ANNUAL REVIEWS

*Annual Review of Vision Science*

# Digital Twin Studies for Reverse Engineering the Origins of Visual Intelligence

Justin N. Wood,[1,2,3] Lalit Pandey,[1] and Samantha M.W. Wood[1,2,3]

[1]Informatics Department, Indiana University Bloomington, Bloomington, Indiana, USA; email: woodjn@indiana.edu, lpandey@iu.edu, sw113@iu.edu

[2]Cognitive Science Program, Indiana University Bloomington, Bloomington, Indiana, USA

[3]Neuroscience Department, Indiana University Bloomington, Bloomington, Indiana, USA

Annu. Rev. Vis. Sci. 2024. 10:145–70

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-101322-103628



## Keywords

newborn, controlled rearing, nativism, empiricism, artificial intelligence, digital twin

## Abstract

What are the core learning algorithms in brains? Nativists propose that intelligence emerges from innate domain-specific knowledge systems, whereas empiricists propose that intelligence emerges from domain-general systems that learn domain-specific knowledge from experience. We address this debate by reviewing digital twin studies designed to reverse engineer the learning algorithms in newborn brains. In digital twin studies, newborn animals and artificial agents are raised in the same environments and tested with the same tasks, permitting direct comparison of their learning abilities. Supporting empiricism, digital twin studies show that domain-general algorithms learn animal-like object perception when trained on the first-person visual experiences of newborn animals. Supporting nativism, digital twin studies show that domain-general algorithms produce innate domain-specific knowledge when trained on prenatal experiences (retinal waves). We argue that learning across humans, animals, and machines can be explained by a universal principle, which we call space-time fitting. Space-time fitting explains both empiricist and nativist phenomena, providing a unified framework for understanding the origins of intelligence.

## 1. INTRODUCTION

The brain is the most powerful learning system in the known universe. Using just a few dozen watts of energy—barely enough to run a dim light bulb—brains learn to perform a wide range of tasks, efficiently transforming streams of high-dimensional sensory data into adaptive behavioral responses. What core algorithms—present in newborn brains—drive these impressive feats of learning and behavior? This question has inspired philosophers and scientists for more than 2,000 years, from Plato and Aristotle to Hume, Locke, von Helmholtz, and many others in the twenty-first century. However, the origins and computational foundations of intelligence remain unsolved mysteries in science and technology. There is disagreement about when intelligence begins, what it consists of, what causes it to emerge, and how it changes with experience.

We propose an approach to resolve this debate by linking methods from developmental psychology, computational neuroscience, virtual reality, and artificial intelligence. We first discuss methodological challenges in resolving the nativist versus empiricist debate (Section 2) and argue that controlled-rearing studies will be essential for characterizing the brain's core learning algorithms (Section 3). We focus on efforts to automate controlled-rearing studies (Section 3.1), which allows newborn animals to be raised in unnatural visual worlds. This methodological advance has allowed us to characterize the initial state of object perception and determine which visual experiences are necessary and sufficient to develop object perception (Section 3.2).

We then describe conceptual challenges in resolving the nativist versus empiricist debate and the need for closed-loop experimental systems for building robust theories of the origins of intelligence (Section 4). Building on recent attempts to reverse engineer sensory systems (Section 4.1), we extend the reverse-engineering paradigm to newborn animals (Section 4.2). The resulting digital twin paradigm involves performing parallel controlled-rearing experiments on newborn animals and artificial agents in a closed-loop scientific system (**Figure 1**). By raising animals and machines in the same environments and testing them with the same tasks, we can directly measure whether nativist or empiricist algorithms learn more like newborn animals (Section 4.3). We show that, when domain-general learning algorithms—including convolutional neural networks (CNNs) (Section 4.4) and vision transformers (ViTs) (Section 4.5)—are trained on the first-person visual experiences of newborn animals, the algorithms learn animal-like object perception, confirming a key empiricist prediction. We also show that domain-general algorithms learn domain-specific visual knowledge when trained solely on retinal waves, illuminating how domain-general algorithms could create innate knowledge in newborn brains (Section 4.6). We posit that visual learning across humans, animals, and machines can be understood in terms of a universal principle that we call space-time fitting, in which visual systems spontaneously adapt (fit) to the spatiotemporal data distributions in the organism's prenatal and postnatal environment. Space-time fitting integrates concepts related to learning in brains (e.g., Hebbian learning, spike timing–dependent plasticity, predictive coding, statistical learning) and learning in machines (e.g., backpropagation, generative modeling, deep reinforcement learning) under a common principle. These different concepts reflect the same general principle of high-dimensional systems iteratively adapting (fitting) to the spatiotemporal data distributions underlying sensory experiences.

We conclude by returning to the nativist versus empiricist debate, emphasizing that space-time fitting aligns with both nativist and empiricist views (Section 5). Empiricists have long posited that powerful domain-general learning mechanisms acquire domain-specific knowledge through experience, consistent with a space-time fitting process (Section 5.1). Nativists have long posited the existence of rich innate knowledge, consistent with a space-time fitting process learning from prenatal training data. We argue that space-time fitting can unite nativist and empiricist

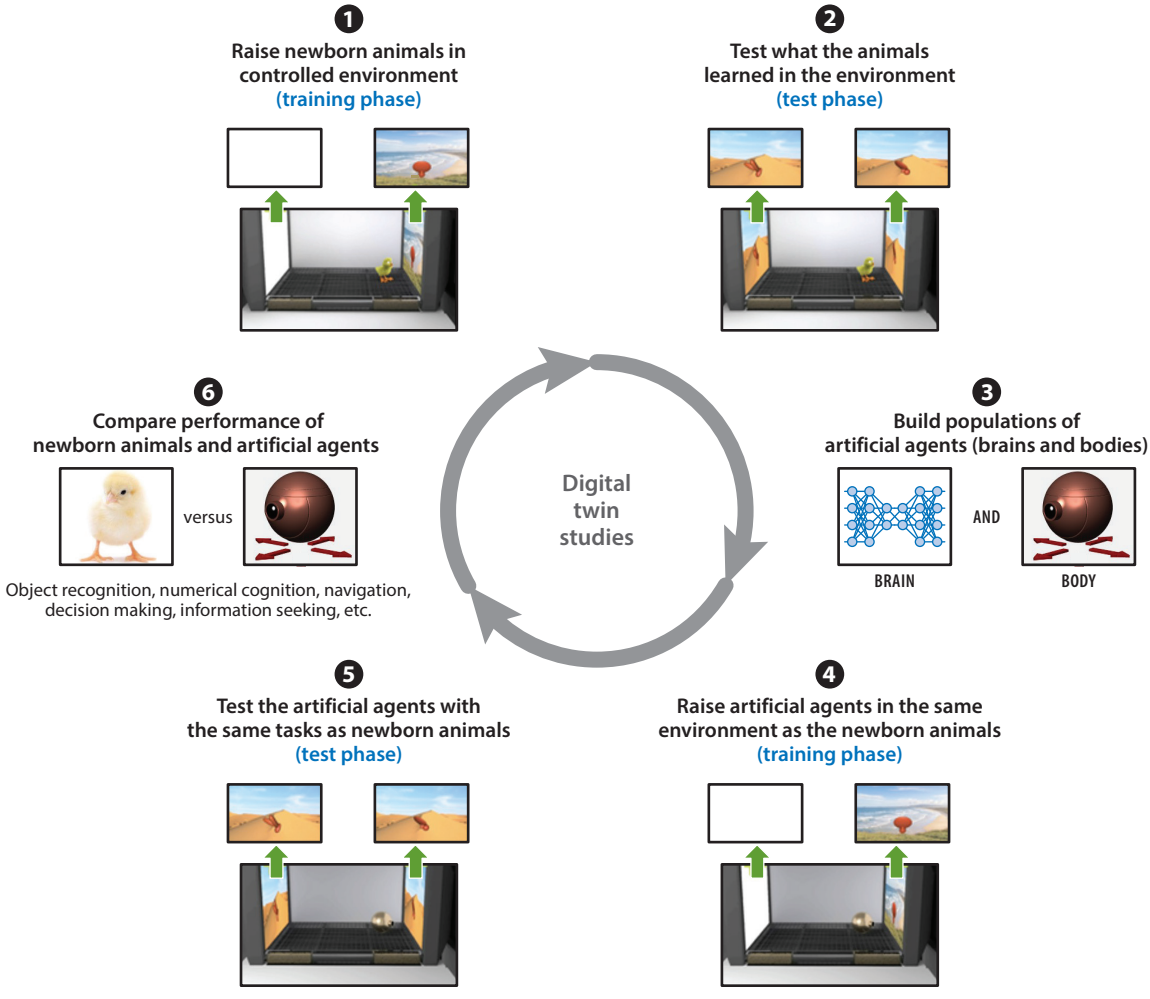



**Figure 1**

Digital twin studies involve raising newborn animals and artificial agents in the same environments and testing them with the same tasks. This allows animals and machines to learn from the same experiences (training data), permitting direct comparison of their learning abilities. Digital twin studies can be created for any perceptual, cognitive, or motor task. Figure adapted from Garimella et al. (2024).

perspectives, leading to a unified understanding of the origins and computational basis of intelligence (Section 5.2).

## 2. NATIVIST VERSUS EMPIRICIST THEORIES

Theories of the origins of intelligence broadly fall along a continuum that differs in terms of the number and domain specificity of core learning systems (Buckner 2023). At one end of the continuum lie nativist theories, which posit that intelligence emerges from a collection of innate, domain-specific systems for learning about different kinds of things. Some researchers have proposed dozens of specialized, evolutionarily ancient learning mechanisms (Pinker 2002), while others have proposed a more modest number of core systems for learning about broad classes of things (e.g., objects, places, agents, number) (Carey 2009, Spelke 2022). Nativist theories typically

assume that a newborn's sensory data are sparse, noisy, and impoverished; as such, nativists argue that the core learning algorithms in newborn brains must have strong domain-specific knowledge to solve perceptual and cognitive tasks.

At the other end of the continuum lie empiricist theories, which posit that intelligence emerges from domain-general learning algorithms. For instance, theorists such as Locke, Hume, Skinner, and Watson proposed that newborns start with a few domain-general faculties, and these faculties serve as the foundation for all other (learned) mental faculties (Hume 1739, Locke 1690, Skinner 1938, Watson 1913). In contrast to nativist theories, empiricists typically assume that newborns have access to richly structured sensory data for learning about the world. This is an important difference between theories because, if training data are sufficiently rich, then domain-general algorithms can learn domain-specific knowledge through experience (Reed et al. 2022). Both empiricist and nativist theories agree that mature animals have domain-specific knowledge; the theories differ in terms of whether domain-specific knowledge is learned versus hardcoded into brains by evolutionary processes.

This debate has generated decades of seminal research about the cognitive abilities of infants, toddlers, and children. However, scientists have long recognized that the only way to determine what knowledge and learning mechanisms are present at birth is to study newborns. Studying newborns circumvents much of the complexity associated with mature organisms, providing a simpler system to understand the core drivers of biological intelligence.

