A Predictive Coding Account for Cognition in Human Children and Chimpanzees: A Case Study of Drawing

Anja Philippsen and Yukie Nagai, Member, IEEE

Abstract—Humans and chimpanzees differ in their cognitive abilities, in particular, in social-cognitive processing; however, the underlying neural mechanisms are still unknown. Based on the theory of predictive coding, we hypothesize that crucial differences in cognitive processing might arise from aberrant reliance on predictions. We test this hypothesis using a recurrent neural network that integrates sensory information with predictions based on the rules of Bayesian inference. Altering a network parameter, we vary how strongly the network relies on its predictions during development. Our model qualitatively replicates findings from a behavioral study on the drawing ability of human children and chimpanzees. Moderate parameter values replicate the ability of human children to complete drawings by adding missing elements. With weak reliance on predictions, the model’s behavior is similar to chimpanzees’ behaviors: trained networks can follow existing lines but fail to complete drawings. Furthermore, with a strong reliance on predictions, networks learn more abstract representations of drawings and confuse different trained patterns. An analysis of the internal network representations reveals that an aberrant reliance on predictions affects the formation of attractors in the network. Thus, appropriate reliance on their own predictions in humans may be crucial for developing abstract representations and acquiring cognitive skills.

Index Terms—Chimpanzees, computational modeling, predictive coding, recurrent neural network (RNN), representational drawing.

I. INTRODUCTION

Hildren acquire the ability to make predictions in complex situational contexts rapidly during their first years of life. From a young age, they are skilled in understanding goal-directed behaviors and predicting the intentions of interaction partners. Nonhuman primates also possess this ability to a degree. In particular, our closest living relatives, chimpanzees, show remarkable performance in many cognitive tasks [1], [2]. In a comparison of the performance of 2.5-year-old human children and chimpanzees using a large number of tests, Herrmann et al. [3] found that both groups exhibited a similar level of intelligence in dealing with the physical world. However, human infants outperformed chimpanzees mostly in tasks that required the usage of communicative cues, gaze following, or reasoning about intentions. Thus, the two species seem to differ crucially in terms of their ability to understand intentions, in particular, in the context of social communication. This difference is potentially caused by a reduced tendency to represent abstract concepts [3]–[5] or a reduced ability to imagine [6]. However, it is currently not known which concrete differences in neural mechanisms could explain such findings.

Computational modeling may help to explore potential mechanisms and their consequences. Being commonly applied to human cognition [7], [8], there have been limited studies involving the computational modeling of cognition of non-human primates. One attempt to compare the behaviors of human children and chimpanzees in a computational model was performed by Lopes et al. [9]. Their model focused on learning by observation. Differences in the degree of social imitation were implemented via tradeoffs between beliefs of the world’s states and perceived actions, which led to either imitation or emulation. Their model explained how behavior arises from switching between predefined mechanisms. In contrast, in this study, we investigate how neural mechanisms can account for the emergence of behavior, in particular, of different types of behavior, such as scribbling behavior versus picture completion.

Our proposed model is based on the principles of predictive coding [10], [11], which is commonly suggested as a unified theory of the brain. The theory states that our brain might be working as a prediction machine. By constantly observing the world, the brain learns internal models [12] of how actions can result in consequences in the world, finally becoming able to make accurate predictions of future sensory signals and the consequences of our and others’ actions. We hypothesize that the degree to which such predictions are utilized during information processing might cause differences in the behaviors of human children and chimpanzees. In particular, we aim to replicate qualitative differences in the drawing abilities of human children and chimpanzees. For this purpose, we first train recurrent neural networks (RNNs) with a number of drawings. During the training of the networks, we subtly alter how sensory information is integrated with predictions of the internal model at the level of network input. Then,
we investigate how this parameter affects the network’s later ability to complete partial drawings of the trained stimuli.

We compare qualitatively whether the results are similar to drawings of human children and chimpanzees from a study by Saito et al. [6], in which spontaneous drawing behavior on an incomplete picture was observed. The striking difference found in their study was that children from the age of two to three years tended to complete the drawing by adding missing features (such as the eyes in the drawing of a face). Chimpanzees, in contrast, despite sufficient motor capabilities in drawing, only followed existing lines or scribbled randomly. In this article, we replicate this task as an exemplary study, which demonstrates a clear difference between the species and can be well emulated in a computational study. Our model, however, is not limited to this task, but may also be used to explain various other differences in cognition (for a discussion, see Section VII).

A preliminary investigation of this hypothesis was presented in our previous study [13]. Here, we confirm the findings by extending the range of parameter conditions and using a larger set of drawings and trained networks. Furthermore, unlike in our previous work, we adjust the network parameters already during the training time, allowing for an analysis of the internal representations acquired in the trained neural network. This analysis provides insights about the networks’ underlying mechanisms, which may provide a more general view on cognitive abilities between human children and chimpanzees.

II. BACKGROUND

First, we discuss general findings on drawing abilities in human children and chimpanzees, and then we introduce the predictive coding idea that is utilized in our model.

A. Drawing in Human Beings and Chimpanzees

Drawing has been commonly used as a tool to gain insight into children’s cognition and, for example, to better understand how children perceive the world and form abstract concepts [14], [15]. Representational drawing, which is creating drawings that represent an object or person, is one behavior that can be used to assess the human ability to build abstract concepts. This behavior emerges in children around the age of two to three years [6]. Furthermore, before acquiring the ability to produce representational drawings themselves, children are also able to recognize them [16].

Similar to human children, chimpanzees also spontaneously engage in drawing when provided with the opportunity [17], [18]. They mostly show scribbling behavior but can also be taught to draw simple shapes [19]. In the study from Saito et al. [6], all participating chimpanzees (two juveniles aged approximately 6 years, and four adults aged approximately 27 years) had some previous drawing experience. A test before the actual experiment revealed that the adult chimpanzees were able to trace existing lines, indicating that they exhibited sufficient motor capabilities for representational drawing. Then, chimpanzees as well as human children from ages 1.5 to 3 years were tested for their representational drawing abilities using the same experimental setup. The drawing stimuli that were presented to both participant groups were drawings of a chimpanzee’s face where some facial features were missing, such as the eyes or mouth. Chimpanzees and human children of all age groups showed scribbling behavior, marked the face or parts of it, or sometimes traced the drawings’ outlines. However, older children showed a higher percentage of completions: despite immature motor skills, the children tended to complete the drawing by adding missing parts. This behavior was only found in human children; none of the chimpanzees spontaneously completed the drawing.

To explain the findings of their study, Saito and colleagues suggested that chimpanzees might miss certain cognitive abilities, in particular, the ability to imagine. For imagination, the ability to learn concepts and to apply these concepts to the real world is crucial. Thus, drawing ability closely relates to general cognitive capabilities, such as abstraction and categorization. In this study, we use a computational model of representational drawing to reproduce cognitive abilities in human children and chimpanzees. With this model, we can directly evaluate how changes in the reliance on own predictions and on sensory input affect behavior. Hereby, we use drawing ability as a metric to exemplify differences in behavior. In Section VII, we discuss how these findings might extend to other types of cognitive tasks.

