From Infants to AI: Incorporating Infant-like Learning in Models Boosts Efficiency and Generalization in Learning Social Prediction Tasks

Shify Treger (shify.treger@weizmann.ac.il)
Department of Computer Science and Applied Mathematics, Weizmann Institute of Science
Rehovot 7610001 Israel

Shimon Ullman (shimon.ullman@weizmann.ac.il)
Department of Computer Science and Applied Mathematics, Weizmann Institute of Science
Rehovot 7610001 Israel

Abstract

Early in development, infants learn a range of useful concepts, which can be challenging from a computational standpoint. This early learning comes together with an initial understanding of aspects of the meaning of concepts, e.g., their implications, causality, and using them to predict likely future events. All this is accomplished in many cases with little or no supervision, and from relatively few examples, compared with current network models. In learning about objects and human-object interactions, early acquired concepts are often used in the process of learning additional, more complex concepts. In the current work, we model how early-acquired concepts are used in the learning of subsequent concepts, and compare the results with standard deep network modeling. We focused in particular on the use of the concepts of animacy and goal attribution in learning to predict future events. We show that the use of early concepts in the learning of new concepts leads to better learning (higher accuracy) and more efficient learning (requiring less data). We further show that this integration of early and new concepts shapes the representation of the concepts acquired by the model. The results show that when the concepts were learned in a human-like manner, the emerging representation was more useful, as measured in terms of generalization to novel data and tasks. On a more general level, the results suggest that there are likely to be basic differences in the conceptual structures acquired by current network models compared to human learning.

Keywords: Computational Modeling; Infant Learning; Artificial Intelligence; Social Cognition; Machine Learning

Introduction

Already in early development, infants learn a broad range of useful concepts and visual tasks, which can be challenging from a computational standpoint. For example, in the domains of objects and human-object interactions (on which we focus), infants learn during the first year of life to recognize hands, their configuration, and their interactions with objects (Woodward, 1998; Gergely, Bekkering, & Király, 2002; Saxe, Tenenbaum, & Carey, 2005), a task where significant and meaningful features can be non-salient and highly variable and therefore difficult to learn.

Infants learn, in an unsupervised manner, to perform figure-ground segmentation, which took years and a large effort to develop (Kirillov et al., 2023). In the domain of dealing with other people, starting at 3-6 months of age, infants learn to detect and follow another person's gaze and establish joint attention, on the basis of head orientation, and later, eye direction (Scaife & Bruner, 1975; Flom, Lee, & Muir, 2017; D'Entremont, Hains, & Muir, 1997). This task is difficult, because 'gaze' is not a physical entity that appears explicitly in the image, and cues for gaze direction can be subtle and difficult to extract and use.

The learning of early concepts comes together with an initial understanding of aspects of the concepts' meaning, in terms of their implications, causality, or use for predicting likely future events. For example, in learning to identify hands, infants also learn that hands cause other objects to move and change location (Saxe et al., 2005; Király, Jovanovic, Prinz, Aschersleben, & Gergely, 2003). All this learning is obtained in many cases with little or no supervision, and from relatively few examples, compared with current models (Ullman, Dorfman, & Harari, 2019). In learning about objects and human-object interactions, early learned concepts are often used in acquiring additional related concepts. For example, the relation of in front/behind appears to be a prerequisite to the subsequent learning of containment (Ullman et al., 2019). Learning to identify the direction of gaze is used to predict future actions (Falck-Ytter, Gredebäck, & Von Hofsten, 2006), and it later plays a role in the development of communication and language (Tomasello, 2009).

In the current work, we model how early-acquired concepts are used in the learning of subsequent concepts. In particular, we focus on using the concepts of animacy and goal attribution in learning to predict future events. The main approach is to compare the learning of new concepts in two similar networks, where only one of the two uses a decomposition, similar to what was shown in infant studies, of learning a visual task by learning a simple concept first and then using it for learning the new task. For example, we compare the representation of the concept 'animate' and how it is used by the model, in the two different learning schemes: learning 'animate' first, and then using it to predict future behavior, compared to combined learning of animacy together with predicting future behavior.