Controlled-rearing studies on newborn animals have been particularly valuable for distinguishing between nativist and empiricist views (Gibson 1963, Held & Hein 1963, Walk et al. 1957). By rearing animals in strictly controlled environments, researchers can systematically manipulate an animal's experiences and measure what the animal learned from those experiences. Thus, researchers can determine which experiences are necessary and sufficient for learning perceptual and cognitive skills.

## 3. CONTROLLED-REARING STUDIES OF NEWBORN CHICKS

To reverse engineer the origins of intelligence, we need an animal model whose environment can be strictly controlled from birth. Primates, rats, and pigeons have been the dominant animal models in psychology and neuroscience; however, these animals must be raised in natural visual worlds (e.g., due to their need for a caregiver). This is problematic for distinguishing between nativist and empiricist theories because the natural world provides ample opportunities for learning domain-specific knowledge through experience. For instance, neuroscientists have shown that object recognition changes rapidly in response to statistical redundancies in the animal's environment (e.g., Cox et al. 2005, Wallis & Bülthoff 2001), with significant neuronal rewiring occurring in as little as one hour of experience with an altered visual world (Li & DiCarlo 2008, 2010). Developmental psychologists, in turn, have discovered that newborn brains can encode statistical relationships soon after birth (e.g., Bulf et al. 2011, Kirkham et al. 2002, Saffran et al. 1996). These findings allow for the possibility that even early developing domain-specific skills (e.g., skills emerging days, weeks, or months after birth) are learned from experience early in postnatal life.

To avoid these roadblocks, we chose an animal model—newborn chicks (*Gallus gallus*)—that is precocial and does not require a caregiver. It is possible to control all of a newborn chick's postnatal experiences from birth,[1] providing causal control over how experience shapes newborn minds (Wood & Wood 2015).

[1]We use the term "birth" colloquially to refer to the event in which an organism comes into the world (including both parturition and hatching).

Studies of chicks also shed light on human learning because avian and mammalian brains share many features. On the circuit level, both contain homologous cortical circuits for processing sensory input (Karten 2013). Although these circuits are organized differently in birds and mammals (nuclear and layered organization, respectively), the circuits share similarities in terms of cell morphology, gene expression, the connectivity pattern of the input and output neurons, and circuit function (Calabrese & Woolley 2015, Dugas-Ford et al. 2012, Jarvis et al. 2005, Wang et al. 2010). On the macro level, avian and mammalian brains share the same large-scale organizational principles: Both are modular, small-world networks with a connective core of hub nodes that includes visual, auditory, limbic, prefrontal, premotor, and hippocampal structures (Applegate et al. 2023, Shanahan et al. 2013). Given the circuit-level and architecture-level similarities between avian and mammalian brains, controlled-rearing studies of chicks can reveal general principles of learning across species.

### 3.1. Automating Controlled-Rearing Studies

Initially, we attempted to study newborn chicks with the same methods that were previously used to study human infants and newborn chicks in other labs: manual testing and direct observation. In these early (unpublished) studies, we reared chicks individually in home cages, then manually transported each chick to a test chamber, where we could present new objects and measure how long they spent with each object, using stopwatches as measuring devices. The experiments were laborious and time consuming, and the results were disappointing. When we moved the chicks from their home chamber to the test chamber, the move often overwhelmed the chicks, and they would freeze for long periods of time. This manual testing approach also limited the precision and amount of data that could be collected from each chick, while opening the possibility of experimenter bias in how the stimuli were presented and/or how the chick's behavior was measured. Overall, the data we collected using manual testing and direct observation were sparse and noisy, leading to datasets with low signal-to-noise ratios.

To make our experimental paradigm more comfortable for chicks and more efficient for researchers, we spent years developing new controlled-rearing methods. In other fields—including physics, chemistry, and astronomy—automation has been invaluable for improving data quality, eliminating experimenter bias, and removing sources of noise by standardizing data collection procedures. To apply these benefits of automation to the study of newborn minds, we invented a fully automated controlled-rearing method (Wood 2013). This method allowed us to raise newborn chicks in strictly controlled environments from the onset of vision. Specifically, after hatching, we move each chick to an automated controlled-rearing chamber (**Figure 2*a***). We move the chicks in darkness using night vision goggles to avoid exposing the chicks to any uncontrolled visual experiences (**Figure 2*b***). At that point, the chicks do not need to be moved again, since the chick's home chamber contains display walls (LCD monitors) for displaying preprogrammed training and test stimuli to the chicks. To track the chick's behavior, we embed microcameras in the ceilings of the chamber, which track the chick's behavior continuously. The camera data are passed to an automated image-based tracking system, which measures the chick's performance on the task. Since stimuli presentation, data collection, and behavioral tracking are performed by computers, this automated method eliminates experimenter bias and allows each chick's behavior to be tracked continuously and precisely across the first weeks of life (**Figure 2*c***). Automation also allows many subjects to be tested simultaneously. With this new tool, we can probe the initial state of object perception and measure how vision changes as a function of particular experiences, using high-precision methods (see the sidebar titled Using Automation to Study Newborn Minds).

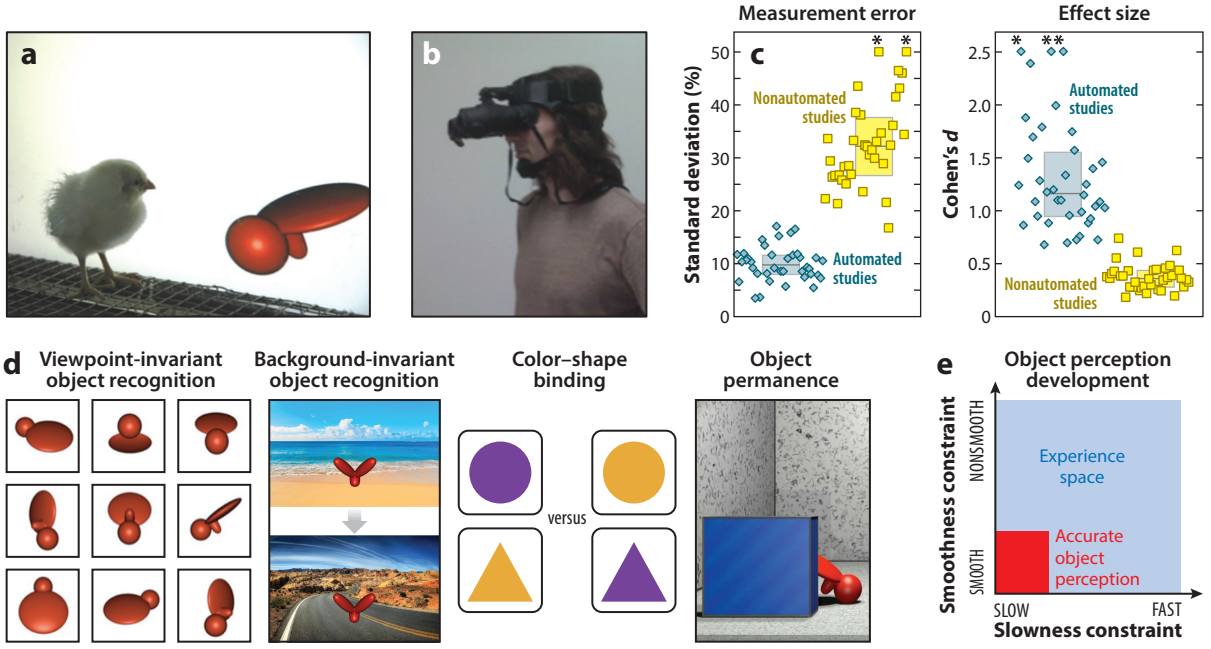



**Figure 2**

(*a*) A newborn chick in a controlled-rearing chamber. (*b*) To avoid exposing the chicks to uncontrolled visual experiences, we use night vision goggles for all animal husbandry. (*c*) Scatterplots and boxplots of the (*left*) measurement error and (*right*) effect sizes from samples of automated (*blue points*) and nonautomated (*yellow points*) controlled-rearing studies. Each point represents the standard deviation or Cohen's *d* from one condition. Across many studies, the effect sizes obtained with automated methods were much larger than the effect sizes obtained with nonautomated methods. Automated methods also produced much more precise measurements than nonautomated methods. Data taken from Wood & Wood (2019). (*d*) Automated controlled-rearing studies have revealed powerful one-shot learning abilities in newborn chicks. After encountering a single object, chicks can recognize the object across new viewpoints and backgrounds. Chicks can also bind color and shape features into integrated object representations and remember objects that have moved out of view (object permanence). (*e*) To develop these object perception skills, newborn chicks need slow and smooth visual experiences with objects (slowness and smoothness constraints). Together, these findings show that newborn visual systems are powerful but constrained.

## USING AUTOMATION TO STUDY NEWBORN MINDS

The accuracy of science depends on the precision of its methods. When fields produce precise measurements, the data can rigorously guide and constrain theory selection. When fields produce noisy measurements, however, the scientific method is not guaranteed to work. In fact, noisy data are regarded as a leading cause of the replication crisis across multiple fields (Loken & Gelman 2017, Munafò et al. 2017, Simmons et al. 2011). Due to the methodological challenges associated with testing newborn subjects, prior studies have tended to produce noisy data with a low signal-to-noise ratio. This has hindered attempts to characterize the brain's core learning algorithms. Automated controlled rearing helps solve this noisy data problem. For instance, in a large sample of nonautomated and automated controlled-rearing studies, automated studies produced measurements that were 3–4 times more precise than nonautomated studies and produced effect sizes that were 3–4 times larger than nonautomated studies (**Figure 2*c***). Automation also eliminates experimenter bias and allows replications to be performed quickly and easily. We argue that automation can be a powerful tool for improving measurement precision, producing high-powered studies, and generating accurate data for distinguishing between candidate models of the origins of intelligence.

### 3.2. Origins of Object Perception

Perhaps the most important benefit of automated controlled rearing is that newborn animals can be raised continuously in altered visual worlds. This is important because nativist and empiricist theories make the same predictions when animals are raised in natural worlds (because knowledge could be either innate or learned from experience early in postnatal life). After all, natural visual environments provide ample evidence about the behavior of objects, so even young animals might have access to sufficient training data to learn object perception from domain-general algorithms. To distinguish between theories, we need a different approach. Rather than raising animals solely in natural worlds (where nativist and empiricist theories make the same prediction), we must also raise animals in impoverished and unnatural worlds (where nativist and empiricist theories make different predictions). By systematically manipulating the object experiences available to newborn animals, we can measure whether those experiences systematically alter the animals' visual knowledge.