B. Predictive Coding and Bayesian Inference

The ability to make predictions is crucial for cognitive development and is often considered as the foundation for various cognitive processes, such as perception and motor learning [10], [11]. To make predictions, we use sensory information as well as the so-called prior information derived from previous experiences. These priors are commonly believed to be generated by internal models in the brain [12], which represent the consequences of our and others’ actions.

Previously, it has been shown that predictive coding can account for many human cognitive abilities [20]. Notably, it may also offer potential explanations of phenomena that were previously considered to be purely socially motivated [21]. However, to live up to the claim of being a unified brain theory, predictive coding also needs to account for impairment or reduced expression of cognitive abilities. In recent years, several conceptual ideas [22]–[24], as well as computational models [8], [25], [26], have focused on aberrant cognitive abilities in developmental disorders. In fact, cognitive abilities including drawing seem to be affected in children with autism spectrum disorder [15], [27]. In this article, we suggest that predictive coding may also be used to explain the cognition of other primate species such as chimpanzees. Specifically, we investigate how the degree of reliance on own predictions might account for differences in the behaviors of chimpanzees and young children.

Our model design is inspired by the hypo-prior hypothesis, originally suggested by Pellicano and Burr [22] for explaining altered perception in subjects with autism spectrum disorder. Their idea was that a weak influence of prior predictions on information processing, a so-called hypo-prior, shifts the focus to sensory information. Here, we investigate how a hypo-prior
or a hyper-prior (stronger influence of priors) affect the ability of the model to complete representational drawings in a task similar to the behavioral study from Saito et al. [6].

In the implementation of our model, we follow the Bayesian brain account [10], [28], suggesting that perception is the result of integrating sensory information with prior predictions. An RNN is used as the internal model of the world which learns to predict sensory input. Thus, the internal model represents the model’s belief about the world. In this study, we extend the concept of previous models [10], [28] by adding a Bayesian inference module that explicitly integrates sensory input with prior predictions at the input level of the network. Consequently, the term prior in our article refers to the prediction made by the internal model instead of the prior that is implicitly encoded in the trained network (see [10] and [28]). This prior prediction is integrated with sensory information, and the integrated (posterior) signal is used as input to the internal model. Thus, the internal model makes its prediction for the current time step based on actual sensory input and based on its prediction that was generated in the previous time step.

This explicit integration at the input level enables us to gradually modify the reliance of models on prior information. In particular, we introduce a network parameter that changes how sensory input is integrated with prior predictions of the internal model during the training process to model differences in cognitive processing, which might exist between different species but also between different individuals. How strongly sensory information and predictions influence the posterior perception depends on the precision of these two signals, which is given by the inverse of the variance values. Signals with higher precision (smaller variance) are considered more strongly than signals with lower precision. For instance, in the depiction of the Bayesian inference module in Fig. 1, sensory information is slightly more precise than the prior prediction, causing the posterior perception to be closer to the sensory input.

The key idea of our model is to artificially modify the variance of the prior prediction to make the network rely more strongly (hyper-prior) or weakly (hypo-prior) on its own predictions.

### III. Computational Model for Drawing

The computational model that we use in the present study was first proposed in [29]. It combines a time-series predictor as an internal model (here, an RNN) with a Bayesian inference module. An overview of the model is presented in Fig. 1. Bayesian inference is used to combine the model’s own predictions, provided by the internal model, with sensory input. As the internal model, we use a stochastic continuous-time RNN (S-CTRNN) [30], a type of RNN that learns to estimate not only the mean of the next time step but also a variance value that indicates how certain the network is about this prediction. This value is a prediction of the amount of variance present in the sensory signal. However, it also reflects the uncertainty of the prediction, relative to the uncertainty in sensory input. In the beginning of the learning process, the network’s estimated variance is high as the network’s predictions are still imprecise and highly variable [see Fig. 3(b)]. The Bayesian inference module thus integrates the signals in a way such that the combined signal mainly consists of the sensory input, which exhibits relatively higher precision. During the course of learning, the network’s predictions become more accurate, leading to a decrease of the predicted variance and a stronger reliance of the network on its predictions.

#### A. Integrating Sensory Input and Predictions

The crucial operation of our model is performed in the Bayesian inference module, which is shown on the left-hand side of Fig. 1. This module integrates raw sensory input with the model’s predictions. The network’s prediction is represented in the form of a Gaussian distribution with mean ($\mu_{\text{net}}$) and variance ($\sigma_{\text{net}}^2$). As this output of the internal model corresponds to the network’s prediction of events in the next time step, we interpret $\mu_{\text{net}}$ as prior information. Sensory input ($x$) corresponds to the observation. Here, we assume that sensory input is not perceived directly by the network. Instead, the network’s perception is biased by its own prediction for this time step. Therefore, the Bayesian inference module integrates sensory input with the network’s prediction $\mu_{\text{pred}} = \mu_{\text{net}}$ according to the variance of these two signals.

The variance of sensory perception ($\sigma_{\text{sensor}}^2$) is not directly observable. We assume here a constant variance of the input signal, which we interpret as the model’s belief about how accurate the perception is, given the environmental conditions.

![Fig. 1. Overview of the computational model. A Bayesian inference model integrates sensory information and its own predictions to a posterior signal that generates the input to the predictor, which is an RNN that learns to predict the mean and variance of the incoming signal.](image-url)
In general, this value can be considered to be low when good training data with minimal noise are available, and training succeeds regardless of the exact choice of $\sigma^2_{\text{sensor}}$ (see [29] for details). Here, we use a small value of around 0.05, computed as the maximum amplitude of variability that is present in the training data. To estimate the variance of the prediction, we make use of the network’s variance output $\sigma^2_{\text{net}}$ that predicts variance in the sensory signal. This variance value is high at the beginning of learning and converges during learning to the actual sensory variance $\sigma^2_{\text{sensor}}$ [see Fig. 3(b)]. A high value of $\sigma^2_{\text{net}}$ thus indicates high variance in the prediction, and a low value close to the actual sensory variance $\sigma^2_{\text{sensor}}$ indicates high precision of the prediction. For usage in the Bayesian inference, we multiply the estimated variance with the prior parameter $H$

$$\sigma^2_{\text{pred}} = H \cdot \sigma^2_{\text{net}}. \quad (1)$$

As explained in Section III-C, we use $H$ to manipulate the network’s tendency to rely strongly or weakly on predictions.