We refer to the version that uses human-like decomposition of concepts as the 'cognitive' model, and the standard, end-to-end training as the 'naive' model. We compare the cognitive and standard approaches along two main directions: one is the performance of the trained models, and the other is a comparison of what was learned and represented in the two types of models. Briefly, in terms of model performance, the results show that the use of early concepts in the learning of new concepts leads to better learning (higher accuracy) and more efficient learning (requiring less data). In terms of the

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

learned concept representations, the results show that when the concepts are learned in a human-like manner, the emerging representation is more useful, as measured in terms of generalization to novel data and tasks. In the final discussion, we examine the possible source of the differences (decomposition and early concepts) and potential implications of the findings to the training of artificial models.

The main contributions of this work are as follows:

- Our study is the first (as far as we know) to test the potential computational advantage of a basic characteristic of infants' learning in terms of systematically using a range of early-acquired concepts in the learning of subsequent, more complex ones.
- We show that the infant-like learning model (in the sense above) reaches better performance on a test task in terms of accuracy as well as efficiency (size of the training set).
- We show evidence that the use of early-learned concepts in the learning of more complex ones plays a role in shaping the final representations produced during the learning process.

Related Work

Infants' Understanding of Animacy and Goal Attribution

An early and influential contribution to understanding infants' goal encoding is Woodward's work (Woodward, 1998), which demonstrated that infants attribute goals to agents or animate actors, but not to inanimate objects. This finding has been validated and elaborated through other experiments and variations (Woodward, 1999; Woodward & Sommerville, 2000; Biro, Verschoor, & Coenen, 2011). A study by Luo and Baillargeon (Luo & Baillargeon, 2005) extended the understanding of goal attribution by showing that infants can attribute goals to self-propelled objects. Further research illustrated that infants can generalize goal attribution to novel actions when provided with sufficient contextual cues (Király et al., 2003). Infants also exhibit a significant understanding of objects and their properties from an early age (Baillargeon, 1987; Spelke, Kestenbaum, Simons, & Wein, 1995; Hespos & Baillargeon, 2001), expecting objects to follow physical rules (Spelke, 2022; Lin, Stavans, & Baillargeon, 2022).

Related Computational Models of Infants' Intuitive Psychology

A recent study (Stojnić, Gandhi, Yasuda, Lake, & Dillon, 2023) investigated goal attribution in infants, both empirically and in models. The study found that infants anticipate agents' actions to be directed at objects rather than locations. In contrast, AI models often target locations, highlighting the need to integrate infants' understanding into models to better replicate human behavior. Li's work (W. Li, Yasuda, Dillon, & Lake, 2024) introduced additional tasks related to agents and objects, along with a self-supervised model. This work demonstrated limitations of current neural network models in goal attribution tasks.

Several works (Bortoletto, Shi, & Bulling, 2024; Hein & Diepold, 2023; Zhi-Xuan et al., 2022) have addressed infant-like tasks, using the Baby Intuitions Benchmark (BIB) (Gandhi, Stojnic, Lake, & Dillon, 2021), a dataset specifically designed for evaluating developmental tasks related to agents. These studies employed various approaches, such as transformer-based architectures and Bayesian models, to tackle these tasks.

The main goal of the studies above was to develop models that replicate infant behavior. In contrast, the current work focuses on the fundamental differences between human learning and standard AI models. Specifically, we focus on the learning process itself: while infants demonstrate an early understanding of core concepts and progressively build on them, most AI models rely on end-to-end training without explicit decomposition of learned concepts. By exploring this difference, we aim to highlight how integrating infant-like learning mechanisms could enhance the efficiency and generalization of AI systems.

Method

In this section, we describe our approach to evaluating learning about goals and action prediction. Inspired by the empirical methods of Woodward et al. (1998), we created a new dataset to compare two models: cognitive and naive.

The Cognitive Model is akin to infant-like learning, where early-acquired concepts are integrated into later learning stages. In contrast, **the Naive Model** serves as a baseline, that does not use such concepts decomposition. This comparison allows us to explore and highlight the differences between these two learning approaches.

