# Bridging Minds and Machines: Leveraging Developmental Psychology to Advance AI

Marvin Theiß
Student ID: [redacted]

S2: Development of Perception and Action
MBB–MA–THM–2: Perception and Action

March 25, 2024

---

## Introduction

Ever since OpenAI unveiled *ChatGPT* in late November 2022, artificial intelligence (AI) has captured the world's attention. In the past year alone, the field has advanced with such velocity that it has prompted significant regulatory discussions, most notably within the European Union which proposed the *Artificial Intelligence Act* in 2023 (Browne, 2023; Satariano, 2023). This rapid development is evidenced by AI's achievements across diverse domains, from image classification (Krizhevsky et al., 2012) and board games (Silver et al., 2017) to language modeling (Brown et al., 2020) and protein folding (Senior et al., 2020). In many tasks, the capabilities of state-of-the-art systems, especially deep neural networks (DNNs), now surpass those of humans.

**A Fundamental Gap.** Yet, a closer examination reveals a fundamental disconnect between these systems[1] and the nuanced intelligence exhibited by humans (Lake et al., 2017). For instance, when it comes to image classification, DNNs predominantly utilize texture cues, diverging from humans who primarily lean on shape (Baker et al., 2018; Brendel & Bethge, 2019; Geirhos et al., 2018). Additionally, these networks often struggle with 3D viewpoint variations (Dong et al., 2022), exhibit vulnerability to

[1] Throughout this essay, I will refrain from adopting the widespread use of the term *model* when referring to DNNs, for the reasons detailed in Wichmann and Geirhos (2023).

image distortions, and make errors that are qualitatively dissimilar to those made by humans (Geirhos et al., 2021). Crucially, unlike humans, who can learn complex tasks with relatively sparse data, DNNs typically require vast quantities of labeled data (Huber et al., 2023), and they lack a fundamental grasp of common sense and causal reasoning, areas where even infants outshine these artificial systems (Piloto et al., 2022; Stojnić et al., 2023). Moreover, when assessed on physical concepts, even advanced video networks trained on large-scale datasets falter, performing no better than random chance (Weihs et al., 2022).

**Toward Intuitive AI.** To bridge these gaps, an emerging body of research is turning to developmental psychology for insights. This includes the creation of benchmarks designed to assess a network's ability to discriminate between physically plausible and implausible scenarios, such as[2] *IntPhys* (Riochet et al., 2020), the *Physical Concepts* dataset (Piloto et al., 2022), and InfLevel (Weihs et al., 2022). These benchmarks notably leverage the *violation-of-expectation* (VoE) paradigm, a cornerstone of infant cognition research (Margoni et al., 2023). In addition, there are promising developments in network design that reflect an emerging understanding of physical principles, and in unsupervised training schemes that attempt to mimic the predominantly exploratory learning characteristic of infancy.

Drawing on basic principles of developmental psychology, in particular the early stages of intuitive physics acquisition in infancy, this essay aims to explore how these concepts are being integrated into AI development. The essay is organized as follows: First, I will lay the groundwork by discussing the role of developmental psychology and the concept of intuitive physics in infancy. I will then review Stetter and Lang's (2021) work on state-action-prediction self-organizing maps, followed by Piloto et al.'s (2022) approach to learning physical concepts through object tracking. Both groups depart from traditional approaches in AI by deliberately basing their network architectures and training schemes on insights from developmental psychology. Finally, I will assess the extent to which these projects draw from developmental psychology, their implications for AI development, as well as challenges and future directions. Ultimately, this essay aims

[2]This list is by no means exhaustive. For a recent overview of benchmarks for physical reasoning see Melnik et al. (2023).

to demonstrate the potential of leveraging insights from developmental psychology to forge AI systems that are not only capable of performing narrow tasks, but also exhibit elements of human-like intelligence.

## Understanding Intuitive Physics: Insights from Infancy

The foundational studies of infant physical reasoning can be traced back to Jean Piaget, who in the mid-20th century posited that infants' cognitive abilities develop only gradually, suggesting a lack of understanding such as object permanence until the age of 8-9 months (Piaget, 1952, 1954). According to Piaget, infants gradually develop these and other aspects of physical reasoning through a series of developmental stages, first interacting with their environment in a very direct and sensorimotor manner, before developing the ability to perform mental operations on abstract concepts (Piaget, 1952, 1954).