The Bayesian inference module generates the posterior perception, namely, $\mu_{\text{perc}}$ and $\sigma^2_{\text{perc}}$, by integrating the input signal with its own predictions via the following formulas (adapted from [31]):

$$\sigma^2_{\text{perc}} = \frac{\sigma^2_{\text{sensor}} \cdot \sigma^2_{\text{pred}}}{\sigma^2_{\text{sensor}} + \sigma^2_{\text{pred}}} \quad (2)$$

$$\mu_{\text{perc}} = \sigma^2_{\text{perc}} \cdot \frac{\mu_{\text{pred}}}{\sigma^2_{\text{pred}}} + \frac{\chi}{\sigma^2_{\text{sensor}}}. \quad (3)$$

These equations are applied independently to each input dimension. From this posterior distribution, the network’s actual input, $x_{\text{perc}}$, is drawn. Note that in this way, the network never receives the raw sensory input, but always an interpreted form of the input that was combined with its own prior experiences.

B. Training the Internal Model

The S-CTRNN is trained using backpropagation through time [30]. Given an input $x^t$, the network estimates $\mu_{\text{net}}$ and $\sigma^2_{\text{net}}$. Then, the likelihood $L$ is computed such that the network’s estimated mean and variance account for the observed input

$$-\ln(L) = \sum_{t=1}^{T} \sum_{i=1}^{D} \left( \ln \left( 2\pi \sigma^2_{\text{net}}^{t,i} \right) + \frac{(x^{t,i} - \mu_{\text{net}}^{t,i})^2}{2\sigma^2_{\text{net}}^{t,i}} \right). \quad (4)$$

Here, $T$ is the total number of time steps and $D$ is the dimensionality of the input vector. The covariance matrix, thus, is diagonal, and each input dimension has a different variance value. This likelihood was maximized by updating the network’s parameters, which is equivalent to minimizing the prediction error between the target signal and the network’s generated output [32]. The estimated variance $\sigma^2_{\text{net}}$ in this formula scales the prediction error such that parts of the input with high variance lead to smaller prediction errors. This avoids an overlearning of noise and makes the learning process more stable [30].

Network parameters are the weights from the input to the recurrent layer, recurrent weights, and output weights. Additionally, the neural activations of the first time step are updated during training. These activations $u_{0}^{(s)}$ are known as initial states and are required so that a single RNN can distinguish different categories of input sensory signals from each other. An appropriate initial state for each category is automatically inferred during learning (starting from zero vectors for all initial states) using the above likelihood function. Additionally, as proposed by Murata et al. [30] and [32], a second likelihood term ensures that the initial states do not diverge forever but maintain a predefined distance $\delta_{\text{dist}} = 10$ from each other

$$-\ln(L_{\text{init}}) = \sum_{s=1}^{S} \sum_{i=1}^{C} \left( \ln(2\pi \delta_{\text{dist}}) + \frac{(u_{0}^{(s)} - \tilde{u}_{0,i})^2}{2\delta_{\text{dist}}} \right). \quad (5)$$

Here, $S$ is the number of different training patterns that the network should distinguish [in the current experiment, the number of categories is $S = 6$ (see Fig. 2)] and $C$ is the number of neurons in the RNN (here, $C = 100$). With this equation, the average distance of the initial states $u_{0}^{(s)}$ to the mean initial state $\tilde{u}_{0}$ is kept at a variance level of around $\delta_{\text{dist}}$.

Training proceeds in epochs. In each epoch, all training trajectories are presented to the network and trainable parameters of the network are updated.

C. Modeling Aberrant Cognitive Processing

To model networks with a weaker or stronger reliance on prior information than usual, we alter the parameter $H$ in Fig. 1, which is a factor that adjusts the network’s estimated variance (1). With a value of $H = 1$, the network learns normally, relying on its own predictions according to the estimated variance of this prediction. By adjusting $H$ to smaller or larger values, we can emulate a hyper-prior or hypo-prior of the network during training. With a hyper-prior ($0 < H < 1$), the network relies more strongly than usual on its own predictions, and with a hypo-prior ($H > 1$), the network relies more strongly on sensory input. In total, seven different values of this parameter were tested in this study: $H \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. Each network used the same $H$ value during and after training when tested with the completion task.

IV. EXPERIMENTAL PROCEDURE

The experiments are performed in two steps. First, networks are trained in different experimental conditions using a set of trajectories representing hand-drawn pictures [see Fig. 2(a)]. For each of the seven conditions of $H$ (see Section III-C), ten networks with different initial connecting weights are trained. Second, the trained networks are asked to complete drawings, given only the first 30 time steps (one third) of a testing trajectory [see Fig. 2(b)]. This second part of the experiment is designed analogously to the study from Saito et al. [6].

A. Training Data

The set of training data is an extended version of the data used in our previous study [13]. In particular, we have added
three new shapes such that the full data set consists of two
groups of three drawings that share similarities (see Fig. 2).
Namely, the drawings of FACE, HOUSE, and CAR start in a
similar drawing direction and roughly cover the same region of
the drawing plane. (Analogously for FLOWER, HUMAN, and
ROCKET.) In total, for each pattern, ten drawings were created
by the experimenter beforehand using a touchscreen. Seven of
these drawings are used for training and the other three for
testing. Each trajectory is resampled such that it consists of
90 time steps.

A drawing is represented as a time series of \((x_0, x_1, x_2)\) coordi-
nates, where \(x_0\) and \(x_1\) indicate the location on the drawing
plane, and \(x_2\) is a binary value switching between “drawing”
and “nondrawing.” \(x_2 = 1\) represents drawing (points \([x_0', x_2']\)
and \([x_0^{-1}, x_2^{-1}]\) are connected with a line) and \(x_2 = 0\)
indicates that the “pen” is moved without drawing a line. The continu-
ous values outputted by the network for \(x_2\) were rounded such
that \(x_2 < 0.5\) equaled 0 and \(x_2 \geq 0.5\) equaled 1. Values of \(x_0\)
and \(x_1\) lie within \([-1, 1]\).

B. Training Procedure

Networks with 100 recurrent units are trained using the
training data. A small Gaussian noise (scaled by 0.01) is added
in each epoch to avoid overfitting. In each training epoch,
all \(7 \times 6\) training trajectories are presented to the network,
and network parameters are updated using backpropagation as
detailed in Section III-B. Training success can be influenced by
the initial weights of the RNN. To make sure that the results
in different conditions of \(H\) are not affected by differences
in weight initialization, the random initial weights are shared
across different parameter conditions.

Training continued for a maximum of 30 000 epochs. After
every 100 epochs, it is evaluated how well the network gener-
at the target patterns, using only the currently trained initial
network states without sensory input. Training is stopped
earlier than after 30 000 epochs if this evaluation error did
not improve within the previous 5000 epochs. The network
from the best performing epoch is used for the subsequent
completion task.

C. Completion Task

In this study, we are interested in how well networks are
able to complete partial drawings of the stimuli. Therefore,
after training, we test the networks’ completion abilities by
providing only the first third of the three testing trajectories,
as shown in Fig. 2(b). The completion is performed in two
steps as follows.