To explore the origins of object perception, we have raised newborn chicks in impoverished and unnatural virtual worlds. These studies provide clear evidence that chicks have an object-based inductive bias at the onset of postnatal visual learning (Wood et al. 2024). This inductive bias predisposes chicks to transform retinal inputs into object-centric scenes, containing bounded objects that persist over space and time. Specifically, two signatures characterize an inductive bias: (*a*) It allows systems to make inferences that go beyond the training data, and (*b*) it constrains the range of input–output functions that can be learned (Lake et al. 2017, Mitchell 1980, Wolpert & Macready 1997). Our experiments provide evidence for both signatures in newborn chicks.

First, newborn vision is powerful: Newborn chicks can generalize far beyond their prior experiences with objects (**Figure 2*d***). Soon after hatching, chicks can segment objects from backgrounds (Wood & Wood 2021a), recognize objects across novel views (Wood & Wood 2017, 2020), bind color and shape features into integrated object representations (Wood 2014), and recognize objects based on motion patterns (Goldman & Wood 2015). Likewise, chicks can rapidly learn to recognize faces (Wood & Wood 2015), including human faces presented from novel views (Wood & Wood 2021b). Chicks also have a robust sense of object permanence, allowing them to remember objects that have moved out of view (Prasad et al. 2019, Wood et al. 2024). All of these skills are present soon after hatching, even when chicks have been raised in impoverished environments containing a single object. Newborn chicks can thus solve challenging object perception tasks in the absence of extensive experience with objects.

Second, newborn vision is constrained (**Figure 2*e***). To characterize the constraints on newborn vision, we raised chicks in unnatural worlds that altered the core visual experiences that animals acquire in natural worlds. We focused on manipulating the slowness and smoothness of object motion. Researchers have long observed that natural visual environments are slow and smooth: Objects are typically present for seconds or longer, and when objects do move, they move smoothly across the retina (e.g., DiCarlo et al. 2012, Feldman & Tremoulet 2006, Földiák 1991, Gibson 1979, Stone 1996, Wallis & Rolls 1997, Wiskott & Sejnowski 2002). In consequence, visual features that co-occur across short periods of time are, on average, more likely to correspond to different images of the same object than to different objects. In principle, newborn visual systems might learn how to see by encoding these slow and smooth signals from their visual environment.

By leveraging automated controlled rearing, we confirmed this prediction. We found that the development of object perception requires visual experiences of object views changing slowly and smoothly, adhering to the spatiotemporal properties of objects in the real world. Without slow and smooth visual object experiences, chicks largely failed to develop object perception. The development of object parsing (Wood & Wood 2021a), visual binding (Wood 2016), view-invariant

object recognition (Wood & Wood 2016, 2018; Wood et al. 2016), face recognition (Wood & Wood 2021b), and object permanence (Prasad et al. 2019) all required visual experience of objects moving slowly and smoothly. If an object moved too quickly when being encoded into memory, the resulting object representation was distorted in the direction of object motion, effectively breaking object recognition (Wood & Wood 2016, 2021b). Similarly, if an object moved nonsmoothly when being encoded into memory, chicks failed to solve simple color and shape recognition tasks (Wood 2016), and their object representations failed to generalize across new views and rotation speeds (Wood & Wood 2016, 2018). Experience with a natural (slow and smooth) environment is necessary for the development of object perception.

Taken together, these findings support both nativist and empiricist claims. Nativists can emphasize that newborn animals have many object perception skills during their first encounters with objects. Empiricists can emphasize that experience heavily shapes object perception, with animals needing specific kinds of experiences (slow and smooth) to develop object perception. To further complicate the debate, one might question whether natural visual experience is required to learn object perception or to maintain object perception skills that are already present at birth (Spelke & Newport 1998). How do we resolve this debate? We suggest that a satisfying scientific explanation for these empirical phenomena will require a reverse engineering perspective, in which we try to build artificial systems (computational models) that learn like newborn animals.

## 4. REVERSE ENGINEERING BIOLOGICAL INTELLIGENCE

Why are the origins of intelligence still unknown? We suspect that, in addition to the methodological challenges discussed above, the field also faces conceptual challenges in understanding the origins of intelligence.

Scientists studying the origins of intelligence have generally relied on verbal theories or simple (low-dimensional) quantitative models. Relying on verbal theories is problematic because verbal theories are underspecified. Any verbal theory can be formalized (computationally) in infinite ways. Given their low-dimensional nature, verbal theories can also make predictions that are vague and difficult to falsify. Similarly, relying on low-dimensional quantitative models is problematic because learning is complex, and its underlying mechanisms likely cannot be understood through simple, low-dimensional models. The brain is a high-dimensional learning system (e.g., with 100 trillion adjustable synapses in the human brain; Azevedo et al. 2009), and during learning, the brain changes as a function of high-dimensional sensory data (e.g., from $10^6$ optic nerve fibers) arising across nested periods of development (Adolph & Hoch 2019). This is a massive amount of complexity to capture in just a few dimensions. Low-dimensional models also do not perform the same tasks as animals (i.e., learning from raw sensory data). Thus, the models cannot be accurate working models of the core learning algorithms in newborn brains. Given the limitations of verbal theories and low-dimensional models, perhaps it is not surprising that scientists have been unable to develop rigorous models of the core learning algorithms in brains. The field simply did not have the right conceptual tools for simulating high-dimensional learning systems interacting with high-dimensional sensory data across time.

Other fields have solved the complexity problem by building closed-loop scientific systems. The idea of closed-loop systems comes from engineering, where the outputs of the system are fed back into the system, as a form of feedback, to allow parameters to adjust in response to environmental changes. In closed-loop scientific systems, experimental data are compared with predictions made by theoretical models. The models, in turn, generate new predictions for refining, validating, and falsifying existing models. In physics, for example, tools like the Large Hadron Collider automate high-energy particle collision experiments, and the resulting data are compared

with predictions made by complex theoretical models, all without human intervention (ATLAS Collab. et al. 2012). Likewise, in chemistry, computer simulations have allowed scientists to study and predict complex chemical reactions. These simulations are guided by autonomous closed-loop systems, where observed outputs are continuously compared with theoretical predictions (Volk et al. 2023). These examples illustrate how humans have overcome the limits of human intuition to build high-dimensional models of complex systems.

This perspective echoes Newell (1973), who famously argued that the most effective way to make theoretical progress in the mind sciences is not to "play 20 questions with nature" by focusing on binary verbal reasoning and simple low-dimensional models (see also Kanwisher et al. 2023, Kriegeskorte & Douglas 2018). Instead, scientists should attempt to build unified task-performing computational systems. These systems (models) should be capable of performing the tasks they aim to explain (e.g., converting high-dimensional sensory signals into actions) and be testable across a wide variety of tasks.

### 4.1. Reverse Engineering Sensory Systems

Computational neuroscience has embraced this approach by attempting to reverse engineer sensory systems. The reverse-engineering paradigm involves comparing brains to artificial neural networks (ANNs) that are trained to perform real-world tasks (e.g., object segmentation, face recognition, navigation). Since the ANNs perform the same tasks as biological systems by using algorithms that are based on brain-like units (e.g., neurons), the ANNs can serve as testable hypotheses about the algorithms underlying biological intelligence (Schrimpf et al. 2020). Brains and ANNs can then be integrated into closed-loop scientific systems, allowing neuroscientists to find ANN models that can accurately predict neural and behavioral patterns produced by biological systems. This paradigm has been successfully applied to vision (Yamins et al. 2014), audition (Kell et al. 2018), olfaction (Wang et al. 2021), visually guided action (Michaels et al. 2020), language (Schrimpf et al. 2021), navigation (Whittington et al. 2022), decision making (Binz & Shulz 2023), and memory (Nayebi et al. 2021).

This closed-loop approach is effective because it allows researchers to discover high-performing models. The ANNs perform the same tasks as animals, so ANNs and animals can be directly compared in a closed loop, allowing effective search through large classes of high-dimensional models. By testing a wide variety of ANN models, researchers can determine which model features are most important for improving model accuracy (and which are not). Features can be progressively added and refined in new models in a continuous feedback loop between model engineering and model testing (Schrimpf et al. 2020).

Since the closed loop links biological and artificial systems, this approach produces models at an engineering level of abstraction: a level close enough to biology to preserve the essential details needed to mimic biological intelligence but abstract enough to discard inessential details (Doerig et al. 2023). This helps researchers focus on the model features that matter for producing brain-like intelligence. For instance, when reverse engineering the ventral visual stream, researchers found that feedforward networks can account for a large proportion of the explainable variance in neural activation in response to particular images (Yamins et al. 2014). While adding additional features to the model (e.g., recurrency, memory, lateral connections, spiking) might improve model performance, it is useful to know that a relatively simple model (i.e., consisting solely of feedforward connections) is sufficient to reproduce much of the neural and behavioral signatures of mature visual systems.

ANN models are also valuable because they can serve as integrative models of intelligence. For example, when researchers find a compelling image-computable model of the ventral visual

system, that model can then be evaluated across a wide range of tasks (e.g., all tasks that take pictures and/or videos as input). This allows researchers to evaluate a single model across many tasks, thereby building up integrative (unified) models of the target domain (e.g., object recognition). There is optimism that the reverse-engineering paradigm (i.e., comparing biological and artificial systems in closed-loop systems) will allow scientists to build unified models of biological intelligence (Doerig et al. 2023, Lindsay 2021, Richards et al. 2019).

## 4.2. The Digital Twin Approach

While reverse engineering has led to success with mature animals, this paradigm has not yet been applied to newborn animals. Ultimately, reverse engineering newborn cognition will be essential for building unified models of intelligence because all biological skills (e.g., object perception, navigation, numerical cognition, social cognition, motor control) are products of the core learning algorithms in newborn brains. To this end, the goal of our research program is to reverse engineer the core learning algorithms in brains. To do so, we developed digital twin studies, which involve performing parallel controlled-rearing experiments on newborn animals and artificial agents (**Figure 1**). We raise newborn animals and artificial agents in the same environments and test them with the same tasks, allowing for a direct comparison of their learning abilities.

Digital twin studies explicitly link newborn animals and artificial agents in a closed-loop scientific system. Newborn animals provide data for guiding the development of artificial agents that learn like animals. Artificial agents, in turn, serve as task-performing models for studying the origins and computational basis of intelligence. Digital twin studies thus allow us to ask questions that cannot be addressed with verbal theories or low-dimensional quantitative models, such as:

- What core learning algorithms are necessary and sufficient to develop psychological skills?
- What experiences are necessary and sufficient to develop psychological skills?
- How do core learning algorithms and experience interact to produce psychological skills?
- Why do newborn animals develop the psychological skills that they do?

Digital twin studies involve raising artificial agents (embodied ANNs) in virtual simulations of the environments faced by newborn animals (**Figure 1**). By raising biological and artificial agents in the same environments, we can give them the same training data and test them with the same tasks. The ANN experiments are performed in a video game engine (Unity 3D), which provides photorealistic accuracy while circumventing the limitations of time and cost associated with physical hardware.