**The Violation-of-Expectation Paradigm.** However, these assumptions have been substantially revised, thanks in large part to research conducted by Renée Baillargeon in the late 1980s. Using the *violation-of-expectation* paradigm[3], Baillargeon's studies offered compelling evidence that infants possess a more sophisticated understanding of the physical world at much earlier stages than Piaget had theorized. For instance, Baillargeon and colleagues showed that already at three to four months of age, infants exhibit surprise at events that violate their expectations of object permanence, suggesting a deeper, more *innate* cognitive processing than previously recognized (Baillargeon, 1987; Baillargeon et al., 1985).

**Core Knowledge.** This paradigm shift, initiated by Baillargeon's findings and subsequently supported by further studies, has led to the development of the *core knowledge hypothesis*, which suggests that infants possess an *innate* understanding of several key physical-reasoning principles (Spelke et al., 1992). These principles are now thought to be gravity, inertia, and persistence; with the principle of persistence (objects persist in time and space as they are) consisting of the five subcategories boundedness, cohesion, continuity, solidity, and unchangeableness (Lin et al., 2022).

**Bridging Infant Cognition and AI.** The intuitive physics evident in infancy provides a blueprint for developing AI systems with a human-like comprehension of the physical

[3]See Margoni et al. (2023) for a recent overview of the paradigm.

world. By emulating the way infants learn about and interact with their environment, AI researchers have the opportunity to create systems that go beyond mere pattern recognition, and that understand the causal relationships inherent in the physical world, mirroring human cognition. The works of Stetter and Lang (2021) and Piloto et al. (2022) exemplify efforts to incorporate developmental insights into AI, with the former focusing on unsupervised learning similar to infant exploration, and the latter emphasizing object-level representations and the application of the VoE paradigm.

## State-Action-Prediction Maps

Stetter and Lang (2021) introduce a minimal network architecture, called *SapSom* (**S**tate-**A**ction-**P**rediction **S**elf-**O**rganizing **M**ap), that mimics the unsupervised, exploratory learning characteristic of human infants. Their network, situated in a simplified cart-pole environment, learns to understand the dynamic properties of its environment solely through interaction, without a predefined task or goal, mirroring the unstructured nature of infant exploration. This setup differs from traditional AI training paradigms, which often rely on supervised learning procedures, predefined tasks, and large labeled datasets, thus embracing a developmental psychology perspective that emphasizes the importance of exploration in infant learning.

**Experimental Design.** All of the experiments conducted by Stetter and Lang (2021) are performed in the cart-pole environment (Brockman et al., 2016). In this environment, a pole is hinged (via an unactuated joint) to a cart that can move left or right on a frictionless track (see Fig. 1). The cart can be pushed from the left or the right with a fixed force, i.e., "push left" and "push right" are the only two actions $a_t$ that can be performed at any given time $t$. The environment is fully described by a four-dimensional vector $\mathbf{u}$ with the following entries: cart position $x$, cart velocity $\dot{x}$, pole angle $\theta$, and pole angular velocity $\dot{\theta}$ (Brockman et al., 2016). Within this environment, Stetter and Lang (2021) conduct three experiments.

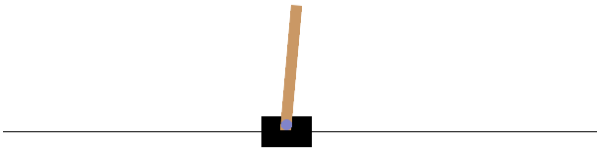

**Figure 1.** The cart-pole environment.

*Note.* From *Gymnasium, Classic Control Environments, Cart Pole,* by Farama Foundation, 2023, (https://gymnasium.farama.org/environments/classic_control/cart_pole/). Copyright 2023 by Farama Foundation.

*Learning Intuitive Physics.* Five random action sequences are executed and the network is asked to predict the state of the environment at time $t + 1$ based on the current state $\mathbf{u}_t$ and the performed action $a_t$. The predicted state $\hat{\mathbf{u}}_{t+1}$ is then compared to the true state $\mathbf{u}_{t+1}$. This experiment tests whether the network is at all capable of learning and internalizing the physical dynamics that govern the cart-pole system.