1) Recognition: The network tries to recognize the shape,
that is, to infer, given the first part of the trajectory,
which of the trained patterns is most likely the cause
for the observation.

2) Generation: The network uses the initial state corre-
sponding to the recognized pattern and tries to generate
the trajectory. During this generation process, sensory
input is available for the first third. Afterward, no input
is available; therefore, the network has to complete the
trajectory freely.

Recognition is implemented as inferring the initial states
similar to the training process, with the exception that network
weights are not adapted \([32]\). Given the first third of the tra-
cjectory, backpropagation through time is performed to gradually
adapt initial states such that they better reproduce the tra-
cjectory. We perform 2000 backpropagation steps (= epochs).
The inferred initial state is then used for the generation step.
The advantage of this recognition procedure is that it proceeds
analogously to the training and follows the predictive coding
idea \([32]\). During the generation of the first third of the tra-
cjectory, the input is available to the network like during training
(with accurate perception with \(\sigma^2_{\text{sens}} \approx 0.05\)). From 30 time
steps onward, all values of the \((x_0, x_1, x_2)\) trajectories are set
to 0, and the variance of the sensory input \(\sigma^2_{\text{sens}}\) is set to infin-
ity to emulate imprecise sensory perception. Accordingly, (2)
and (3) reduce to \(\sigma^2_{\text{perc}} = \sigma^2_{\text{pred}}\) and \(\mu_{\text{perc}} = \mu_{\text{pred}}\),
respectively.

For each of the ten trained networks per prior condition,
three individual completions (in the following, referred to as
trials) of all six patterns are performed, each using one set of
testing trajectories from Fig. 2(b).

V. RESULTS I: BEHAVIOR

We first present the results of the training (Section V-A).
Then, we show the behavior of networks for the completion
task qualitatively (Section V-B) and by quantifying the results
(Section V-C).

A. Training Results

Fig. 3 displays how the prediction error [Fig. 3(a)] and
estimated variance [Fig. 3(b)] decreased during the learning
from the first epoch to the best performing epoch. The average

![Fig. 2](image-url)
across all ten networks trained for each parameter condition is shown.

The left plot displays the networks’ ability to predict the next time step of the trajectory, given the previous time step. The error of all networks successfully converged. The higher error value for hyper-prior conditions (0 < $H < 1$) can be explained by the relatively lower influence of sensory input on training. As the network relies more strongly on its own predictions, less information from the training data is utilized. Therefore, training requires more time, and trained trajectories tend to be less accurate. The right plot, analogously, shows the estimated variance of the networks $\sigma_{\text{net}}^2$, which converged to values close to the actual sensory variance $\sigma_{\text{sensor}}^2$.

**B. Completion Results: Qualitative Inspection**

After training, we tested all networks for their ability to complete trajectories using the first 30 time steps of the three testing trajectories. Some examples of the drawings of two networks per parameter condition are presented in Fig. 4. The orange lines show the initial parts of the trajectories that were presented to the networks. The black and gray lines show the replication of this first part for drawing and nondrawing lines, respectively. Analogously, the green and light-green lines show the networks’ completions.

First, it can be observed that the black lines, in most cases, closely follow the orange lines. The networks, thus, were properly able to reproduce existing parts of the drawing. However, the ability of the networks to complete the patterns drastically changed over the spectrum of the parameter conditions. Generally, networks with values of $H$ close to 1 were able to accurately draw the intended patterns. In contrast, networks with a hypo-prior ($H \gg 1$) were not able to properly complete the patterns. Instead, they mostly exhibited scribbling behavior. The reason is that these networks rely more strongly on external input and not so much on their own predictions. As a result, they can follow existing trajectories, but when the sensory input becomes imprecise or is not available, the networks are unable to complete the patterns: they consider neither the sensory input nor their own predictions as sufficiently precise. Hyper-prior networks ($H < 1$) were able to complete the drawings, but the completions tended to differ from the original training patterns. As these networks only rely weakly on sensory input during training, they abstract from the original patterns, drawing the shapes sometimes in a simplified,
stereotypical way (e.g., column 1 in Fig. 4). In addition, patterns were sometimes confused with other patterns. For example, in Fig. 4, the networks drew a HOUSE instead of a FACE in column 5 and a FACE instead of a CAR in column 2. In some cases, the pattern could not be successfully completed (e.g., in columns 3, 5, and 6).

C. Completion Results: Quantitative Inspection

The results from the qualitative analysis can be confirmed quantitatively by computing the errors of the network’s productions when comparing them to the training trajectories. The distance between trajectories is assessed using dynamic time warping (DTW) to alleviate the influence of small shifts in timing that may be present in the data due to differences in drawing speed in the training data. Since we are interested in the general quality of the produced drawings, we designed the error measure such that confusions between different training patterns are ignored: the error of the generated trajectory to the closest training trajectory is computed.

Fig. 5(a) shows the errors for the first (visible) third of the trajectories. Every point represents the error for one inference trial with one network. The boxes show the median and 80% percentile of the error variations. The error for all $H$ conditions was low. Thus, networks of all prior conditions could follow existing lines. However, there was a small but significant error value increase in networks with a hyper-prior ($H < 1$): the reproduction of the existing lines became less accurate due to a weaker reliance on sensory input.

Fig. 5(b) displays the errors for the completed part of the trajectories where no external input was available. Whereas networks with $H = 1$ and moderate deviations of $H$ still achieved low error values, hypo-prior networks with $H = 100$ and $H = 1000$ showed a significant increase in error for all training patterns. This result corresponds to the scribbling that we observed in Fig. 4 (right-hand side). Therefore, hypo-prior networks have a generally reduced ability to complete representational drawings in a meaningful way.

Additionally, for networks trained in the hyper-prior condition, the error increased in Fig. 5(b), although the corresponding networks did not show scribbling behavior. The reason for this error increase may be the relatively strong reliance of these networks on its own predictions, which causes the networks to disregard some information from the sensory input. Already during training, fine details of the trajectories might not have been taken into account, leading to more errors in the drawn trajectories (see Fig. 3). Therefore, drawings of hyper-prior networks exhibit a higher degree of abstraction and sometimes
fail to generate a pattern, which results in higher error values. Although hypo-prior and hyper-prior networks both showed more errors during completion, the increase in errors in the hypo-prior conditions was more drastic.

Fig. 5 does not consider the fact that the network might confuse some patterns with other training patterns. Therefore, we separately calculated the total number of confusions occurring in each parameter condition by computing to which of the training patterns the network’s drawing had the smallest distance (using DTW distance as above). If this pattern did not equal the intended pattern, we counted it as confusion. Fig. 6 shows the percentage of confusions in different $H$ conditions in blue color. The lowest amount of confusion was found for $H = 1$ and $H = 10$. The number of confusions increased in the hyper-prior conditions to about 23%. Also, in the hypo-prior conditions, the confusion percentage increased. However, as patterns drawn in this condition cannot be clearly recognized as either of the training shapes, high percentages indicate a general inability to complete the patterns, not the occurrence of confusions.