Virtual studies offer many other advantages. First, unlike physical robots, ANN models can be easily shared between scientists. This allows scientists to rapidly test competing models, fostering a culture of open verification of scientific findings. Second, not only the ANN models, but also entire experiments, can be exchanged between researchers. These virtual experiments—formatted as games in video game engines—can be disseminated widely and used by other research groups, enabling rapid replication. Third, the virtual nature of both the models and experiments makes it possible to build large-scale scientific testbeds. These testbeds will allow a single model to be evaluated across a range of psychological tasks, enriching our understanding of models and producing unified models of biological intelligence.

## 4.3. Machine Learning Algorithms and the Nativist–Empiricist Debate

Digital twin studies involve building artificial agents that have machine learning (ML) algorithms. This approach therefore requires ML algorithms that resemble the learning systems described in

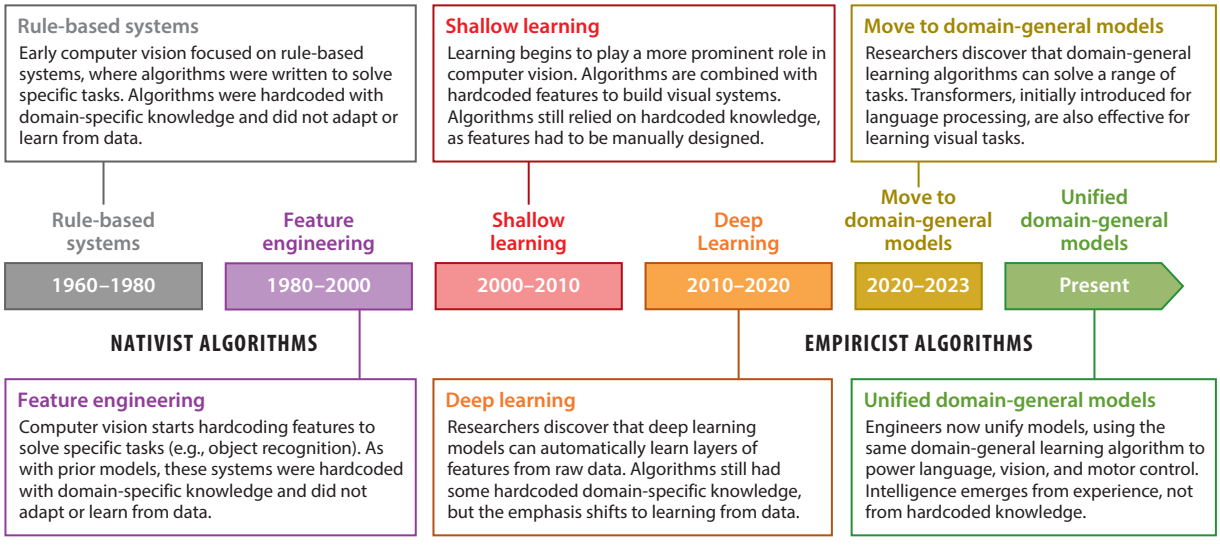



**Figure 3**

In computer vision, there has been a progression from hardcoded, domain-specific algorithms to flexible, domain-general algorithms. This trend reflects the field's evolving understanding of how systems can learn effectively from data. In the 1960s–1980s, early computer vision focused on rule-based systems where algorithms were written to solve specific tasks. For example, edge detection algorithms identified edges in images by looking at areas where there was a sharp change in intensity or color. These algorithms were hardcoded with domain-specific knowledge and did not adapt or learn from data. In the 1980s–2000s, the next stage of computer vision focused on hand-engineering features to solve specific tasks (e.g., object and face recognition). This led to the development of hardcoded feature descriptors (e.g., Scale-Invariant Feature Transform, Histogram of Oriented Gradients). As with prior models, these systems were hardcoded with domain-specific knowledge and did not adapt or learn from data. In the 2000s–2010s, machine learning began to play a more prominent role in computer vision. Algorithms like Support Vector Machines were combined with hardcoded features to build object detection systems. These systems still largely relied on domain-specific knowledge, as features had to be manually designed. In the 2010s–2020s, deep learning models, unlike previous approaches, could automatically learn hierarchical features from raw data. Convolutional neural network architectures (e.g., AlexNet, VGGNet) did have some hardcoded domain-specific features (convolutional layers), but these models learned the filters and features directly from the data rather than relying on hand-engineered features. From 2020 to 2023, as deep learning matured, researchers explored architectures that were less reliant on hardcoded domain-specific inductive biases. Transformers, initially introduced for language processing tasks, were found to be highly effective for vision tasks. This showed that domain-general architectures could be effective in solving complex visual tasks. At present, researchers are now unifying models, where the same architecture is used for multiple domains, such as language, vision, and motor control. This signifies a shift toward domain-general learning algorithms that are agnostic to data type. This mirrors the empiricist view in the nativist–empiricist debate, as learning is guided more by data and less by hardcoded, domain-specific knowledge.

theories of the origins of intelligence. To what extent does ML rely on nativist (domain-specific) versus empiricist (domain-general) learning algorithms?

In computer vision, there has been a gradual progression from hardcoded domain-specific algorithms to domain-general algorithms that learn from data (**Figure 3**). During the early days of artificial intelligence (1960s–1980s), researchers built rule-based systems that were designed to solve specific tasks (e.g., edge detection) (Canny 1986, Marr & Hildreth 1980, Sobel & Feldman 1968). The algorithms were hardcoded with specific heuristics and did not learn from data. As computer vision progressed (1980s–2000s), researchers realized that some features were better than others for tasks like object detection and face recognition, which led to the development of feature descriptors, such as Scale-Invariant Feature Transform (Lowe 1999) and Histogram of Oriented Gradients (Dalal & Triggs 2005). These feature descriptors, while still human engineered, provided a way to extract complex representations from images. During the

early 2000s, learning began to play a more prominent role in computer vision. Algorithms like Support Vector Machines (Boser et al. 1992) were combined with handcrafted features to build object detection systems. However, these systems still largely relied on hardcoded domain-specific knowledge.

With the rise of deep learning, particularly convolutional neural networks (CNNs), the field experienced a shift toward empiricist-like learning principles. Unlike prior approaches, CNNs could automatically learn features from raw sensory data. CNNs still had some hardcoded knowledge in the form of convolutional layers, which are effective at processing grid-like data (e.g., images). However, CNNs learned the features directly from the data, rather than relying on hand-engineered features. As deep learning matured, researchers became even less reliant on hardcoded knowledge. For example, convolutional layers are now unnecessary for computer vision (Chen et al. 2020, Dosovitskiy et al. 2020). Vision transformers (ViTs) do not have a CNN's hardcoded bias toward local spatial structure, but instead are based entirely on the flexible (learned) allocation of attention. Nevertheless, transformer architectures are still effective on vision tasks. Transformer architectures are also effective across a variety of domains, including natural language processing, their initial purpose (Vaswani et al. 2017). Transformers show that domain-general algorithms can learn to solve many real-world tasks.

Most recently, researchers have started using the same architectures and learning objectives across domains, including language, vision, speech, navigation, and decision making. Using transformer architectures, which can serve as computational building blocks for both language and vision, researchers have developed a unified self-supervised learning technique, called masked autoencoding (MAE) (He et al. 2022). MAE involves masking random patches from the input image and reconstructing the missing patches in the pixel space. This learning objective encourages ANNs to learn the underlying data distributions producing the images, so the networks can predict features in the masked patches. The MAE learning objective is simple and domain general, and it is highly effective for both language learning (Devlin et al. 2019) and visual learning (Feichtenhofer et al. 2022, Tong et al. 2022). Hardcoding less domain-specific knowledge appears to give ANNs more flexibility and power.

Using digital twin studies, we can directly test where newborn visual systems fall on this spectrum from nativist to empiricist algorithms. Do artificial agents need innate (hardcoded) knowledge to learn like newborn animals? Or are domain-general algorithms sufficient to learn animal-like object perception?

### 4.4. Digital Twin Studies with Convolutional Neural Networks

CNNs are ideal starting points for testing whether domain-general algorithms can learn the same object perception skills as newborn animals. As discussed above, CNNs learn visual features from data. However, CNNs still have some innate knowledge built into their architecture because (*a*) they have a strong spatial inductive bias encoded in their retinotopic architecture and (*b*) they are typically considered to be specialized models for sensory processing (e.g., vision, audition, olfaction).

Research on CNNs has demonstrated that domain-specific visual knowledge can emerge from these domain-general algorithms when they are optimized for specific tasks. For example, behavioral signatures of face recognition emerge when CNNs are trained on face recognition (Dobs et al. 2023). Domain-specific learning has also been observed for letter perception (Janini et al. 2022), shape perception (Ritter et al. 2017), and scene recognition (Zhou et al. 2014). CNNs can also learn to cluster images according to distal properties such as reflectance and illumination, despite receiving no explicit information about these properties (Storrs et al. 2021). Another example of domain-specific learning in CNNs comes from visual illusions. CNNs trained for

low-level visual tasks show a human-like propensity to fall prey to visual illusions (Gomez-Villa et al. 2019), indicating that visual illusions may be a form of learned domain-specific knowledge.

To directly test whether CNNs learn like newborn animals, we performed digital twin studies (Lee et al. 2021; L. Pandey, D. Lee, S. Wood and J. Wood, manuscript under review). We first raised newborn chicks in strictly controlled visual environments and measured the chicks' view-invariant object recognition performance (Wood 2013). We then simulated the training data available to the chicks by creating virtual replicas of the controlled-rearing chambers in a video game engine and recording the first-person images acquired by agents moving through the chambers. Finally, we trained self-supervised CNNs with the simulated first-person images from the virtual animal chambers and tested the CNNs with the same images used to test the chicks. This approach allowed us to train newborn chicks and CNNs in the same environment and test them with the same test stimuli, enabling direct comparison of their learning abilities.

We found that self-supervised CNNs spontaneously learn view-invariant object features when trained on the first-person visual experiences of newborn chicks. For both chicks and CNNs, impoverished environments (e.g., containing a single object) provide sufficient visual experience for learning view-invariant features. We also found that CNNs produce well-structured representations, containing information about both object identity and other latent variables of interest (e.g., object distance, viewing position). When CNNs receive the same visual experiences as chicks, we observe parallel development of view-invariant object recognition in CNNs and chicks.

To what extent did performance depend on hardcoded versus learned features of the model? One possibility is that CNNs learned view-invariant features in impoverished environments because CNNs have a strong hardcoded inductive bias. The convolutional operation reflects the spatial structure of natural images, including local connectivity, parameter sharing, and hierarchical structure (LeCun et al. 2015). This spatial bias allows CNNs to generalize well from small datasets and learn useful feature hierarchies that capture the structure of visual images (Cao & Wu 2022, Liu & Deng 2015). This innate spatial knowledge might also explain how CNNs learn like newborn chicks (i.e., both animals and machines might have a strong inductive bias supporting spatial learning).