*Playing Virtual Episodes.* The network is given a starting state $\mathbf{u}_0$ along with action sequences $a_t$, and must predict the environment's state $u_t$ at multiple consecutive future times $t$. This is a more rigorous test than the first experiment, as small errors accumulate over time, since the prediction of $\hat{\mathbf{u}}_{t+1}$ is based on $\mathbf{u}_0$ and the previous *predictions* $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_t$, rather than on the *true* state $\mathbf{u}_t$.

*One-Shot Imitation.* The network is tested on its ability "to perform one-shot imitation, i.e., its ability to learn from very few demonstrations of any given task, and instantly generalize to new situations of the same task, without requiring task-specific engineering" (Duan et al., 2017). To do this, the network is presented with a sequence of states $\mathbf{u}_t$, and it must find a sequence of actions $a_t$ that drives the environment's state toward the desired target state (e.g., keeping the pole tilted at a constant angle).

**Network Architecture.** The network implemented by Stetter and Lang (2021) is based on the *self-organizing map* (SOM) framework (Kohonen, 1982). In an SOM, neurons are arranged in a 2D map, and each neuron carries with it a representation $\mathbf{w}(\mathbf{s})$ of the environment, where $\mathbf{s} \in \mathbb{R}^2$ denotes the spatial location of the neuron within the 2D map. The vector $\mathbf{w}(\mathbf{s})$ has the same dimension as $\mathbf{u}$ so that, for each state $\mathbf{u}$, there exists a neuron $\mathbf{s}^*(\mathbf{u})$, called the *winning unit*, which *best* represents the environment in the sense that its representation $\mathbf{w}(\mathbf{s}^*(\mathbf{u}))$ is closest to

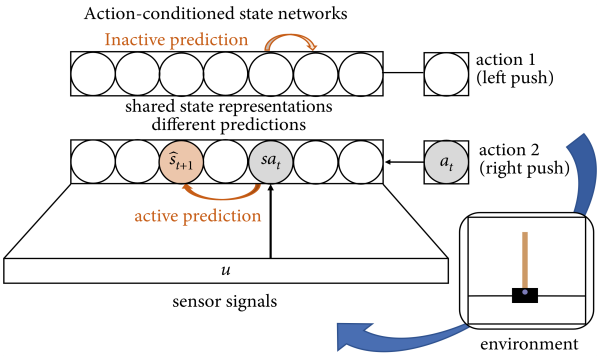


**Figure 2.** The architecture implemented by Stetter and Lang (2021). Two action-conditioned state networks learn to predict the most likely future state $\hat{\mathbf{s}}_{t+1}$ based on the current state representation $\mathbf{s}_t$ and an action $a_t$. The current state representation $\mathbf{s}_t$ is inferred from an input vector $\mathbf{u}$ representing the environmental state.

*Note.* From "Learning intuitive physics and one-shot imitation using state-action-prediction self-organizing maps," by M. Stetter and E. W. Lang, 2021, *Computational Intelligence and Neuroscience, 2021*(5590445) (https://doi.org/10.1155/2021/5590445). CC BY 4.0.

$\mathbf{u}$ in the Euclidean sense, i.e.,

$$\mathbf{s}^*(\mathbf{u}) = \text{argmin}_{\mathbf{s}} \|\mathbf{u} - \mathbf{w}(\mathbf{s})\|_2.$$

Besides this 2D map, the network learns a set of state transition matrices $T_a$, where $T_a(\mathbf{s}, \tilde{\mathbf{s}})$ denotes the probability that unit $\mathbf{s}$ is active one time step after unit $\tilde{\mathbf{s}}$ was active and action $a$ was performed. In this setup, one-step prediction is performed as follows: after presenting the current state of the environment $\mathbf{u}_t$ and an action $a_t$ to the network, the *winning unit* $\mathbf{s}^*(\mathbf{u}_t)$ is determined. The next map state $\hat{\mathbf{s}}_{t+1}$ is predicted to be the unit $\mathbf{s}$ that maximizes the probability $T_{a_t}(\mathbf{s}, \mathbf{s}^*(\mathbf{u_t}))$ of being active after action $a_t$ is performed. Finally, the predicted environment's state $\hat{\mathbf{u}}_{t+1}$ at time $t+1$ is given by $\mathbf{w}(\hat{\mathbf{s}}_{t+1})$. This predicted state can be used as a starting point for yet another one-step prediction, and by chaining multiple one-step predictions together, SOMs can predict entire (environment) state sequences $\mathbf{u}_t$, given an action sequence $a_t$ (Stetter & Lang, 2021).