Goal-Directed Dataset

We designed a dataset inspired by Woodward's experiments (Woodward, 1998), extending it to include tasks beyond those studied in prior research. The dataset consists of sequences of simple scenes (Figure 1), created using icons from 'flaticon.com'. Each frame contains three entities: two inanimate objects and an actor, which can be either animate (e.g., hands, animals, people) or inanimate (e.g., books, flowers).

The primary task in most experiments was to predict the future motion of the actor in the next step of the sequence (which is unseen). For animate actors, the prediction is determined by goals—meaning they are expected to move toward the same object they previously interacted with, even if its location has changed. In contrast, for non-animate actors, the prediction is determined by prior locations—meaning they are expected to return to their prior location, regardless of whether the object has changed. Figure 1 provides an example of the first task. This dataset allows us to test the ability of network models to predict goals and actions in scenes that involve both animate and inanimate entities.

The Two Models

We compared two models to evaluate the different types of learning. Both models follow the same two-step process:

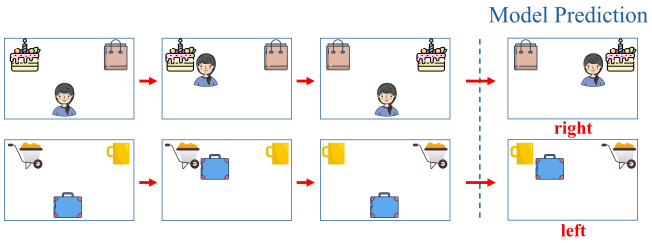

Figure 1: Input data for Experiment 1: The model receives a sequence of three images and predicts the location ('left' or 'right') of the actor in the next unseen step of the sequence. Top: animate actors; Bottom: inanimate actors. The person follows the object, the suitcase goes to the previous location.

(Figure 2) First, frames are processed to create representations of the scene, encoding the class and location of entities. Then, using these scene representations, the models predict the future location of the actor ('left' or 'right').

The key difference between the models lies in their treatment of conceptual information: the **Cognitive Model** incorporates additional concepts, e.g. distinguishing between animate and inanimate actors, into its representation, whereas the **Naive Model** relies solely on raw scene representations, excluding such conceptual information. Below, we elaborate on each of the steps in more detail.

First Step - Scene Representations The objective of the first step is to create scene representations for the image sequences. To achieve this, we fine-tuned the BLIP model (J. Li, Li, Xiong, & Hoi, 2022) to generate representations for each frame, encoding the class and location of the three entities (upper-left, upper-right, bottom). For example, the representation for the first frame in the top row of Figure 1 would be: 'cake', 'bag', 'girl'. The meaning of the words (e.g., 'cake') is not used by the model. Instead, the words only serve as placeholders to encode the entities' identities.

In the cognitive model, the actor type (animate/inanimate) is explicitly included during this step. For example, the cognitive model's representation for the previous scene would be: 'cake', 'bag', 'animate girl'.

Second Step - Prediction Generation The prediction task is framed as a classification problem using the BERT model (Devlin, 2018). The input to the model is the full sequence representation of the frames, and the model's output is a prediction of the actor's future location ('left' or 'right') in the next step of the sequence. The results in the following sections correspond to the model's accuracy in this prediction.

We used large models (BLIP, BERT) in our cognitive model to reflect the complexity of the infant brain—infants, too, are born with a sophisticated neural architecture. We fine-tuned all layers in both models for our specific tasks. To ensure that the models' pre-trained linguistic knowledge did not drive the observed effects, we also ran a version of the experiment using binary representations (e.g., 0/1 vectors) instead of natural language and obtained similar results. Furthermore, we confirmed that their prior knowledge did not bias the results: the models' initial performance on our tasks was not significantly different from chance level.