**D. Conclusion About Network Behavior**

Overall, these results show that networks trained with $H = 1$ or $H = 10$ are capable of completing drawings with a low error rate and make fewer mistakes than higher or lower values of $H$. Therefore, these conditions are qualitatively comparable to the cognitive abilities of human children from the age of two to three years, who tend to complete missing parts of a figure according to their experience.

Networks trained with weak reliance on priors (hypo-prior) can follow existing parts of the trajectories, but do not complete them properly. Instead, they show scribbling behavior similar to how chimpanzees and very young children scribble on incomplete drawing stimuli.

Also, aberrations in the opposite directions, namely, too strong reliance on priors, cause certain deficits, such as confusion between patterns and a higher level of abstractions of the patterns. This failure in estimating the intended pattern might correspond to another type of immaturity of drawing behavior which could not be discovered in the experimental setting in [6].

These results suggest that a well-balanced integration of the network’s own predictions and sensory input might be crucial for acquiring the ability to infer the intention of what a drawing should represent and to appropriately complete it by the addition of missing parts.

**VI. RESULTS II: INTERNAL NETWORK REPRESENTATION**

Next, we investigated the precise mechanism in the learning framework that caused the impairment in completion in hypo-prior networks and the occurrence of confusions in hyper-prior networks. In our computational model, there are two potential sources for such failures. The first source is the recognition process, that is, the inference of an initial state given the first part of the trajectory. The second source is the actual generation of the drawing: the generation of the trajectory consisting of the visible part and the new part. The key question is whether already the recognition fails (i.e., the network does not recognize the pattern as one of the learned ones) or whether the network recognizes the shape but does not utilize this information while drawing (due to a general lower tendency to rely on own predictions).

To answer this question and pinpoint the reason for different types of failures in completion, we looked at the internal representations of the networks. In particular, we investigated: 1) how the $H$ parameter influences the inference of the initial states during recognition (Section VI-A) and 2) how $H$ influences the dynamics of the neural activation patterns while generating the drawings (Section VI-B).

For this analysis, we collected all neural activations for each of the ten trained networks (with three trials each) while generating the six target patterns. Then, we standardized the activations to mean 0 and standard deviation of 1, and projected them to the principal component space using principal component analysis.

**A. Analysis of Initial State Inference**

We first evaluated the initial states that were inferred during the recognition process (i.e., the neural activations of the recurrent network layer at time step 0). The initial states acquired during training and three inferred initial states for networks from different prior conditions are shown in Fig. 7, plotted in the space of the first two principal components of the neural activation space. It can be observed that in all conditions, the inferred initial states for one pattern roughly clustered in similar regions, but these regions were overlapping and not necessarily close to the training initial states. Fig. 8 quantitatively compares how close the inferred initial states were to other initial states inferred for the same drawing pattern (blue) or to initial states inferred for different drawing patterns (red). The average distances for all ten networks per prior condition are shown. The results confirm that clustering by pattern during the inference occurred for networks with all values of $H$. There were no clear differences between prior conditions.

Next, we computed for all inferred initial states the closest training initial state. If the closest training initial state did not correspond to the training initial state of the intended pattern, we counted the inferred initial state as incorrectly estimated.
The percentage of incorrectly estimated initial states for different prior conditions is displayed in red color in Fig. 6. It can be observed that the closest training initial state often did not correspond to the correct initial state. The reason seems to be generally large distances of the inferred initial states from the training initial states as shown in Fig. 7 which result from the high-dimensional nature of the network activation space. Recognition using backpropagation inference tends to get stuck in local minima, in particular, in high-dimensional spaces such that the original training initial state cannot be successfully recovered. Furthermore, it is likely that the initial states were incorrectly estimated due to the (intentionally designed) similarity between the initial parts of the trajectories of some of the categories.

Despite the high percentage of incorrectly inferred initial states, additional experiments showed that, nevertheless, the recognition step is crucial for achieving the completion: when replacing the inferred initial states with zero initial states or the mean of the training initial states, the networks are not able to differentiate between different input patterns; they can draw neither the visible part nor the new part of the trajectories accurately.

The percentage of cases where a wrong initial state was inferred was higher in the hyper-prior conditions, indicating that the higher number of confusions in the trajectory (blue bars in Fig. 6) to a degree might have been caused by a tendency to infer a wrong initial state. Notably, there was no better inference in the case of normal prior compared to the hypo-prior conditions. On the contrary, networks trained with hypo-prior were better at inferring the correct initial state, probably because they generally pay more attention to small details of the input trajectory, which helps to differentiate, for example, between FACE and HOUSE.

We can conclude that a correct inference of the exact training initial state was not required for correct completion, and an incorrectly inferred initial state did not necessarily cause a failure in completion and could be alleviated during the actual generation phase. Therefore, differences in the inference of initial states are not the main factors affecting the completion ability but may contribute to the higher number of confusions in the hyper-prior conditions.

**B. Analysis of Neural Activation Patterns**

Fig. 9 displays the neural activation traces (in principal component space) of example networks of three prior conditions. The traces of three independent generations from each network are shown. For clarity of presentation, only three drawing categories are displayed: FACE and HOUSE, which were commonly confused with each other, and FLOWER, a distinct pattern that was rarely confused with other patterns. The top row shows the neural activation traces during the first 30 time steps, and the bottom row shows the continuation of the trajectory when no sensory input was available. Time step 30 of all trajectories is marked in the top and bottom plots with a square for better orientation.

During the first 30 time steps, the trajectories of different target patterns showed distinct activations in all $H$ conditions. Thus, even if incorrect initial states might have been estimated (see Section VI-A), the sensory input available at the beginning of the generation step helps to drive the network dynamics in a way such that the drawing categories are separated. In the upper row, it can be observed that activation patterns were less diverse in the hypo-prior condition compared to the other conditions. The reason is that hypo-prior networks generate the output primarily in an input-driven way; the network dynamics contribute only little to the learning process.
The differences between the conditions became clearly observable once sensory input was not available anymore. In the hypo-prior condition ($H = 100$), although the activation traces separated during the first 30 time steps, the neural activation traces diverged when input became unavailable. Due to a weak reliance on its own predictions, the network is strongly influenced by sensory input. Without input, network activations cannot separate, and the activation traces of different patterns intermingle.