**Training.** During training, the network has to learn to (a) correctly represent the environment by the activations $\mathbf{w}(\mathbf{s})$ of the neurons that make up the 2D state map, and (b) learn the action-conditioned state transition matrices $T_a$ for all possible actions $a$. To do this, the network explores the environment by performing random action sequences $a_t$. At each time $t$, it compares the environment's state $\mathbf{u}_t$ with the activation $\mathbf{w}(\mathbf{s}^*(\mathbf{u}_t))$ of the winning unit $\mathbf{s}^*(\mathbf{u}_t)$, and updates its activation and the activations of nearby neurons to better reflect the environment's state $\mathbf{u}_t$. It then performs a one-step prediction to predict the next state $\hat{\mathbf{u}}_{t+1}$, executes the action $a_t$, and compares the resulting environment's state $\mathbf{u}_{t+1}$ with its prediction $\hat{\mathbf{u}}_{t+1}$ to update the action-conditioned state transition matrices $T_{a_t}$.

**Results.** Based on the observation that, in their first experiment, SapSom's predicted motion directions are quite close to the actual motion directions, Stetter and Lang (2021) conclude that the network is indeed able to learn a "reasonably accurate representation" of the dynamics of the cart-pole environment. For the multi-step prediction experiment, the authors find that, while the absolute deviation between the true and predicted pole angle $\theta$ increases over time, SapSom still correctly predicts the general behavior of the pole for multiple action sequences. Figure 3 illustrates the results for

one of the multi-step prediction experiments.

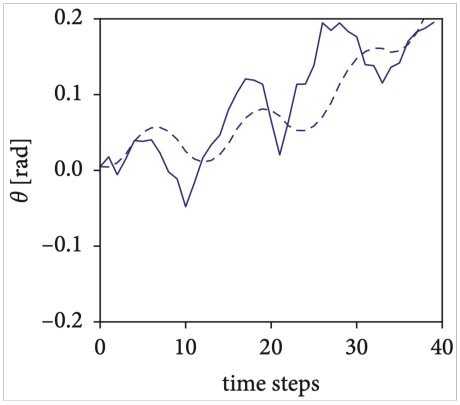


**Figure 3.** Real (dashed) and predicted (solid) evolution of the pole angle $\theta$ over time. Three initial left pushes are followed by alternating sequences of six right and six left pushes.

*Note.* Adapted from "Learning intuitive physics and one-shot imitation using state-action-prediction self-organizing maps," by M. Stetter and E. W. Lang, 2021, *Computational Intelligence and Neuroscience, 2021*(5590445) (https://doi.org/10.1155/2021/5590445). CC BY 4.0.

Finally, Stetter and Lang (2021) also attest SapSom a satisfactory performance in the one-shot imitation experiments.

## Learning Physics by Tracking Objects

In contrast to the elementary architecture proposed by Stetter and Lang (2021), Piloto et al. (2022) present a more sophisticated system designed to learn intuitive physics from visual data alone. Their approach differs from traditional approaches in AI by prioritizing object-level representations—a concept deeply rooted in developmental psychology, which posits that an understanding of physics emerges most naturally from the interactions between discrete objects. To evaluate their network, Piloto et al. leverage the violation-of-expectation paradigm, and apply it to five different physical concepts. The network's training and evaluation are conducted using the *Physical Concepts* dataset, a dataset curated by Piloto et al. (2022) themselves.