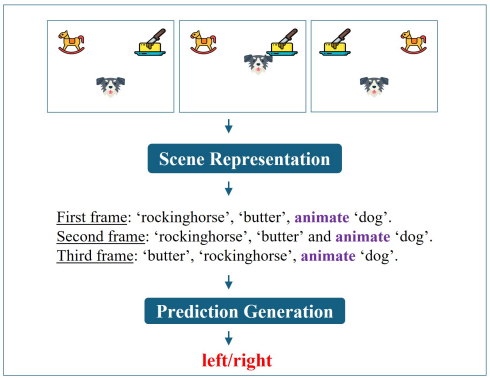

Figure 2: Two-step process of the cognitive and naive models: First, scene representations are created, and then predictions are made. Concepts specific only to the cognitive model are in bold and purple.

Experiments

Experiment 1: Prediction

Experiment 1 used the three-frame paradigm shown in Figure 1, where the task was to predict the actor's location (left vs. right) following three frames. We compared how the cognitive and naive models learned this task. The only difference between the models is the inclusion of the animate/non-animate label in the cognitive model.

The primary question was how the two models compare in learning the prediction task. Two important points regarding this comparison are: 1. The naive model could theoretically learn to distinguish between the two types of actors based on the data it observes, as the actor type is consistently correlated with its behavior in both training and test scenarios. 2. Note that the addition, such as the labels 'animate' and 'non-animate', is provided in a 'bare' form, without additional context about its implications or how it relates to the model's predictions. For the model, these labels are arbitrary markers (just as e.g. 'goo' vs. 'not goo').

Data The models were trained using sequences of three frames (Figure 1), featuring various combinations of icons, with half of the sequences showing animate actors and half showing non-animate actors. Two dataset sizes were used: a small dataset (320 training examples and 80 test examples) and a large dataset (1280 training, 320 test examples). The test data used the same actors as the training data but introduced novel target objects to evaluate generalization. Both the cognitive and naive networks were trained for 2000 epochs. Each experiment was repeated 15 times with ran-

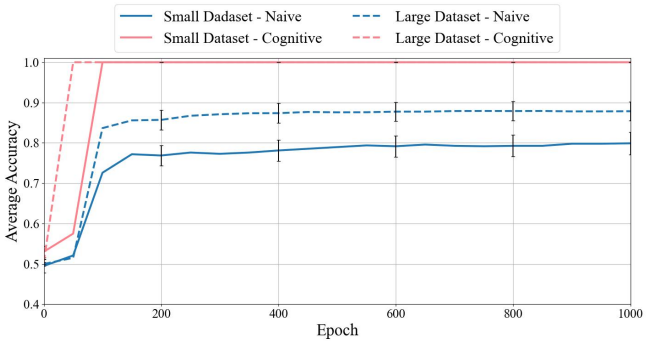

Figure 3: Experiment 1: Prediction Results. Average test accuracy of the naive and cognitive networks for both the small and large datasets, with the Standard Error of the Mean (SEM) included.

domly generated sequences, and the results were averaged.

Results The averaged results of Experiment 1 for both models are shown in Figure 3, comparing their performance on the small and large datasets. Only the first 1000 epochs are shown, as the results stabilized within this range. The cognitive network achieved perfect accuracy after 100 epochs with the smaller dataset and converged to perfect accuracy after 50 epochs with the larger dataset. In contrast, the naive network achieved approximately 80% accuracy after 1000 epochs with the smaller dataset and 87% accuracy with the larger dataset.

These results demonstrate that the cognitive network consistently achieves higher accuracy (perfect performance) and faster convergence, even with a small dataset.

Experiment 2: Generalization

Generalization to new tasks and domains is a crucial aspect and a useful measure of intelligent behavior (Lake & Baroni, 2018; Chollet, 2019). In the previous experiment, the test data used the same actors as in training but introduced new target objects. Here, we extended this to assess the networks' ability to generalize by evaluating their performance with entirely new actors.

To address generalization, a 5-frame paradigm was introduced (Figure 4). This paradigm first presents the actor's behavior in the initial three frames and then introduces the same actor with new target objects, requiring the model to predict the actor's behavior in the next step. Unlike Experiment 1, this setup includes information about the actor's behavior within the data itself, theoretically enabling both networks to predict behavior from a single example. As a result, it is well suited for testing generalization to new actors.