The stronger the prior, the better the different patterns can be separated, leading to a clear separation with $H = 1$. The network activations fell into the acquired attractors of the RNN [33]. In the hyper-prior condition ($H = 0.001$), the network relies almost entirely on its own predictions. Therefore, less variation between different trajectories of the same pattern can be observed in the $H = 0.001$ condition (in particular, for HOUSE), compared to the $H = 1$ condition. The three individual generations of one pattern tended to collapse to similar neural activation traces. Furthermore, in the hyper-prior condition, the neural activations of FACE and HOUSE overlapped: the patterns were drawn in a very similar way and could easily be confused with each other.

To quantify these observations, we investigated the relative distances of the neural activation traces in the 100-D space. Fig. 10(a) shows the average distance between trajectories that generate the same target pattern. This measure assesses the level of diversity in generating individual patterns. The hyper-prior networks exhibited slightly higher diversity at the very beginning of generation, probably caused by the lower influence of external input such that it required more time until a decision was made about which pattern should be generated. However, once this decision was made, the network adhered to it: the variability steadily decreased. This decrease led to the stable generation of patterns, but also entailed rather prototypical patterns with little variability. Changes in the environmental conditions, such as the sudden absence of sensory information, did not affect the generation. Therefore, confusions probably already occur in the first time steps of the generation where a high variability exists between the neural traces. The occurrence of confusions is facilitated by the weak influence of sensory input, which causes the network not to pay close attention to sensory input, on the one hand, and the higher probability of incorrectly inferred initial states, on the other hand (see Fig. 6).

Networks with moderate values around $H = 1$ are able to switch to their own predictions when the sensory input suddenly becomes unavailable. Therefore, the diversity increased after time step 30, but then again converged to a small value. Networks trained with the hypo-prior condition ($H \gg 1$) do not have this flexibility. The sudden change of environmental conditions directly affects the network’s performance, leading to unstable behavior; the distance between same-pattern activation traces did not converge.

To analyze how well the networks can differentiate patterns, we additionally looked at the distances between neural activations that generated different patterns, as shown in Fig. 10(b). A higher value means that different patterns are better differentiated in the neural activation space. For networks in the hyper-prior condition, the differentiation gradually decreased during generation, indicating that the network activations of...
some of the patterns collapsed onto the same pattern as it happened in Fig. 9 with FACE and HOUSE. With moderate values such as \( H = 1 \), the networks better differentiated between the patterns. For hypo-prior networks with \( H \gg 1 \), the distance between the activation traces of different pattern increased during the generation of the first 30 time steps, but became unstable afterward, indicating that the generated activations did not follow distinct trajectories but were governed by random fluctuations.

C. Conclusion About Internal Representations

Our findings indicate that the differences between hyper-prior, normal, and hypo-prior conditions mainly arise during the generation process. While there are differences in the inference of initial states, these differences do not reflect differences in behavior. Instead, network activations seem to be more strongly determined by the network dynamics in response to the first 30 time steps of the trajectory.

The higher number of confusions in the hyper-prior condition probably is caused by the decreased attention to sensory input at the beginning of the generation process while input is still available [see Fig. 10(a)]. Additionally, confusions of the initial states during the inference process might contribute to a larger percentage of confusions overall (see Fig. 6).

These results indicate that the main difference between the conditions lies in the formation of the attractors for different patterns. Specifically, hypo-prior networks do not form sufficiently strong attractors such that the neural activations diverge in the absence of input. In contrast, hyper-prior networks form strong attractors such that neural activations in the wide surroundings of the attractor of one pattern are drawn toward the current attractor. The evidence for these strong attractors can be found in Figs. 9 and 10(a), where a strong convergence of patterns toward similar patterns can be observed in the hyper-prior condition.

To substantiate the hypothesis that the prior condition affects attractor formation, we investigated more closely how network behavior is affected by disturbances of the initial state. Specifically, we looked at how trajectories are generated when using initial states that were not inferred but linearly interpolated between the training initial state of one pattern and of another pattern. Like during the completion, the first 30 time steps were available, after which no external input was provided. Fig. 11 displays the trajectory generation error when the initial state is interpolated between HOUSE (left-hand side of the graphs, interpolation factor 0) and FACE, CAR, and FLOWER (right-hand side of the graphs, interpolation factor 1). The solid lines show the error of the generated pattern to the HOUSE pattern and the dashed lines show the error to the other pattern. With this analysis, we can visualize, to a degree, the structure of the initial state space. If the attractors for different stimuli categories are well distributed, we would expect a gradual change of the error when interpolating between two training initial states: the closer the used initial state is to the training initial state, the more similar should the resulting drawing be to the drawing pattern corresponding to this initial state.

It can be observed that in the normal prior condition \((H = 1)\), all patterns were clearly differentiated. When the initial state of HOUSE is modified into the direction of other initial states, the HOUSE pattern can still be generated with a low error. This result indicates that a good and stable attractor was acquired.

The hypo-prior condition \((H = 100)\) clearly differs from the normal prior condition. Regardless of the used initial state, the generated pattern for HOUSE cannot be differentiated from the generated pattern for FACE or CAR. The difference to FLOWER is still visible owing to the difference in the visible part of the HOUSE and FLOWER trajectories.

The differences between the normal and hyper-prior condition were subtler. However, the error level that was achieved in the hyper-prior condition \((H = 0.001)\) was higher than in \( H = 1 \), indicating that patterns were generally reproduced less accurately regardless of the used initial state. This result was most likely caused by the higher level of abstractions in the generated drawings. Furthermore, it can be observed that the solid and dashed lines were approximately symmetrical to the \( x \)-axis position 0.5 in the \( H = 1 \) condition, but not

\(^{3}\)Note that we can only look at a limited part of the initial state space, namely, at modifications along the interpolation between the trained initial states.
in the $H = 0.001$ condition. This result indicates that attractors were unequally distributed in the hyper-prior condition. Some patterns form stronger attractors than other patterns. As a result, there may be positions in the attractor space that are closer to one of the patterns while reproducing a trajectory that is closer to another pattern. For example, with an interpolation factor of 0.7, the used initial state was closer to FACE than to HOUSE. Still, the error value indicates that the generated patterns were closer to HOUSE than to FACE (see solid blue line in Fig. 11 left). Due to this asymmetry, wrongly inferred initial states (see Fig. 6) might lead more easily to an actual confusion between different patterns.

In conclusion, this analysis demonstrates that hypo-prior networks fail to form proper attractors, whereas networks with hyper-prior or normal prior succeed. Furthermore, networks with a normal prior seem to exhibit a better balance between the trained attractors.