**Experimental Design.** Central to the research conducted by Piloto et al. (2022) is the introduction of the *Physical Concepts* dataset[4], a novel machine-learning video dataset inspired by the violation-of-expectation framework in developmental psychology. The dataset is made up of separate training and test sets. Videos in the training set display a variety of unstructured physical events. The test set specifically targets five physical concepts adapted from developmental psychology: continuity, object persistence, solidity, unchangeableness, and directional inertia. For each of these concepts, multiple probe quadruplets consisting of four videos each are available. In each probe quadruplet, two videos obey the physical concept under consideration, while the remaining two do not. These quadruplets are obtained by a

[4] https://github.com/deepmind/physical_concepts

splicing procedure illustrated in Figure 4.

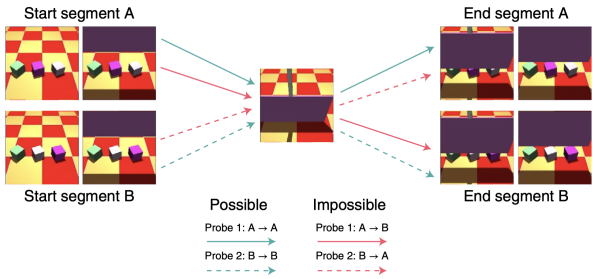


**Figure 4.** Illustration of the splicing procedure used by Piloto et al. (2022) to create the physical concept probes. Two videos, each representing a physically possible event, and sharing a common frame, are split into start and end segments at the common frame. The resulting two sets of start and end segments are recombined to create a total of four videos, two physically possible and two impossible.

*Note.* From "Intuitive physics learning in a deep-learning model inspired by developmental psychology," by L. S. Piloto, A. Weinstein, P. Battaglia, and M. Botvinick, 2022, *Nature Human Behaviour, 6*(9), p. 1260 (https://doi.org/10.1038/s41562-022-01394-8). CC BY 4.0.

Piloto et al. (2022) feed these videos to their network, at each frame ask the network to predict the next frame based on all previous frames, and then derive a measure of the network's surprise from these predictions. By comparing the network's surprise for physically possible scenes with its surprise elicited by physically impossible scenes, Piloto et al. (2022) can infer whether the network is able to discriminate between the two.

**Network Architecture.** Piloto et al. (2022) developed a two-component system, named PLATO (**P**hysics **L**earning Through **A**uto-Encoding and **T**racking **O**bjects), consisting of a feedforward *perception module* and a recurrent *dynamics module*. The perception module (Fig. 5) is based on the autoencoder architecture, which consists of an *encoder* $\theta$ and a *decoder* $\phi$. The encoder transforms the input data (single-object image-mask pairs) into an efficient representation (i.e., "object codes" $z_k$), which is then decoded back into an image-mask pair, the goal being to match

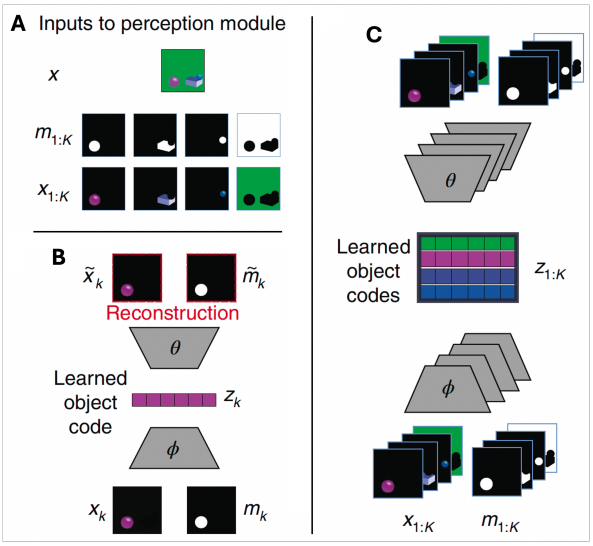


**Figure 5.** A *perception module* consisting of an encoder $\theta$ and a decoder $\phi$ is trained to encode useful features of individual image-mask pairs $(\tilde{x}_k, \tilde{m}_k)$ in object codes $z_k$.

