2003) Constructing a Language: A Usage-Based Theory of Language Acquisition, Harvard University Press.

87. Goldberg, A.E. (2024) Usage-based constructionist approaches and large language models. Constr. Frames 16, 220–254.

88. Kuhl, P.K. (2007) Is speech learning ‘gated’ by the social brain? Dev. Sci. 10, 110–120.

89. Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017) Neuroscience-inspired artificial intelligence. Neuron 95, 245–258.

90. Kriegeskorte, N. (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. 1, 417–446.

91. Zador, A. et al. (2023) Catalyzing next-generation artificial intelligence through NeuroAI. Nat. Commun. 14, 1597.

92. Zipf, G.K. (1949) Human Behavior and the Principle of Least Effort, Addison-Wesley.

93. Gomez-Marin, A. and Ghazanfar, A.A. (2019) The life of behavior. Neuron 104, 25–36.

94. Pandey, L., Lee, D., Wood, S.M.W., and Wood, J.N. (2024) Parallel development of object recognition in newborn chicks and deep neural networks. PLoS Comput. Biol. 20, e1012600.

95. Mathis, A. et al. (2018) DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. 21, 1281–1289.

96. Diekelmann, S. and Born, J. (2010) The memory function of sleep. Nat. Rev. Neurosci. 11, 114–126.

97. Dudai, Y. (2012) The restless engram: consolidations never end. Annu. Rev. Neurosci. 35, 227–247.

98. McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. 102, 419–457.

99. Kumaran, D., Hassabis, D., and McClelland, J.L. (2016) What learning systems do intelligent agents need? Complementary learning systems theory updated. Trends Cogn. Sci. 20, 512–534.