### 4.5. Digital Twin Studies with Vision Transformers

Is hardcoded spatial knowledge necessary for algorithms to learn like newborn chicks? To test whether algorithms that are more domain general can learn like newborn visual systems, we turned to ViTs. Unlike CNNs, ViTs lack convolutional processing (hardcoded knowledge about spatial relationships) and hierarchical feature extraction (hardcoded knowledge about local relationships between features). Instead, ViTs process an image by dividing it into patches (with positional encodings) and applying a domain-general attention mechanism (the same employed in language transformers) to encode spatial relationships.

This minimalistic approach to hardcoded knowledge generally improves performance, since ViTs often outperform CNNs and other models with stronger hardcoded inductive biases. For instance, ViTs demonstrate state-of-the-art performance on visual tasks, including object segmentation and recognition (Dosovitskiy et al. 2020, Zhou et al. 2021), face recognition (Zhou et al. 2021), and action recognition (Ulhaq et al. 2022, Yang et al. 2022), while also generating high-quality readouts for estimation of optical flow, occlusions, object segments, and relative depth (Bear et al. 2023). ViTs thus provide a powerful existence proof that domain-general algorithms can be a strong foundation for vision.

To compare learning across ViTs and newborn chicks, we used the digital twin approach described above, where we trained and tested ViTs with the simulated first-person visual experiences

from the virtual animal chambers (Pandey et al. 2023). We built ViTs that—like chicks—learned by leveraging the temporal structure of natural visual experience, without relying on labeled data. This temporal ViT algorithm, which we call Vision Transformers with Contrastive Learning through Time (ViT-CoT), learns representations that maximize similarity between temporally adjacent images and minimize similarity between nonadjacent images. These representations reflect the underlying dynamics, context, and patterns across time. We found that ViT-CoT learns view-invariant object features when trained on the visual experiences of newborn chicks. This ViT architecture also learned well-structured visual representations containing information about both object identity and viewing position. For both chicks and ViTs, impoverished environments (with a single object) contained sufficient visual experience for learning view-invariant object features.

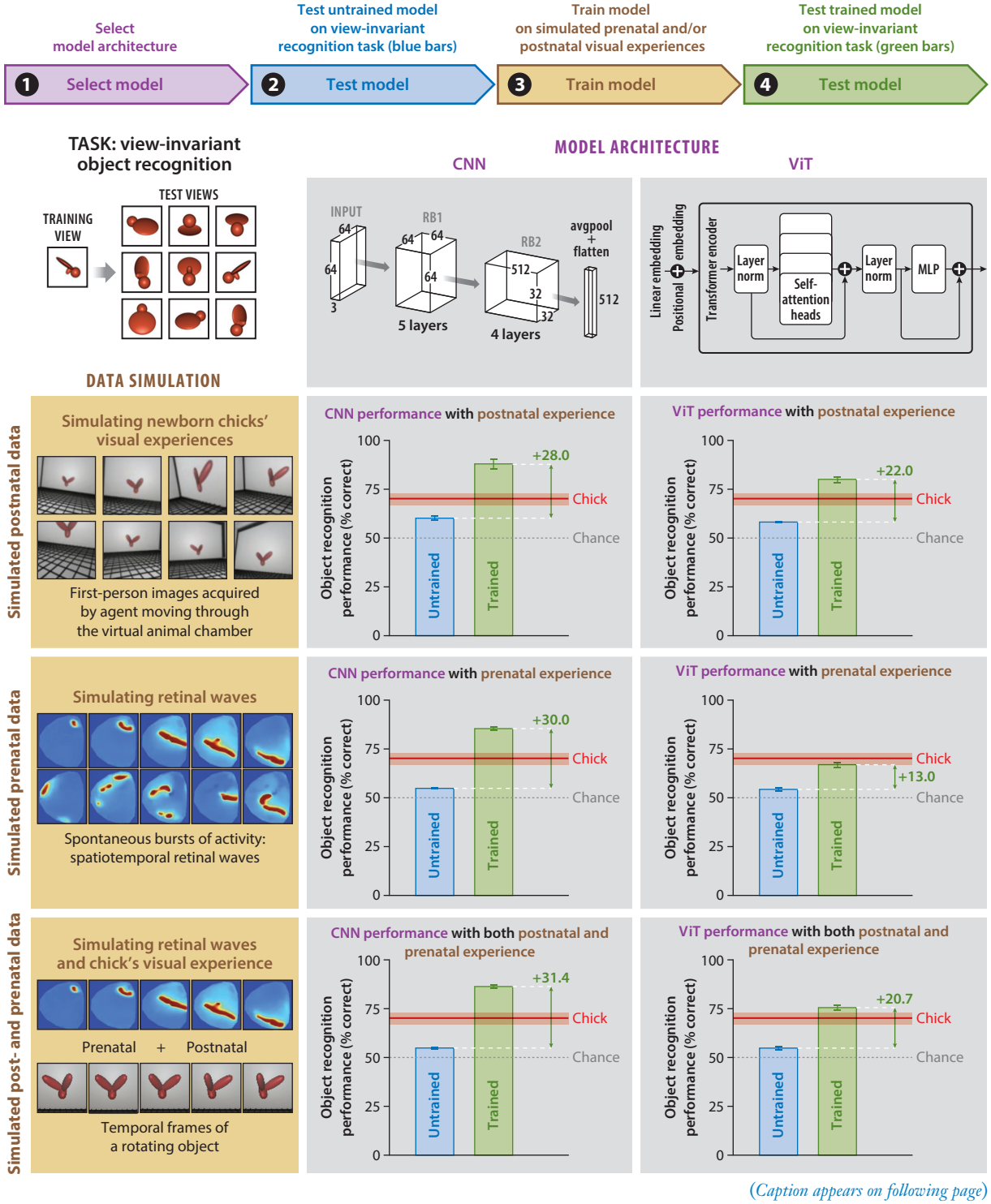
We conclude that neither CNNs nor ViTs are more data hungry than newborn chicks. This finding reinforces the possibility that CNNs and ViTs can be used as image-computable models of visual learning and development. More generally, these studies show that domain-general algorithms, combined with the embodied data streams available to newborn animals, are sufficient to drive the development of animal-like object recognition.

### 4.6. Simulating Prenatal Learning in Machines

The above results support the central claim of empiricism: The core learning algorithms supporting biological vision can be domain general in nature. However, these results do not necessarily oppose the central claim of nativism: that domain-specific knowledge is present and functional at birth. Although researchers often consider birth (or hatching) to be the starting point of learning, prenatal learning plays a critical role in the development of vision. In principle, domain-general algorithms might produce domain-specific knowledge if those algorithms are trained on prenatal experiences.

Researchers have been particularly interested in retinal waves as a potential source of innate (prenatal) visual development. During prenatal development, neurons in the retina generate spontaneous, synchronized clusters of activity among neighboring groups of cells (Arroyo et al. 2016, Blankenship & Feller 2010, Wang & Bergles 2015, Wenner 2012). These clusters of activation are called waves because they propagate smoothly over space and time and contain spatiotemporal statistics similar to those found in the natural visual world (Ge et al. 2021). Thus, these object-like patterns could predispose newborns to perceive the world in terms of enduring objects that persist over space and time. In support of this view, Albert et al. (2008) showed that efficient learning systems trained on simulated retinal waves develop neurons with a localized, oriented, bandpass structure, similar to neurons in the primary visual cortex. We extended this result by asking whether retinal waves are sufficient to learn view-invariant object features (L. Pandey, S.M.W. Wood and J.N. Wood, unpublished results).

To test whether domain-general algorithms can learn view-invariant features when trained solely on retinal waves (prenatal sensory data), we performed digital twin experiments training CNNs and ViTs on retinal waves (**Figure 4**). When these domain-general algorithms were trained solely on simulated retinal waves, the algorithms developed high-level object features, allowing the networks to solve the same view-invariant object recognition task as newborn chicks. Thus, when domain-general algorithms learn from prenatal experiences, the networks develop domain-specific knowledge. These results simultaneously provide evidence for a core empiricist claim—domain-specific knowledge can be learned from domain-general learning systems—and a core nativist claim—domain-specific knowledge can develop from prenatal experiences.




*(Caption appears on following page)*

**Figure 4** (*Figure appears on preceding page*)

To evaluate whether ML algorithms learn like newborn chicks, we ① select a ML model (e.g., a CNN or ViT), ② test the untrained model's performance (*blue bars*), ③ train the model on the visual experiences of newborn chicks, and ④ test the trained model's performance (*green bars*). Comparing the untrained and trained models reveals whether the models could learn from the same visual experiences as chicks. The algorithms were tested on the view-invariant recognition task from Wood (2013), in which chicks were reared in environments containing a single object seen from a single viewpoint range, then tested on their ability to recognize that object across novel views. Each row shows the performance of the algorithm when trained on different types of simulated data. Both algorithms had the same learning objective (contrastive learning through time), which leverages time as a teaching signal to learn representations, akin to biological visual systems. Both CNNs and ViTs learned animal-like object recognition when trained on prenatal experiences and when trained on the first-person views acquired by newborn chicks in controlled-rearing chambers. Abbreviations: CNN, convolutional neural network; ML, machine learning; MLP, multilayer perceptron; ViT, vision transformer.

## 5. THE ORIGINS OF INTELLIGENCE AS SPACE-TIME FITTING

What core learning algorithms underlie visual intelligence? Above, we argue that distinguishing between nativist and empiricist theories requires a closed-loop scientific system (digital twin studies), in which newborn animals and artificial agents are raised in the same environments and tested with the same tasks. Through parallel controlled-rearing studies of newborn animals and artificial agents, we can test which learning algorithms—and which experiences—are necessary and sufficient to develop visual intelligence. To date, we have discovered that sparse visual experiences with objects are sufficient for newborn animals to develop object perception, provided that the objects move slowly and smoothly. Without slow and smooth experiences, newborn visual systems develop distorted object perception. Likewise, we discovered that domain-general algorithms (e.g., CNNs and ViTs) are sufficient for learning object perception when the algorithms are trained on the first-person visual experiences of newborn animals. CNNs and ViTs also show evidence for a slowness constraint when trained on the visual experiences of human infants; like newborn visual systems, these algorithms learn better representations when they are trained in more slowly changing visual environments (Sheybani et al. 2023). Therefore, we posit that visual learning across humans, animals, and machines can be parsimoniously understood in terms of a domain-general principle, called space-time fitting, in which visual systems spontaneously adapt (fit) to the spatiotemporal data distributions of the visual environment (**Figure 5**).

We use space-time fitting to refer to a class of direct-fit learning models that become adapted to their training data through brute-force fitting processes (for a detailed discussion of direct-fit models, see Hasson et al. 2020). This term is inspired by Gibson's (1979) use of the term direct perception. Similar to evolutionary processes (a brute-force fitting process by which organisms become adapted to their environment), direct-fit models use a brute-force fitting process to learn how to perceive and act on the world. By optimizing millions of parameters (connection weights) across millions of samples (experiences), direct-fit models learn to solve real-world tasks by fitting their internal parameters to the data distributions in the environment.