*Note.* Adapted from "Intuitive physics learning in a deep-learning model inspired by developmental psychology," by L. S. Piloto, A. Weinstein, P. Battaglia, and M. Botvinick, 2022, *Nature Human Behaviour, 6*(9), p. 1259 (https://doi.org/10.1038/s41562-022-01394-8). CC BY 4.0.

the initial input as close as possible. Entire video sequences (multiple frames along with their segmentation masks) are then run through the encoder of the perception module to generate a sequence of object codes

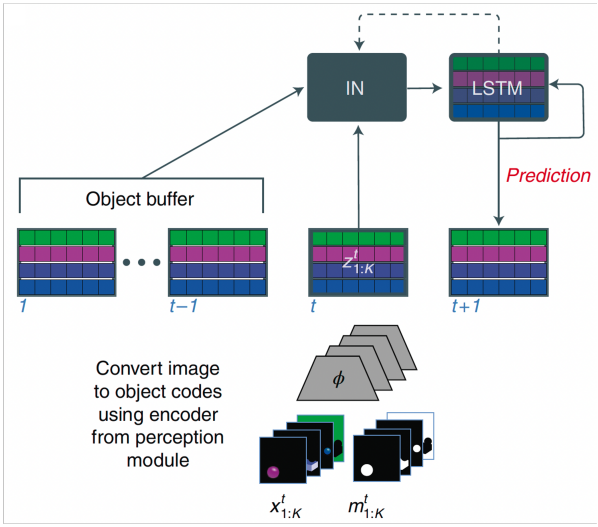


**Figure 6.** A *dynamics module* is trained to predict object codes $z^{t+1}_{1:K}$ at time $t+1$ based on the sequence of object codes $z^{1:t}_{1:K}$ obtained by passing video sequences through the encoder of the perception module.

*Note.* Adapted from "Intuitive physics learning in a deep-learning model inspired by developmental psychology," by L. S. Piloto, A. Weinstein, P. Battaglia, and M. Botvinick, 2022, *Nature Human Behaviour, 6*(9), p. 1259 (https://doi.org/10.1038/s41562-022-01394-8). CC BY 4.0.

$z^{1:t}_{1:K}$. These object codes are fed into the interaction network (IN) of the dynamics module (Fig. 6), which, in combination with a long short-term memory (LSTM) network (Hochreiter & Schmidhuber, 1997), computes interactions of the objects in a scene to predict future object codes $z^{t+1}_{1:K}$.

**Training.** PLATO's training consists of two distinct, sequential phases, each phase focusing on one of PLATO's components. First, the perception module is trained via an unsupervised method commonly used for autoencoders. In this process, training samples are fed into the network, which then encodes these inputs into efficient representations. These representations are then decoded back into outputs that match the dimensionality of the inputs. The network learns to represent the inputs efficiently by minimizing the reconstruction error (e.g., the mean squared error in pixel space), thereby increasing the fidelity of the output relative to the original input. Second, the dynamics module is trained to predict object codes $\hat{z}^{t+1}_{1:K}$ based on a sequence of previously "viewed" object codes $z^{1:t}_{1:K}$. The network learns by comparing its predictions to the true object codes $z^{t+1}_{1:K}$. The object codes used to train the dynamics module are obtained by passing video sequences through the perception module's encoder, whose weights remain fixed throughout the dynamics module's training.

**Results.** PLATO's understanding of intuitive physics is assessed by testing it on the *Physical Concepts* dataset, also developed by Piloto et al. (2022). As described in the *Experimental Design* section, Piloto et al. present both physically plausible and implausible scenes to PLATO, and measure

[5]Essentially, this is taken to be the squared error in pixel space, after decoding the true and predicted object codes $z^{t+1}_k$ and $\hat{z}^{t+1}_k$, respectively, using the perception module's decoder $\theta$. See Piloto et al. (2022) for

PLATO's surprise for each scene. *Surprise* is computed as the network's prediction error[5] for each frame, accumulated over all frames of a scene. An individual probe is "classified" correctly if the surprise for the two impossible scenes ("physically impossible surprise") is greater than the surprise caused by the two physically possible scenes ("physically possible surprise"). The proportion of correctly classified probes yields an accuracy score per physical concept. Additionally, the network's *relative surprise* is computed as the difference between the physically impossible and the physically possible surprise, normalized by the sum of the two.