Data The 5-frame paradigm incorporates the second and third frames from Experiment 1 (Prediction task), along with the frame that features the next step of the sequence, which was not shown in Experiment 1. It extends this sequence by adding two new frames presenting the same actor with new target objects. The prediction task focuses on the actor's

Figure 4: Input Data for Experiments 2 and 3: The five rightmost frames are used in Experiment 2, while the full seven frames are used in Experiment 3.

behavior in relation to these new targets. The models were trained using 5-frame sequences featuring various combinations of icons, with half of the sequences containing animate actors and the other half containing non-animate actors.

Two task types were generated: Task 1 (T1): A simpler task where the test data use the same actors as in the training data but introduce new target objects. Task 2 (T2): A more challenging task because the test data include entirely new actors and new target objects.

The networks were first trained on T1 and then further fine-tuned under one of two conditions: T1-T1: The same simple task (T1) with additional data. This condition served as a control to assess the effect of additional data without introducing new challenges. T1-T2: The more challenging generalization task (T2). Each task used 640 training examples and 160 test examples. Both networks were trained for 1000 epochs. Each experiment was repeated 15 times with randomly generated sequences, and the results were averaged.

Results Figure 5 shows the performance of both models under the following conditions: 1. Learning the initial task (T1). 2. Retraining on the same task with additional data (T1-T1). 3. Generalization to a new task (T1-T2). Only the first 300 epochs are shown, as there was no significant change afterward. The cognitive network outperformed the naive network in both accuracy and learning speed across all tasks: T1-T1 Condition: The cognitive network achieved near-perfect accuracy in the first run and 100% accuracy almost immediately in the second run. The naive network stabilized at 56% accuracy in the first run and reached 75% accuracy in the second run. T1-T2 Condition: For the challenging generalization task, the cognitive network achieved perfect accuracy after just a few epochs. The naive network, however, reached only 65% accuracy, performing worse than in the T1-T1 condition.

We conclude that in addition to a difference in final performance, there is also a marked difference in the ability to generalize to a somewhat different, more challenging task.

Experiment 3: Decomposition

We address the following question: How is the additional information provided to the cognitive model learned? In earlier experiments, the actor's type (animate/inanimate) was explicitly provided to the cognitive network. In the current experiment, however, the same information is available to both models. Inspired by evidence that self-propelled motion is a strong indicator of animacy in infants (Luo & Baillargeon, 2005), the experiment introduced additional frames to enable the learning of this concept.

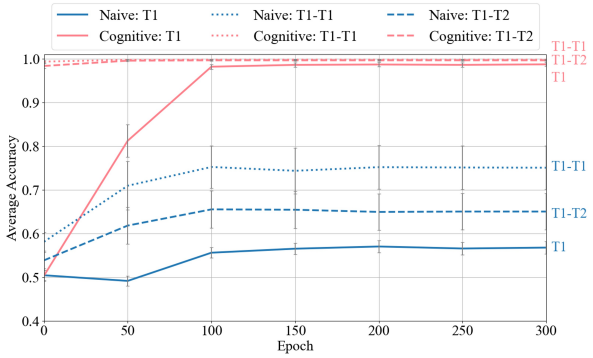

Figure 5: Experiment 2: Generalization Results. Average test accuracy of the naive and cognitive networks on Task 1 (T1), retraining on Task 1 (T1-T1), and retraining on Task 2 (T1-T2), with SEM included.

The cognitive network emulates infant learning by first acquiring simple concepts (e.g., animacy) and then integrating this knowledge into downstream tasks. In contrast, the naive network processes the entire dataset directly, without explicitly decomposing the task.

Data The experiment builds on the previous 5-frame paradigm (Figure 4) by adding two additional frames at the start of each sequence. These new frames depict the actor's movement: If the actor changes location between the first and second frames, it is classified as self-propelled (indicating animacy). If no change in position is observed, the actor is classified as non-self-propelled (indicating inanimacy). The models were trained using 7-frame sequences with various icon combinations, half involving animate actors and half non-animate.

For the cognitive model, training first focuses on learning the concept of animacy (animate vs. inanimate), which is associated with the actor and subsequently used in the five-frame task. In contrast, the naive network receives all seven frames concatenated and learns the prediction task directly, without explicit concept decomposition.