Unlike classic models in developmental psychology, direct-fit models do not learn simple, human-interpretable rules or representations of the world. Instead, direct-fit models build complex, high-dimensional representations by iteratively adjusting large numbers of parameters to fit (adapt) to the structure of the environment. These representations approximate the distal variables (e.g., objects, scenes) that produce proximal retinal images. We emphasize that, although the learning algorithms that implement direct fitting are complex (e.g., millions of adjustable parameters), direct-fit models are conceptually simple and parsimonious. In fact, the processes driving direct fit mirror the processes driving natural selection, in which organisms become adapted to their environment through iterative selection (Hasson et al. 2020). Both evolution and development can be conceptualized as blind, brute-force fitting processes in which organisms gradually adjust to the environment.

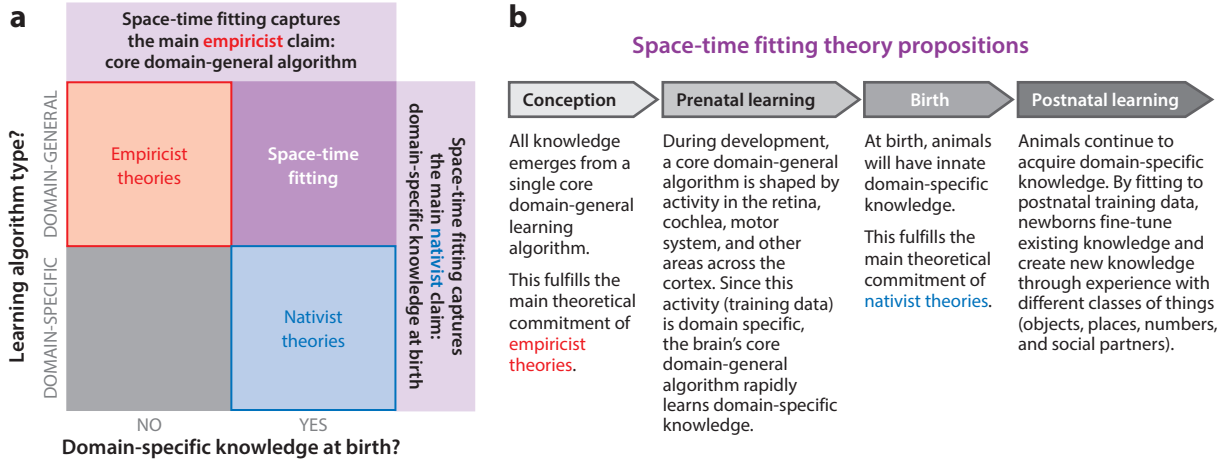



**Figure 5**

Space-time fitting theory. (*a*) Space-time fitting explains the core empiricist claim (i.e., that domain-specific knowledge emerges from domain-general learning mechanisms) and the core nativist claim (i.e., that domain-specific knowledge exists at birth). (*b*) A timeline showing how a single core domain-general learning algorithm develops innate domain-specific knowledge.

Space-time fitting models are the subset of direct-fit algorithms that perform unsupervised learning from spatiotemporal data. Many direct-fit algorithms for vision learn from supervision (e.g., vanilla CNNs, vanilla ViTs) and/or from spatial statistics (e.g., SimCLR, MAEs; Chen et al. 2020, He et al. 2022). However, a limited number of algorithms—the space-time fitters—learn from spatiotemporal data without supervision [e.g., SimCLR-CLTT (Schneider et al. 2021), ViT-CoT (Pandey et al. 2023), VideoMAEs (Tong et al. 2022)]. We hypothesize that this subset of direct-fit models learns the most like newborn visual systems. We speculate that space-time fitters will both (*a*) show the same generalization abilities and (*b*) show the same learning constraints (e.g., slowness and smoothness constraints) as newborn animals.

Space-time fitting is a new term intended to integrate learning in brains and machines under a common principle (**Figure 6*a***). For brains, Hebbian learning, spike timing–dependent plasticity, reinforcement learning, and predictive coding are all ways to fit brains to spatiotemporal data distributions via iterative, brute-force learning. Likewise, for machines, backpropagation, generative modeling, and deep reinforcement learning are popular artificial intelligence approaches that involve iterative, brute-force learning of underlying data distributions. While these concepts are typically treated separately, we argue that they reflect the same general principle of high-dimensional systems iteratively adapting (fitting) to spatiotemporal data distributions. As such, space-time fitting provides a unified conceptual framework for understanding learning in brains and machines.

Space-time fitting models are flexible (**Figure 6*b***), spontaneously learning representations by adapting to the spatiotemporal statistics of the environment. Accordingly, space-time fitting models of visual learning make a key prediction: Since representations are learned by fitting to the data distributions in the environment, the representations in the brain will ultimately mirror the data distributions in the environment (after prolonged periods of development). Thus, manipulating the spatiotemporal data distributions in an animal's environment should systematically alter the newborn's visual representations.

We have confirmed this prediction across a range of controlled-rearing studies. As reviewed in Section 3.2, we found that it is possible to systematically alter a newborn chick's object perception behavior simply by varying the speed at which objects move (Wood & Wood 2016, 2018). When

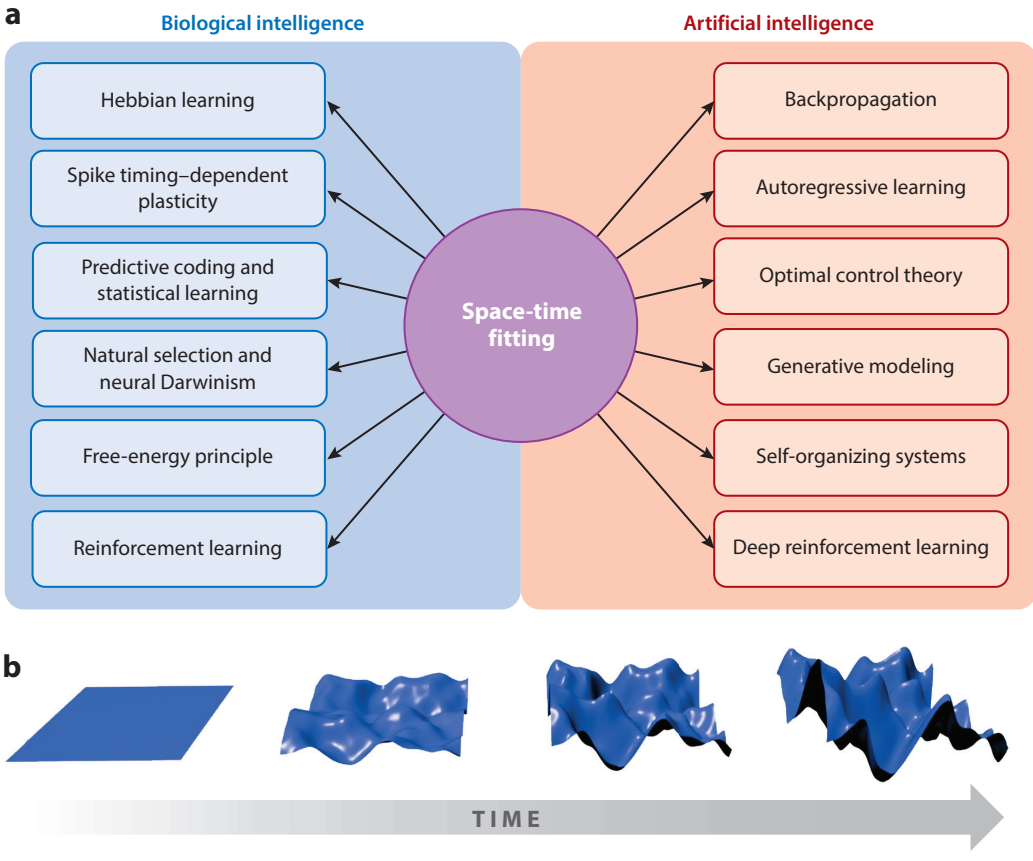



**Figure 6**

(*a*) Space-time fitting links concepts from learning in brains (*blue box*) and machines (*pink box*) under a common principle. These different concepts reflect the same general principle of high-dimensional systems iteratively adapting (fitting) to the spatiotemporal data distributions in the visual environment. This brute-force learning approach can be implemented in different ways across biological and artificial learning systems. (*b*) A visualization of space-time fitting in a simplified three-dimensional space. Space-time fitting is a flexible learning process: During prenatal and postnatal development, the representational landscape (*blue sheet*) in brains and machines gradually adapts to the spatiotemporal data distributions in the environment.

reared with slowly rotating objects, chicks build abstract object representations that are selective for object identity and tolerant to identity-preserving image changes; conversely, when reared with quickly rotating objects, chicks build view-dependent representations that are selective for familiar motion features. We found the same pattern for face recognition (Wood & Wood 2021a). When animals are raised in environments with altered space-time data distributions, the animals learn altered forms of object perception, confirming a key prediction of space-time fitting models of visual learning.

## 5.1. Implications for the Nativist Versus Empiricist Debate

Digital twin studies provide an engineering-level framework for developmental psychology, in which models learn about the world in their own right. By attempting to build systems that learn like animals, we can systematically explore which core algorithms and experiences matter for reproducing biological intelligence in machines. This engineering-level framework opens new

possibilities for using high-performing ANN models as runnable, computationally precise, and neurally mechanistic models of the origins of intelligence. Ultimately, we suspect that digital twin studies will support both empiricist and nativist claims (**Figure 5*a***). Because the verbal models that have dominated the nativist–empiricist debate are low-dimensional models intended to explain high-dimensional processes, both nativist and empiricist observations can be explained through a unified high-dimensional model.

**5.1.1. Space-time fitting explains empiricist phenomena.** Empiricists have long posited powerful domain-general learning mechanisms capable of acquiring domain-specific knowledge from experience, consistent with a generic space-time fitting process. Space-time fitting models thus provide existence proofs that empiricist principles can underlie visual intelligence. There is also evidence that direct-fit models (such as space-time fitting) can support intelligence in other domains, including audition, language, navigation, decision making, and motor control (Chen et al. 2021, Li et al. 2022, Radosavovic et al. 2023, Schrimpf et al. 2021). Space-time fitting can even be used to build single unified systems that solve a range of tasks. For example, Gato—a transformer agent—learned how to perform hundreds of real-world tasks, developing a form of general intelligence by leveraging direct-fit learning principles. The existence of systems like Gato shows that space-time fitters can be viable models of embodied intelligence. These models suggest that a large fraction of biological intelligence is the direct consequence of brains fitting to the multimodal data streams acquired by animals. Space-time fitters thus fulfill a core promise of empiricist thinking: the discovery of a domain-general system that can learn to solve many tasks.