Piloto et al. (2022) find that PLATO displays robust VoE effects across all five categories of physical concepts, both in terms of classification accuracy (significantly above chance) and mean relative surprise (significantly larger than 0). The evolution of relative surprise across individual frames is highly dependent on the physical concept tested, see Figure 7 for details.

PLATO was tested on three additional physical concepts using the ADEPT dataset (Smith et al., 2019) and found qualitatively identical results. Finally, Piloto et al. (2022) also tested two object-agnostic networks on the *Physical Concepts* dataset and found that they did not perform nearly as well as PLATO across the five different categories.

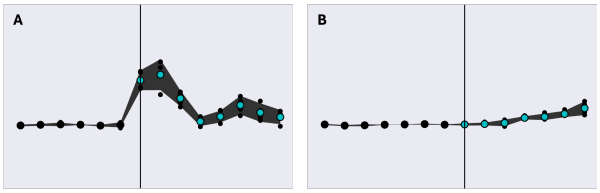


**Figure 7.** Frame-by-frame analysis of PLATO's relative surprise for two distinct physical concepts. **A** For the *continuity* condition, relative surprise increases steeply when a rolling ball fails to appear between two adjacent pillars. **B** For the *unchangeableness* condition, relative surprise increases gradually as an occluder slowly rises to reveal a physically impossible scenario.

*Note.* Adapted from "Intuitive physics learning in a deep-learning model inspired by developmental psychology," by L. S. Piloto, A. Weinstein, P. Battaglia, and M. Botvinick, 2022, *Nature Human Behaviour, 6*(9), p. 1261 (https://doi.org/10.1038/s41562-022-01394-8). CC BY 4.0.

## Discussion

The development of AI systems that emulate human cognition, in particular the seemingly effortless understanding of intuitive physics, has been a recent focal point of AI research. This essay has examined two contributions to this field, by Stetter and Lang (2021) and Piloto et al. (2022), from the perspective of developmental psychology.

Stetter and Lang (2021) have introduced a network that learns the dynamics of its environment in an unsupervised manner, mir-

details.

roring the exploratory learning observed in human infants. Their approach contrasts with traditional AI networks, which often rely heavily on supervised learning and large datasets, and highlights a shift toward more human-like learning processes in machines. In particular, unlike traditional reinforcement learning approaches, the learning process they implement does *not* depend on an extrinsic reward structure in the form of a laboriously hand-crafted loss function. This also allows their network to generalize reasonably well to previously unseen tasks.

On the other hand, Piloto et al. (2022) extend this discourse (a) by incorporating object-level representations into the design of their network PLATO, and (b) by explicitly employing the violation-of-expectation paradigm to test their network's understanding of intuitive physics. The authors also show that object-agnostic networks that are matched in learnable parameters and representational capacity do not perform nearly as well, highlighting the "strong facilitative role for object-level representation in the acquisition of intuitive physics concepts" (Piloto et al., 2022).

Both studies highlight the potential of using insights from developmental psychology to enhance the learning capabilities of AI systems. By employing unsupervised learning strategies and emphasizing object-level interactions, these approaches demonstrate a shift away from reliance on extensive labeled data and move closer to replicating the nuanced and flexible learning mechanisms inherent in human cognition.

However, translating the complexity of human learning, characterized by its exploratory nature and minimal reliance on explicit instruction, into AI systems remains a significant challenge. The path to creating AI systems with a comprehensive understanding of the physical world, similar to that of human infants, requires not only technical advances, but also a deeper integration of cognitive science principles.

In essence, the use of developmental psychology offers a promising avenue for advancing AI beyond narrow tasks to more generalized and adaptive systems. Stetter and Lang (2021) and Piloto et al. (2022) provide valuable frameworks for this endeavor, suggesting that future AI development could greatly benefit from further interdisciplinary research. This exploration points to the possibility of AI systems that not only perform tasks with high proficiency, but also possess a fundamental understanding of the world that parallels human cognition.

## References

Baillargeon, R. (1987). Object permanence in 3 1/2- and 4 1/2-month-old infants. *Developmental Psychology*, *23*(5), 655–664. https://doi.org/10.1037/0012-1649.23.5.655

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208. https://doi.org/10.1016/0010-0277(85)90008-3

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, *14*(12), e1006613. https://doi.org/10.1371/journal.pcbi.1006613

Brendel, W., & Bethge, M. (2019). Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. https://doi.org/10.48550/arXiv.1904.00760

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). OpenAI Gym. https://doi.org/10.48550/arXiv.1606.01540

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 1877–1901, Vol. 33). Curran Associates, Inc.