The naive network was tested on the 7-frame paradigm with progressively larger datasets, starting from 640 training and 160 test examples, doubling in size at each step. The cognitive network, in contrast, learned actor types using smaller datasets of 320 training and 80 test examples. For the 5-frame task, a dataset of 640 training and 160 test examples was used. Each experiment was repeated 15 times with randomly generated sequences, and the results were averaged. Both networks were trained for 300 epochs due to the larger datasets.

Results The cognitive network learned the first task of classifying the actor's type within 20 epochs, achieving perfect accuracy (Figure 6). For the subsequent task, the cognitive network performs the 5-frame paradigm discussed earlier, so the results for this stage are identical to those of the 5-frame test. The performance of the naive network is shown in Figure 7, alongside the cognitive network's results from the 5-frame

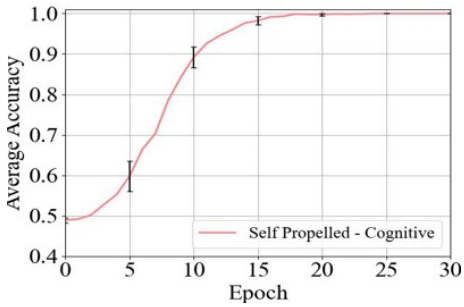

Figure 6: The cognitive network's performance in the preceding step, where it learned to distinguish between animate and inanimate entities based on self-propulsion.

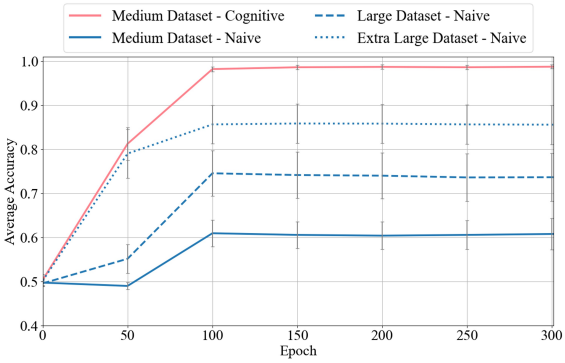

Figure 7: Experiment 3: Decomposition Results. Average test accuracy of the naive network trained with varying dataset sizes, compared to the cognitive model's performance on the 5-frame test with medium dataset size, with SEM included.

stage. For the smallest dataset, the naive network achieved approximately 60% accuracy. When the dataset size was doubled, its accuracy increased to 74%. With the largest dataset, accuracy improved to 85%. In comparison, the cognitive network achieved near-perfect accuracy on the same 5-frame test even with the smallest dataset.

The results show that the decomposition of the task into subtasks, performed by the cognitive network, leads to substantial improvement in performance.

Experiment 4: Goal Attribution

In this experiment, a related yet distinct concept, 'Goal,' is investigated. Similar to the idea of actor type, it is assumed, based on the empirical literature, that infants can attribute goals to animate entities, based on cues such as certain types of contact or gaze direction (Woodward, 1998; Phillips, Wellman, & Spelke, 2002). Similar to the 'animacy' concept, the concept of 'having a goal' is provided to the cognitive model, and the results are compared with a naive model that can infer the goal, but is not directly provided with the concept.

Data The Goal Attribution task uses a one-frame paradigm, where each frame depicts an actor with two target objects,

similar to the first frame of the three-frame paradigm introduced earlier. Unlike previous experiments, all actors in this task are animate, as goal attribution applies only to animate entities. Each training dataset includes 10 repetitions of an actor appearing with the same goal object, but with varying distractor objects across frames. The data demonstrates that the actor has a global goal, consistently preferred over all distractors. During testing, the actors and their associated goals remain the same as in training, but the distractors are new.

For the cognitive model, the goal is provided as part of the actor's representation (e.g.: 'girl with goal cake'), but its implications are not specified and must be learned during training. In contrast, the naive model must infer the goal directly from the data. The only difference between the models is the inclusion of the goal representation in the cognitive model. The training size was 320, and the testing size was 80. Both models were also evaluated on their ability to generalize to a more challenging task (T2), involving new actors and new goals. The experiment ran for 1000 epochs, but only results from the first 100 epochs are presented, as there was no significant change beyond that point. Each experiment was repeated 15 times with randomly generated sequences, and the results were averaged.