Digital twin studies allow researchers to rigorously evaluate domain-general models of the core learning algorithms in brains. Researchers can directly test what is learnable—and what is not—from a visual environment, thereby grounding empiricist theories in a closed-loop scientific system. The space of empiricist theories is vast, so researchers need strategies to effectively search through the model space. By directly comparing learning across animals and machines, we can efficiently test which algorithms (models) learn like newborn animals and falsify incorrect algorithms.

**5.1.2. Space-time fitting explains nativist phenomena.** Space-time fitting also accords with nativist theories. According to most nativists, knowledge present at birth qualifies as innate knowledge. However, innate knowledge could either be hardwired through evolutionarily predetermined neural circuitry or learned during prenatal development. This opens the possibility that innate knowledge develops from domain-general learning algorithms. Digital twin studies provide computationally explicit evidence for this claim, illuminating both how and why innate knowledge exists in the first place. By simulating core algorithms learning from prenatal experiences, we can see that innate visual knowledge emerges spontaneously during development.

As reviewed in Section 4.6, domain-general algorithms learn high-level visual knowledge when trained solely on retinal waves, which are widely available during prenatal development. Retinal waves are highly structured across space and time, mimicking the second-order correlations of natural images (Albert et al. 2008). When domain-general algorithms are trained on retinal waves, the systems learn domain-specific knowledge about the visual world. This finding (partially) explains how newborn animals (e.g., chicks, babies) could have an object-based inductive bias at birth (Section 3.2): An object-based inductive bias is an emergent property of a core domain-general algorithm (space-time fitting) learning from prenatal experiences.

In consequence, we do not see space-time fitting as a direct challenge to nativist theories (**Figure 5*b***). Instead, space-time fitting provides a unifying explanation for why innate knowledge exists in the first place. Innate knowledge might be the outcome of a core domain-general learning system adapting (fitting) to prenatal training data (e.g., retinal waves). We hypothesize

that space-time fitting algorithms are the medium from which core knowledge emerges, akin to DNA being the medium from which animal bodies emerge. Mental skills and animal species are both emergent phenomena of high-dimensional systems (brains and DNA, respectively) fitting to the environment on different timescales (development versus evolution).

## 5.2. Conclusions and Next Steps

These are the early days of this research program. We have only tested a handful of models and tasks with the digital twin approach. Despite promising initial findings with domain-general algorithms, it may turn out that more domain-specific algorithms will best match newborn animals. To promote a community-wide effort to address this classic debate and reverse engineer the core learning algorithms in brains, we developed a public website—the Origins of Intelligence Testbed—that allows researchers to directly test whether ML algorithms learn like newborn animals across a range of tasks. Researchers can download virtual environments that mimic the environments of the newborn chicks, insert ML algorithms into artificial chicks, and raise artificial chicks in the same training and test environments as biological chicks. Our hope is that this testbed will link nativist and empiricist views, creating a unified framework for studying the origins of intelligence.

### SUMMARY POINTS

1. The nativist versus empiricist debate (also known as the nature versus nurture debate) is one of the oldest debates in the mind sciences. This debate concerns the learning algorithms underlying intelligence and the role of experience in building knowledge.
2. Nativists argue that biological intelligence emerges from a collection of innate, domain-specific systems for learning about different kinds of things (e.g., objects, agents, places, and numbers). Empiricists argue that biological intelligence emerges from domain-general learning faculties that develop domain-specific knowledge from experience.
3. We address this debate by introducing digital twin studies designed to reverse engineer the learning algorithms in newborn brains. In digital twin studies, newborn animals and artificial agents are raised in the same environments and tested with the same tasks, permitting direct comparison of their learning abilities.
4. Supporting empiricism, digital twin studies show that domain-general algorithms from artificial intelligence learn animal-like object perception when trained on the first-person visual experiences of newborn animals. Supporting nativism, digital twin studies show that domain-general algorithms learn innate domain-specific knowledge when trained on prenatal experiences (retinal waves).
5. We argue that these findings—and visual learning more generally—can be explained by a universal principle that we call space-time fitting. Space-time fitting provides a common framework for understanding learning across humans, animals, and machines.
6. Space-time fitting unifies concepts related to learning in brains and machines under a common principle. Different concepts across biological and artificial intelligence reflect the same general principle of high-dimensional systems iteratively adapting (fitting) to the spatiotemporal data distributions in the visual environment.
7. We conclude that space-time fitting explains both empiricist and nativist phenomena, providing a unified framework for understanding the origins of intelligence.

## FUTURE ISSUES

1. Digital twin studies suggest that domain-general algorithms can explain visual learning in newborn chicks. Can domain-general algorithms also explain visual learning in other species, including humans?
2. Space-time fitting provides a unified framework for understanding visual learning across humans, animals, and machines. Can space-time fitting also explain learning in other sensory domains (e.g., audition, proprioception) and cognitive capacities (e.g., navigation, social cognition, language, decision making)?
3. Space-time fitting algorithms trained on prenatal visual experience (retinal waves) develop domain-specific knowledge, providing an explanation for why innate knowledge exists in the first place. Can learning from prenatal experiences also explain other canonical innate abilities documented in developmental psychology and animal behavior?
4. Digital twin studies show that domain-general models can be transformed into domain-specific models by fitting them to prenatal data distributions, thereby producing innate object knowledge (i.e., knowledge emerging before models gain visual experience with objects). However, it is unclear how postnatal visual learning interacts with prenatally trained networks to produce mature object knowledge.
5. The blind, brute-force learning processes underlying space-time fitting share much in common with the blind, brute-force fitting processes underlying natural selection. Can evolution and development be united under a common framework in which organisms adapt to the environment at different timescales (i.e., evolution as slow fitting and development as faster fitting)?
6. Space-time fitting algorithms iteratively adapt to environmental data distributions. How much do space-time data distributions vary within and between species? How might differences in natural perceptual experiences generate different perceptual knowledge across species?

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

This work was funded by a National Science Foundation CAREER grant (BCS-1351892 to J.N.W.), a James S. McDonnell Foundation Understanding Human Cognition Scholar award (to J.N.W.), and a Facebook Artificial Intelligence Research award (to J.N.W.).

## LITERATURE CITED

Adolph KE, Hoch JE. 2019. Motor development: embodied, embedded, enculturated, and enabling. *Annu. Rev. Psychol.* 70:141–64

Albert MV, Schnabel A, Field DJ. 2008. Innate visual learning through spontaneous activity patterns. *PLOS Comput. Biol.* 4(8):e1000137

Applegate MC, Gutnichenko KS, Mackevicius EL, Aronov D. 2023. An entorhinal-like region in food-caching birds. *Curr. Biol.* 33(12):2465–77.e7

Arroyo DA, Kirkby LA, Feller MB. 2016. Retinal waves modulate an intraretinal circuit of intrinsically photosensitive retinal ganglion cells. *J. Neurosci.* 36(26):6892–905

ATLAS Collab., Aad G, Abajyan T, Abbott B, Abdallah J, et al. 2012. A particle consistent with the Higgs boson observed with the ATLAS detector at the Large Hadron Collider. *Science* 338(6114):1576–82

Azevedo FA, Carvalho LR, Grinberg LT, Farfel JM, Ferretti RE, et al. 2009. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* 513(5):532–41

Bear DM, Feigelis K, Chen H, Lee W, Venkatesh R, et al. 2023. Unifying (machine) vision via counterfactual world modeling. arXiv:2306.01828 [cs.CV]

Binz M, Schulz E. 2023. Using cognitive psychology to understand GPT-3. *PNAS* 120(6):e2218523120

Blankenship AG, Feller MB. 2010. Mechanisms underlying spontaneous patterned activity in developing neural circuits. *Nat. Rev. Neurosci.* 11(1):18–29

Boser BE, Guyon IM, Vapnik VN. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–52. New York: ACM

Buckner C. 2023. *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us About the Future of Artificial Intelligence*. Oxford, UK: Oxford Univ. Press

Bulf H, Johnson SP, Valenza E. 2011. Visual statistical learning in the newborn infant. *Cognition* 121(1):127–32

Calabrese A, Woolley SM. 2015. Coding principles of the canonical cortical microcircuit in the avian brain. *PNAS* 112(11):3517–22

Canny J. 1986. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intel.* (6):679–98

Cao YH, Wu J. 2022. A random CNN sees objects: one inductive bias of CNN and its applications. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI-22)*, pp. 194–202. Washington, DC: Assoc. Adv. Artif. Intel.

Carey S. 2009. *The Origin of Concepts*. Oxford, UK: Oxford Univ. Press

Chen L, Lu K, Rajeswaran A, Lee K, Grover A, et al. 2021. Decision transformer: reinforcement learning via sequence modeling. *Adv. Neural Inform. Process. Syst.* 34:15084–97

Chen T, Kornblith S, Norouzi M, Hinton G. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–607. Red Hook, NY: Curran Assoc.

Cox DD, Meier P, Oertelt N, DiCarlo JJ. 2005. "Breaking" position-invariant object recognition. *Nat. Neurosci.* 8(9):1145–47

Dalal N, Triggs B. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 886–93. Piscataway, NJ: IEEE

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL]

DiCarlo JJ, Zoccolan D, Rust NC. 2012. How does the brain solve visual object recognition? *Neuron* 73(3):415–34

Dobs K, Yuan J, Martinez J, Kanwisher N. 2023. Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *PNAS* 120(32):e2220642120

Doerig A, Sommers RP, Seeliger K, Richards B, Ismael J, et al. 2023. The neuroconnectionist research programme. *Nat. Rev. Neurosci.* 24(7):431–50

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. 2020. An image is worth 16×16 words: transformers for image recognition at scale. arXiv:2010.11929 [cs.CV]

Dugas-Ford J, Rowell JJ, Ragsdale CW. 2012. Cell-type homologies and the origins of the neocortex. *PNAS* 109(42):16974–79

Feichtenhofer C, Li Y, He K. 2022. Masked autoencoders as spatiotemporal learners. *Adv. Neural Inform. Process. Syst.* 35:35946–58

Feldman J, Tremoulet PD. 2006. Individuation of visual objects over time. *Cognition* 99(2):131–65

Földiák P. 1991. Learning invariance from transformation sequences. *Neural Comput.* 3(2):194–200

Garimella M, Pak D, Wood SMW, Wood JN. 2024. *A newborn embodied Turing test for comparing object segmentation across animals and machines*. Paper presented at the 12th International Conference on Learning Representations, Vienna, May 7–11