VII. DISCUSSION

The striking differences between the drawing behavior of chimpanzees and human children seem to suggest fundamental differences in their cognitive mechanisms. In this article, we evaluated whether the differences between the species could be explained using the same underlying mechanism, changing only the degree to which sensory information and own predictions are utilized. In our model, the hypo-prior condition qualitatively replicates the behavior of chimpanzees and of very young human children: the model can follow the existing parts of the trajectory, but the drawings are not completed; instead, scribbling occurs. A normal integration of predictions with sensory input, in contrast, replicates the behavior of older human children who develop the ability to complete representational drawings. If the reliance on predictions is stronger than normal, completion is possible; however, the network's drawings are generally more abstract, and some patterns are confused with other patterns. Thus, our results demonstrate that reliance on prior predictions is an important parameter that can result in vastly different behaviors. To achieve the great behavioral flexibility that human children develop, the appropriate integration of predictions and sensory input during the course of development might be a crucial factor. The differences in the cognitive mechanisms of chimpanzees compared to humans could be that they exhibit a weaker tendency to rely on their internal predictions.

An investigation of the internal mechanisms suggests that differences in the integration between sensory input and predictions affect the strength of the emerged attractors in the neural activation space. With a hypo-prior, weaker attractors are formed. Weak attractors indicate a weak tendency to categorize and a generally higher sensitivity to (disturbances in) sensory input.

Although networks with hypo-prior cannot complete drawings, they are better able than networks from other conditions to infer the initial states correctly. This indicates that even when weakly relying on predictions, recognition is possible. In accordance with these results, many studies have shown that chimpanzees are generally able to predict and recognize pictures [3], [34]. For instance, one of the chimpanzees participating in the study from Saito et al. [6] was able to mark human and animal figures in picture books. However, their study could not answer whether chimpanzees failed to complete the pictures because they did not recognize the incomplete face, whether they did not notice that parts were missing, or whether they were not motivated to complete it for other reasons. Our findings suggest that it is possible that the chimpanzees recognized the figure and noticed the missing parts; however, a generally weaker tendency to categorize sensory input into distinct categories might have prevented them from completing the drawing.

In terms of general cognitive capabilities, it is still debated whether and in how far chimpanzees possess a theory of mind [4], [5]. However, most studies suggest that chimpanzees have a reduced tendency to ascribe psychological states to their interaction partners and that they also do not reason about abstract concepts [5]. This fits our finding that hypo-prior networks clearly show less abstraction of sensory input and therefore do not create abstract concepts of the stimuli. Also, cues about intentions in a social context are generally more subtle sensory information than other types of sensory inputs such as physical movements of objects. For example, while observing the goal-directed actions of an actor, it is possible
to understand the goal of the action by observing the action itself or by observing the actor’s face (e.g., looking direction) and inferring the intention of the action. It has been shown by Myowa-Yamakoshi et al. [35] that although chimpanzees and human children are able to infer the goal of a goal-directed action, how they attend to the actor while performing the action is different. Human children from a young age show more attention to the actor’s face when presented with a goal-directed action. Chimpanzees, in contrast, focus strongly on the movements of the arms—preferring to attend to the more salient information. Our computational model provides a possible explanation for this result. The chimpanzee’s behavior might be explained by a weaker reliance on prior predictions. Naturally, when utilizing less prediction ability, it is necessary to receive more accurate input to compensate for this deficit. Therefore, a weak reliance on predictions could also explain differences in the visual attention of chimpanzees and a reduced tendency to utilize subtle social cues in an interaction [3].

While our computational experiment was designed to replicate the original experiment from [6], there are some important differences between the settings. Specifically, it was necessary to pretrain our model with a set of trajectories before performing the completion task. In contrast, participants in [6] were asked to draw more spontaneously, without a clear goal in mind, and did not previously receive training on the particular stimuli. This additional level of abstraction from a drawing to a real-world object made the task generally more difficult for the study participants than for our network model. To validate the effect of this abstraction step, it would be interesting to investigate whether the findings in [6] would be the same in a setting that is closer to our computational model setting, for instance, when explicitly showing them the completed drawing in advance.

Finally, it is important to emphasize that the model proposed here is only one potential implementation of cognitive differences between human children and chimpanzees. Presumably, many different model architectures and hyperparameters could be utilized to achieve similar effects. We cannot judge whether the proposed model is more valid than any other model. However, we based the model on the ideas of predictive coding, which is nowadays a commonly accepted theory for explaining the fundamental mechanisms of cognition. Therefore, our model proposes a plausible explanation of how cognitive differences in human children and chimpanzees may arise.

REFERENCES

[1] A. G. Rosati, L. R. Santos, and B. Hare, “Primate social cognition: Thirty years after Premack and Woodruff,” *Primate Neuroethol.*, vol. 1, no. 9, pp. 117–144, 2010.

[2] J. Call, B. Hare, M. Carpenter, and M. Tomasello, “Unwilling” versus ‘unable’: Chimpanzees’ understanding of human intentional action,” *Dev. Sci.*, vol. 7, no. 4, pp. 488–498, 2004.

[3] E. Herrmann, M. V. Hernández-Lloreda, J. Call, B. Hare, and M. Tomasello, “The structure of individual differences in the cognitive abilities of children and chimpanzees,” *Psychol. Sci.*, vol. 21, no. 1, pp. 102–110, Dec. 2009.

[4] M. Tomasello, M. Carpenter, and R. P. Hobson, “The emergence of social cognition in three young chimpanzees,” *Monographs Soc. Res. Child Dev.*, vol. 70, no. 1, pp. 107–122, 2005.

[5] D. J. Povinelli and S. Giambra, “Reasoning about beliefs: A human specialization?” *Child Dev.*, vol. 72, no. 3, pp. 691–695, 2001.

[6] A. Saito, M. Hayashi, H. Takeshita, and T. Matsuzawa, “The origin of representational drawing: A comparison of human children and chimpanzees,” *Child Dev.*, vol. 85, no. 6, pp. 2232–2246, 2014.

[7] A. Testolin and M. Zorzi, “Probabilistic models and generative neural networks: Towards an unified framework for modeling normal and impaired neurocognitive functions,” *Front. Comput. Neurosci.*, vol. 10, p. 73, Jul. 2016.

[8] P. Lanillos, D. Oliva, A. Philippens, Y. Yamashita, Y. Nagai, and G. Cheng, “A review on neural network models of schizophrenia and autism spectrum disorder,” *Neural Netw.*, vol. 122, pp. 338–363, Feb. 2020.

[9] M. Lopes, F. S. Melo, B. Kenward, and J. Santos-Victor, “A computational model of social-learning mechanisms,” *Adapt. Behav.*, vol. 17, no. 6, pp. 467–483, Aug. 2009.

[10] K. Friston, “The free-energy principle: A rough guide to the brain?” *Trends Cogn. Sci.*, vol. 13, no. 7, pp. 293–301, 2009.

[11] A. Bubic, D. Y. Von Cramon, and R. I. Schubotz, “Prediction, cognition and the brain,” *Front. Hum. Neurosci.*, vol. 4, p. 25, Mar. 2010.

[12] D. M. Wolpert, R. C. Miall, and M. Kawato, “Internal models in the cerebellum,” *Trends Cogn. Sci.*, vol. 2, no. 9, pp. 338–347, 1998.