Results The results for the Goal Attribution task are shown in Figure 8. The cognitive network achieved perfect accuracy after around 20 epochs on the first task (T1) and maintained perfect accuracy almost immediately in the generalization task (T2). The naive network, in contrast, reached approximately 82% accuracy for both tasks after 100 epochs. Although the learning rate was faster for T2 than T1, the accuracy of the naive network did not improve significantly.

Generalization behavior is assessed by examining the transition from T1 to T2. In the cognitive model, there is full transfer: after learning T1, the accuracy of T2 (dashed curve) starts at the same level that T1 reached at the end of training. In contrast, the naive model shows limited transfer. The accuracy of T2 begins at chance level, similar to T1, but exhibits an accelerated learning rate, indicating partial transfer of knowledge from the first task to the second. These results demonstrate that the cognitive model learns faster, achieves higher accuracy, and transfers knowledge seamlessly to the generalization task (T2). In contrast, the naive model learns less efficiently and struggles with knowledge transfer.

Discussion

The current study examined a basic characteristic of early human learning, by modeling some aspects of how early-acquired concepts are used in the learning of subsequent concepts and comparing the results with standard deep network modeling. The results show that the use of early concepts in the learning of new concepts can lead to higher accuracy and more efficient learning. The results of comparing generalization to new tasks also suggest that the integration of early and new concepts during the learning process can shape the representation of the concepts acquired by the model. Although

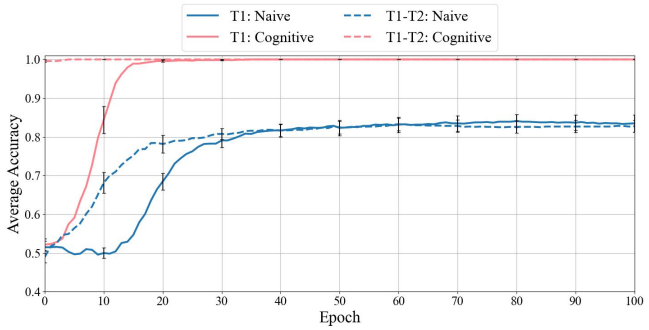

Figure 8: Experiment 4: Goal Attribution Results. Average test accuracy of the cognitive and naive networks on the first and second tasks, with SEM included.

the dataset is simplified compared to natural images, prior work has shown that infants succeed in tasks with similarly simplified stimuli (Stojnić et al., 2023).

What can cause the 'cognitive' models to learn better than the 'naïve' ones? There is empirical and theoretical evidence suggesting that the difficulty of learning a function F by a deep network model depends on the complexity of F: the learning will require more data, and the probability of finding F, or a close approximation, starting from a random initialization, will decrease (Valle-Perez, Camargo, & Louis, 2018). If this notion is correct, then learning by the cognitive network, where the learning process is divided into components, may help the learning by decomposing the overall task into components, where the complexity of each component is lower than the complexity of the full task. The general suggestion raised by this possibility is that the training of models could benefit from some form of hierarchical decomposition, where simple concepts are learned first and can be used in the learning of more complex concepts.

Current large models are trained on huge data sets, and they strive to create so-called foundational models, which can deal with a broad range of tasks. Such models have achieved impressive results. However, multiple studies have shown examples where the learning by current large models appears to be different from human leaning; they 'show failures on surprisingly trivial problems' (Dziri et al., 2024), or 'show basic errors in understanding that would not be expected even in non-expert humans' (Oh, Kim, Cha, & Oh, 2024). It is possible that shortcomings of this kind will be overcome by additional data and training. Alternatively, it may prove useful to combine current large models and human learning in some manner. For example, it may be possible to learn early concepts by adapting methods similar to self-supervised learning of early concepts in humans, and use early-learned concepts in the learning of more complex ones. If such training proves feasible, it would be of interest to test the effects of making the model more human-like both on its performance, as well as its similarity to human behavior in different visual tasks.