Ge X, Zhang K, Gribizis A, Hamodi AS, Sabino AM, Crair MC. 2021. Retinal waves prime visual motion detection by simulating future optic flow. *Science* 373(6553):eabd0830
Gibson EJ. 1963. Perceptual learning. *Annu. Rev. Psychol.* 14:29–56
Gibson JJ. 1979. *The Ecological Approach to Visual Perception: Classic Edition*. Boston: Houghton Mifflin
Goldman JG, Wood JN. 2015. An automated controlled-rearing method for studying the origins of movement recognition in newly hatched chicks. *Anim. Cogn.* 18:723–31
Gomez-Villa A, Martin A, Vazquez-Corral J, Bertalmío M. 2019. Convolutional neural networks can be deceived by visual illusions. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12309–17. Piscataway, NJ: IEEE
Hasson U, Nastase SA, Goldstein A. 2020. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* 105(3):416–34
He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–9. Piscataway, NJ: IEEE
Held R, Hein A. 1963. Movement-produced stimulation in the development of visually guided behavior. *J. Comp. Physiol. Psychol.* 56(5):872–76
Hume D. 1739. *A Treatise of Human Nature*. London: John Noon
Janini D, Hamblin C, Deza A, Konkle T. 2022. General object-based features account for letter perception. *PLOS Comput. Biol.* 18(9):e1010522
Jarvis ED, Güntürkün O, Bruce L, Csillag A, Karten H, et al. 2005. Avian brains and a new understanding of vertebrate brain evolution. *Nat. Rev. Neurosci.* 6(2):151–59
Kanwisher N, Khosla M, Dobs K. 2023. Using artificial neural networks to ask "why" questions of minds and brains. *Trends Neurosci*. 46(3):240–54
Karten HJ. 2013. Neocortical evolution: Neuronal circuits arise independently of lamination. *Curr. Biol.* 23(1):R12–15
Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98(3):630–44
Kirkham NZ, Slemmer JA, Johnson SP. 2002. Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition* 83(2):B35–42
Kriegeskorte N, Douglas PK. 2018. Cognitive computational neuroscience. *Nat. Neurosci.* 21(9):1148–60
Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. 2017. Building machines that learn and think like people. *Behav. Brain Sci.* 40:e253
LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44
Lee D, Gujarathi P, Wood JN. 2021. Controlled-rearing studies of newborn chicks and deep neural networks. arXiv:2112.06106 [cs.CV]
Li N, DiCarlo JJ. 2008. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* 321(5895):1502–7
Li N, DiCarlo JJ. 2010. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* 67(6):1062–75
Li Y, Anumanchipalli GK, Mohamed A, Lu J, Wu J, Chang EF. 2022. Dissecting neural computations of the human auditory pathway using deep neural networks for speech. *Nat. Neurosci.* 26:2213–25
Lindsay G. 2021. *Models of the Mind: How Physics, Engineering and Mathematics Have Shaped Our Understanding of the Brain*. London: Bloomsbury Publ.
Liu S, Deng W. 2015. Very deep convolutional neural network based image classification using small training sample size. In *Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition*, pp. 730–34. Piscataway, NJ: IEEE
Locke J. 1690. *An Essay Concerning Human Understanding*. London: Thomas Basset
Loken E, Gelman A. 2017. Measurement error and the replication crisis. *Science* 355(6325):584–85
Lowe DG. 1999. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision*, Vol. 2, pp. 1150–57. Piscataway, NJ: IEEE
Marr D, Hildreth E. 1980. Theory of edge detection. *Proc. R. Soc. Lond. B* 207(1167):187–217

Michaels JA, Schaffelhofer S, Agudelo-Toro A, Scherberger H. 2020. A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *PNAS* 117(50):32124–35

Mitchell TM. 1980. *The need for biases in learning generalizations*. Work. Pap. CBM-TR-117, Rutgers Univ., New Brunswick, NJ

Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, et al. 2017. A manifesto for reproducible science. *Nat. Hum. Behav.* 1:0021

Nayebi A, Attinger A, Campbell M, Hardcastle K, Low I, et al. 2021. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *Adv. Neural Inform. Process. Syst.* 34:12167–79

Newell A. 1973. *You can't play 20 questions with nature and win: projective comments on the papers of this symposium*. Tech. Rep., School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA

Pandey L, Wood SMW, Wood JN. 2023. Are vision transformers more data hungry than newborn visual systems? arXiv:2312.02843 [cs.CV]

Pinker S. 2002. *The Blank Slate: The Modern Denial of Human Nature*. London: Penguin

Prasad A, Wood SMW, Wood JN. 2019. Using automated controlled rearing to explore the origins of object permanence. *Dev. Sci.* 22(3):e12796

Radosavovic I, Xiao T, James S, Abbeel P, Malik J, Darrell T. 2023. Real-world robot learning with masked visual pre-training. In *Proceedings of the 2023 Conference on Robot Learning*, pp. 416–26. Cambridge, MA: PMLR

Reed S, Zolna K, Parisotto E, Colmenarejo SG, Novikov A, et al. 2022. A generalist agent. arXiv:2205.06175 [cs.AI]

Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, et al. 2019. A deep learning framework for neuroscience. *Nat. Neurosci.* 22(11):1761–70

Ritter S, Barrett DG, Santoro A, Botvinick MM. 2017. Cognitive psychology for deep neural networks: a shape bias case study. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2940–49. Cambridge, MA: PMLR

Saffran JR, Aslin RN, Newport EL. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294):1926–28

Schneider F, Xu X, Ernst MR, Yu Z, Triesch J. 2021. *Contrastive learning through time*. Paper presented at the NeurIPS Shared Visual Representations in Humans and Machines Workshop, virtual, Dec. 13

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, et al. 2021. The neural architecture of language: integrative modeling converges on predictive processing. *PNAS* 118(45):e2105646118

Schrimpf M, Kubilius J, Lee MJ, Murty NAR, Ajemian R, DiCarlo JJ. 2020. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* 108(3):413–23

Shanahan M, Bingman VP, Shimizu T, Wild M, Güntürkün O. 2013. Large-scale network organization in the avian forebrain: a connectivity matrix and theoretical analysis. *Front. Comput. Neurosci.* 7:89

Sheybani S, Hansaria H, Wood JN, Smith LB, Tiganj Z. 2023. *Curriculum learning with infant egocentric videos*. Paper presented at the 37th Annual Conference on Neural Information Processing Systems, New Orleans, LA, Dec. 10–16

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22(11):1359–66

Skinner BF. 1938. *The Behavior of Organisms: An Experimental Analysis*. New York: Appleton-Century

Sobel I, Feldman G. 1968. *A 3×3 isotropic gradient operator for image processing*. Talk, Stanford Artificial Intelligence Laboratory, Stanford, CA

Spelke ES. 2022. *What Babies Know: Core Knowledge and Composition*, Vol. 1. Oxford, UK: Oxford Univ. Press

Spelke ES, Newport EL. 1998. Nativism, empiricism, and the development of knowledge. In *Handbook of Child Psychology: Theoretical Models of Human Development*, ed. W Damon, RM Lerner, pp. 275–340. Hoboken, NJ: John Wiley & Sons

Stone JV. 1996. Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Comput.* 8(7):1463–92

Storrs KR, Anderson BL, Fleming RW. 2021. Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* 5(10):1402–17

Tong Z, Song Y, Wang J, Wang L. 2022. VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. *Adv. Neural Inform. Process. Syst.* 35:10078–93

Ulhaq A, Akhtar N, Pogrebna G, Mian A. 2022. Vision transformers for action recognition: a survey. arXiv:2209.05700 [cs.CV]
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. arXiv:1706.03762 [cs.CL]
Volk AA, Epps RW, Yonemoto DT, Masters BS, Castellano FN, et al. 2023. AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nat. Commun.* 14:1403
Walk RD, Gibson EJ, Tighe TJ. 1957. Behavior of light- and dark-reared rats on a visual cliff. *Science* 126(3263):80–81
Wallis G, Bülthoff HH. 2001. Effects of temporal association on recognition memory. *PNAS* 98(8):4800–4
Wallis G, Rolls ET. 1997. Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51(2):167–94
Wang HC, Bergles DE. 2015. Spontaneous activity in the developing auditory system. *Cell Tissue Res.* 361:65–75
Wang PY, Sun Y, Axel R, Abbott LF, Yang GR. 2021. Evolving the olfactory system with machine learning. *Neuron* 109(23):3879–92
Wang Y, Brzozowska-Prechtl A, Karten HJ. 2010. Laminar and columnar auditory cortex in avian brain. *PNAS* 107(28):12676–81
Watson JB. 1913. Psychology as the behaviorist views it. *Psychol. Rev.* 20(2):158–177
Wenner P. 2012. Motor development: Activity matters after all. *Curr. Biol.* 22(2):R47–48
Whittington JC, McCaffary D, Bakermans JJ, Behrens TE. 2022. How to build a cognitive map. *Nat. Neurosci.* 25(10):1257–72
Wiskott L, Sejnowski TJ. 2002. Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* 14(4):715–70
Wolpert DH, Macready WG. 1997. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1:67–82
Wood JN. 2013. Newborn chickens generate invariant object representations at the onset of visual object experience. *PNAS* 110(34):14000–5
Wood JN. 2014. Newly hatched chicks solve the visual binding problem. *Psychol. Sci.* 25(7):1475–81
Wood JN. 2016. A smoothness constraint on the development of object recognition. *Cognition* 153:140–45
Wood JN, Prasad A, Goldman JG, Wood SMW. 2016. Enhanced learning of natural visual sequences in newborn chicks. *Anim. Cogn.* 19:835–45
Wood JN, Ullman TD, Wood BWW, Spelke ES, Wood SMW. 2024. Object permanence in newborn chicks is robust against opposing evidence. arXiv:2402.14641 [q-bio.NC]
Wood JN, Wood SMW. 2016. The development of newborn object recognition in fast and slow visual worlds. *Proc. R. Soc. B* 283(1829):20160166
Wood JN, Wood SMW. 2017. Measuring the speed of newborn object recognition in controlled visual worlds. *Dev. Sci.* 20(4):e12470
Wood JN, Wood SMW. 2018. The development of invariant object recognition requires visual experience with temporally smooth objects. *Cogn. Sci.* 42(4):1391–406
Wood JN, Wood SMW. 2020. One-shot learning of view-invariant object representations in newborn chicks. *Cognition* 199:104192
Wood SMW, Wood JN. 2015. A chicken model for studying the emergence of invariant object recognition. *Front. Neural Circuits* 9:7
Wood SMW, Wood JN. 2019. Using automation to combat the replication crisis: a case study from controlled-rearing studies of newborn chicks. *Infant Behav. Dev.* 57:101329
Wood SMW, Wood JN. 2021a. Distorting face representations in newborn brains. *Cogn. Sci.* 45(8):e13021
Wood SMW, Wood JN. 2021b. One-shot object parsing in newborn chicks. *J. Exp. Psychol. Gen.* 150(11):2408–20
Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* 111(23):8619–24

Yang J, Dong X, Liu L, Zhang C, Shen J, Yu D. 2022. Recurring the transformer for video action recognition. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14063–73. Piscataway, NJ: IEEE

Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. 2014. *Learning deep features for scene recognition using places database*. Paper presented at the 28th Annual Conference on Neural Information Processing Systems, Montreal, Can., Dec. 8–13

Zhou HY, Lu C, Yang S, Yu Y. 2021. Convnets versus transformers: Whose visual representations are more transferable? In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*, pp. 2230–38. Piscataway, NJ: IEEE