[13] A. Philippens and Y. Nagai, “A predictive coding model of representational drawing in human children and chimpanzees,” in *Proc. Joint IEEE 9th Int. Conf. Develop. Learn. Epigenet. Robot. (ICDL-EpiRob)*, Oslo, Norway, 2019, pp. 171–176.

[14] C. C. Peterson, “Drawing insight from pictures: The development of concepts of false drawing and false belief in children with deafness, normal hearing, and autism,” *Child Dev.*, vol. 73, no. 5, pp. 1442–1459, 2002.

[15] R. M. Ford and E. L. Rees, “Representational drawing and the transition from intellectual to visual realism in children with autism,” *Brit. J. Develop. Psychol.*, vol. 26, no. 1, pp. 82–89, 1987.

[16] K. Yamagata, “Emergence of representational activity during the early drawing stage: Process analysis,” *Jpn. Psychol. Res.*, vol. 43, no. 3, pp. 130–140, 2001.

[17] P. H. Schiller, “Figural preferences in the drawings of a chimpanzee,” *J. Comp. Physiol. Psychol.*, vol. 44, no. 2, pp. 101–111, 1951.

[18] S. T. Boysen, G. G. Berntson, and J. Prentice, “Simian scribbles: A reappraisal of drawing in the chimpanzee (pan troglodytes),” *J. Comp. Psychol.*, vol. 101, no. 1, pp. 82–89, 1987.

[19] I. H. Iversen and T. Matsuzawa, “Visually guided drawing in the chimpanzee (pan troglodytes) 1,” *Jpn. Psychol. Res.*, vol. 38, no. 3, pp. 126–135, 1996.

[20] Y. Nagai, “Predictive learning: Its key role in early cognitive development,” *Philos. Trans. Roy. Soc. B, Biol. Sci.*, vol. 374, no. 1771, 2019, Art. no. 20180033.

[21] J. Baraglia, Y. Nagai, and M. Asada, “Emergence of altruistic behavior through the minimization of prediction error,” *IEEE Trans. Cogn. Develop. Syst.*, vol. 8, no. 3, pp. 141–151, Sep. 2016.

[22] E. Pellicano and D. Burr, “When the world becomes ‘too real’: A Bayesian explanation of autistic perception,” *Trends Cogn. Sci.*, vol. 16, no. 10, pp. 504–510, 2012.

[23] J. Brock, “Alternative Bayesian accounts of autistic perception: Comment on Pellicano and Burr,” *Trends Cogn. Sci.*, vol. 16, no. 12, pp. 573–574, Dec. 2012.

[24] R. P. Lawson, G. Rees, and K. J. Friston, “An aberrant precision account of autistic perception,” *Trends Cogn. Sci.*, vol. 16, no. 12, pp. 562–571, Dec. 2012.

[25] A. Philippens and Y. Nagai, “Understanding the cognitive mechanisms underlying autistic behavior: A recurrent neural network study,” in *Proc. Joint IEEE 8th Int. Conf. Develop. Learn. Epigenet. Robot. (ICDL-EpiRob)*, Tokyo, Japan, Sep. 2018, pp. 84–90.

[26] H. Idei, S. Murata, Y. Chen, Y. Yamashita, J. Tani, and T. Ogata, “A neurorobotics simulation of autistic behavior induced by unusual sensory precision,” *Comput. Psy.*, vol. 2, pp. 164–182, Dec. 2018.

[27] C. Hartley and M. L. Allen, “Is children’s naming and drawing of pictures mediated by representational intentions? Evidence from typical development and autism,” *Cogn. Develop.*, vol. 36, pp. 52–67, Oct.–Dec. 2015.

[28] D. C. Knill and A. Pouget, “The Bayesian brain: The role of uncertainty in neural coding and computation,” *Trends Neurosci.*, vol. 27, no. 12, pp. 712–719, 2004.
[29] D. Oliva, A. Philippsen, and Y. Nagai, “How development in the Bayesian brain facilitates learning,” in Proc. Joint IEEE 9th Int. Conf. Develop. Learn. Epigenet. Robot. (ICDL-EpiRob), Oslo, Norway, Aug. 2019, pp. 1–7.

[30] S. Murata, J. Namikawa, H. Arie, S. Sugano, and J. Tani, “Learning to reproduce fluctuating time series by inferring their time-dependent stochastic properties: Application in robot learning via tutoring,” IEEE Trans. Auton. Mental Develop., vol. 5, no. 4, pp. 298–310, Dec. 2013.

[31] K. J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, “Classical and Bayesian inference in neuroimaging: Theory,” NeuroImage, vol. 16, no. 2, pp. 465–483, 2002.

[32] S. Murata, H. Arie, T. Ogata, J. Tani, and S. Sugano, “Learning and recognition of multiple fluctuating temporal patterns using S-CTRNN,” in Proc. Int. Conf. Artif. Neural Netw., 2014, pp. 9–16.

[33] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in Proc. Int.Conf. Mach. Learn., 2013, pp. 1310–1318.

[34] M. Tanaka, “Recognition of pictorial representations by chimpanzees (Pan troglodytes),” Anim. Cogn., vol. 10, no. 2, pp. 169–179, 2007.

[35] M. Myowa-Yamakoshi, C. Scola, and S. Hirata, “Humans and chimpanzees attend differently to goal-directed actions,” Nat. Commun., vol. 3, no. 1, p. 693, Feb. 2012.

Anja Philippsen received the master’s and Ph.D. degrees in cognitive computer science and intelligent systems from Bielefeld University, Bielefeld, Germany.

In 2017, she joined the Cognitive Developmental Robotics Lab, National Institute of Information and Communications Technology, Osaka, Japan, as a Postdoctoral Researcher. In 2019, she moved with the lab to the International Research Center for Neurointelligence, University of Tokyo, Bunkyo, Japan. Her research interests include the application of machine learning techniques and robotics for gaining an understanding of human skill development, ranging from the acquisition of speech and language skills to general sensorimotor development.

Yukie Nagai (Member, IEEE) received the Ph.D. degree from Osaka University, Suita, Japan, in 2004.

She was a Postdoctoral Researcher with the National Institute of Information and Communications Technology (NICT), Kyoto, Japan, from 2004 to 2006, and Bielefeld University, Bielefeld, Germany, from 2006 to 2009. Then, she became a Specially Appointed Associate Professor with Osaka University in 2009 and a Senior Researcher with NICT in 2017. Since 2019, she has been a Project Professor with the International Research Center for Neurointelligence, University of Tokyo, Bunkyo, Japan. She has been serving as the Research Director of JST CREST Cognitive Mirroring since December 2016. She has been investigating underlying neural mechanisms for social cognition, such as self-other cognition, estimation of others’ intention and emotion, and altruism based on the theory of predictive coding. Her research interests include computational modeling of human cognitive development and design of assistant systems for developmental disorders.