Acknowledgments

ST acknowledges support from the Ariane de Rothschild Women Doctoral Program. SU acknowledges the Weizmann Institute of Science for its support.

References

Baillargeon, R. (1987). Object permanence in 31/2-and 41/2-month-old infants. *Developmental psychology*, *23*(5), 655.

Biro, S., Verschoor, S., & Coenen, L. (2011). Evidence for a unitary goal concept in 12-month-old infants. *Developmental science*, *14*(6), 1255–1260.

Bortoletto, M., Shi, L., & Bulling, A. (2024). Neural reasoning about agents' goals, preferences, and actions. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 38, pp. 456–464).

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

D'Entremont, B., Hains, S. M., & Muir, D. W. (1997). A demonstration of gaze following in 3-to 6-month-olds. *Infant Behavior and Development*, *20*(4), 569–572.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., ... others (2024). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, *36*.

Falck-Ytter, T., Gredebäck, G., & Von Hofsten, C. (2006). Infants predict other people's action goals. *Nature neuroscience*, *9*(7), 878–879.

Flom, R., Lee, K., & Muir, D. (2017). *Gaze-following: Its development and significance*. Psychology Press.

Gandhi, K., Stojnic, G., Lake, B. M., & Dillon, M. R. (2021). Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *Advances in neural information processing systems*, *34*, 9963–9976.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*(6873), 755–755.

Hein, A., & Diepold, K. (2023). Comparing intuitions about agents' goals, preferences and actions in human infants and video transformers. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Hespos, S. J., & Baillargeon, R. (2001). Reasoning about containment events in very young infants. *Cognition*, *78*(3), 207–245.

Király, I., Jovanovic, B., Prinz, W., Aschersleben, G., & Gergely, G. (2003). The early origins of goal attribution in infancy. *Consciousness and cognition*, *12*(4), 752–769.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... others (2023). Segment anything. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 4015–4026).

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning* (pp. 2873–2882).

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888–12900).

Li, W., Yasuda, S. C., Dillon, M. R., & Lake, B. (2024). An infant-cognition inspired machine benchmark for identifying agency, affiliation, belief, and intention. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Lin, Y., Stavans, M., & Baillargeon, R. (2022). Infants' physical reasoning and the cognitive architecture that supports it. *Cambridge handbook of cognitive development*, 168–194.

Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? psychological reasoning in 5-month-old infants. *Psychological science*, *16*(8), 601–608.

Oh, J., Kim, E., Cha, I., & Oh, A. (2024). The generative ai paradox on evaluation: What it can solve, it may not evaluate. *arXiv preprint arXiv:2402.06204*.

Phillips, A. T., Wellman, H. M., & Spelke, E. S. (2002). Infants' ability to connect gaze and emotional expression to intentional action. *Cognition*, *85*(1), 53–78.

Saxe, R., Tenenbaum, J., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10-and 12-month-old infants. *Psychological science*, *16*(12), 995–1001.

Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature*, *253*(5489), 265–266.

Spelke, E. S. (2022). *What babies know: Core knowledge and composition volume 1* (Vol. 1). Oxford University Press.

Spelke, E. S., Kestenbaum, R., Simons, D. J., & Wein, D. (1995). Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British journal of developmental psychology*, *13*(2), 113–142.

Stojnić, G., Gandhi, K., Yasuda, S., Lake, B. M., & Dillon, M. R. (2023). Commonsense psychology in human infants and machines. *Cognition*, *235*, 105406.

Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard university press.

Ullman, S., Dorfman, N., & Harari, D. (2019). A model for discovering 'containment'relations. *Cognition*, *183*, 67–81.

Valle-Perez, G., Camargo, C. Q., & Louis, A. A. (2018). Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant behavior and development*, *22*(2), 145–160.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science*, *11*(1), 73–77.

Zhi-Xuan, T., Gothoskar, N., Pollok, F., Gutfreund, D., Tenenbaum, J. B., & Mansinghka, V. K. (2022). Solving the baby intuitions benchmark with a hierarchically bayesian theory of mind. *arXiv preprint arXiv:2208.02914*.