Browne, R. (2023, June 14). EU lawmakers pass landmark artificial intelligence regulation. *CNBC*. https://www.cnbc.com/2023/06/14/eu-lawmakers-pass-landmark-artificial-intelligence-regulation.html

Dong, Y., Ruan, S., Su, H., Kang, C., Wei, X., & Zhu, J. (2022). ViewFool: Evaluating the robustness of visual recognition to adversarial viewpoints. https://doi.org/10.48550/arXiv.2210.03895

Duan, Y., Andrychowicz, M., Stadie, B., Jonathan Ho, O., Schneider, J., Sutskever, I., Abbeel, P., & Zaremba, W. (2017). One-shot imitation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 1087–1098, Vol. 30). Curran Associates, Inc.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. https://doi.org/10.48550/arXiv.2106.07411

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. https://doi.org/10.48550/arXiv.1811.12231

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huber, L. S., Geirhos, R., & Wichmann, F. A. (2023). The developmental trajectory of object recognition robustness: Children are like small adults but unlike big deep neural networks. *Journal of Vision*, *23*(7), 1–30. https://doi.org/10.1167/jov.23.7.4

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 1097–1105, Vol. 25). Curran Associates, Inc.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Lin, Y., Stavans, M., & Baillargeon, R. (2022). Infants' physical reasoning and the cognitive architecture that supports it. In O. Houdé & G. Borst (Eds.), *The Cambridge handbook of cognitive development* (1st ed., pp. 168–194). Cambridge University Press. https://doi.org/10.1017/9781108399838.012

Margoni, F., Surian, L., & Baillargeon, R. (2023). The violation-of-expectation paradigm: A conceptual overview. *Psychological Review*. https://doi.org/10.1037/rev0000450

Melnik, A., Schiewer, R., Lange, M., Muresanu, A., Saeidi, M., Garg, A., & Ritter, H. (2023). Benchmarks for physical reasoning AI. https://doi.org/10.48550/arXiv.2312.10728

Piaget, J. (1952). *The origins of intelligence in children* (M. Cook, Trans.). W. W. Norton & Co. https://doi.org/10.1037/11494-000

Piaget, J. (1954). *The construction of reality in the child* (M. Cook, Trans.). Basic Books. https://doi.org/10.1037/11168-000

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, *6*(9), 1257–1267. https://doi.org/10.1038/s41562-022-01394-8

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2020). IntPhys: A framework and benchmark for visual intuitive physics reasoning. https://doi.org/10.48550/arXiv.1803.07616

Satariano, A. (2023, June 14). Europeans take a major step toward regulating A.I. *The New York Times*. https://www.nytimes.com/2023/06/14/technology/europe-ai-regulation.html

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. https://doi.org/10.48550/arXiv.1712.01815

Smith, K., Mei, L., Yao, S., Wu, J., Spelke, E., Tenenbaum, J., & Ullman, T. (2019). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 8985–8995, Vol. 32). Curran Associates, Inc.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605–632. https://doi.org/10.1037/0033-295X.99.4.605

Stetter, M., & Lang, E. W. (2021). Learning intuitive physics and one-shot imitation using state-action-prediction self-organizing maps. *Computational Intelligence and Neuroscience*, *2021*(5590445). https://doi.org/10.1155/2021/5590445

Stojnić, G., Gandhi, K., Yasuda, S., Lake, B. M., & Dillon, M. R. (2023). Commonsense psychology in human infants and machines. *Cognition*, *235*, 105406. https://doi.org/10.1016/j.cognition.2023.105406

Weihs, L., Yuile, A., Baillargeon, R., Fisher, C., Marcus, G., Mottaghi, R., & Kembhavi, A. (2022). Benchmarking progress to infant-level physical reasoning in AI. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=9NjqD9i48M

Wichmann, F. A., & Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, *9*(1), 501–524. https://doi.org/10.1146/annurev-vision-120522-031739