Towards Cognitive Mirroring: Identifying stimuli and features to investigate hypo-prior in Autism Spectrum Disorder

Master Thesis in Artificial Intelligence

Laura Tigchelaar
s1013029
August 18, 2019

Supervised by:
Yukie Nagai[1]
Anja Philippsen[1]
Pim Haselager[2]

Third assessor:
Franc Grootjen[2]

[1]International Research Center for Neurointelligence, The University of Tokyo, Tokyo, Japan

[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Radboud Universiteit

Radboud University Nijmegen
The Netherlands

1 Acknowledgements

For this project I traveled all the way to Japan, starting my internship in Osaka at the Center for Information and Neural Networks (CiNet) of the National Institute of Information and Communications Technology (NICT) at Suita campus of Osaka University, and ending my internship at the International Research Center for Neurointelligence (IRCN) at The University of Tokyo, where the lab moved halfway through this project.

For providing me with this amazing opportunity, I would like to thank Dr. Yukie Nagai, who has made all of this possible by accepting me as an intern in her research group.

Specific thanks go out to Dr. Anja Philippsen, who closely supervised my work and was always understanding and willing to help.

I want to thank Yoshimoto-san and Ijiri-san for helping me with everything regarding administration and residency. I lived in beautiful places both in Osaka and in Tokyo and I am incredibly grateful for that.

I am grateful to Dr. Pim Haselager, my internal supervisor, for always being patient, showing faith in my work, and guiding me in the right direction, especially in terms of mindset.

Furthermore, I want to thank my parents for all their support, financially, but more important, morally. Without you by my (digital) side, I don't know how I would have managed on the other side of the world.

Lastly, I would like to thank everyone else who was involved in my project, one way or another, because all of you helped me complete this thesis.

Affiliation: International Research Center for Neurointelligence (WPI-IRCN), UTIAS, The University of Tokyo.
This work was supported by JST CREST 'Cognitive Mirroring' (Grant Number: JPMJCR16E2) and by World Premier International Research Center Initiative (WPI), MEXT, Japan.

Abstract

Autism Spectrum Disorder (ASD) is characterized not only by problems with social interaction and stereotyped repetitive behaviours, but studies have shown that people with ASD also have problems with perception and categorization. Following the hypo-prior account by Pellicano and Burr, I suggest that problems in ASD are caused by a hypo-prior, and propose an experiment in which a computational neural network with an adjustable prior-influence imitates participants in a human-robot interaction to investigate this idea. In this thesis, I conduct an experiment in which typically developed (TD) participants draw trajectories they have observed one step at a time, to explore how they respond to the task and the different stimuli. The aim is to identify stimuli and features on which TD participants show generalization behaviour. Any identified stimuli and features can be used in further experiments, and the behaviour by TD participants can be used as a baseline to compare with behaviour by participants with ASD, to investigate the possible presence of a hypo-prior in ASD. In short, if participants with ASD show less generalization behaviour and more accurate replication on trajectories that TD participants do show generalization behaviour on, that would indicate a hypo-prior in ASD. I discuss the results from the experiment and make several recommendations for further experiments investigating hypo-prior in ASD.

Contents

2 Introduction

There has been a lot of research into the developmental Autism Spectrum Disorder (ASD), that is characterized by problems with social communication and social interactions, and restricted repetitive patterns of behaviour [2]. Most of the research, however, focuses only on the social problems. Studies that do investigate perception seem to focus only on perception by itself, and not on perception in a behavioural task. In this thesis I design and conduct an experiment that investigates perception in a behavioral task. For this experiment, I make use of a new idea: using computational models to imitate human behaviour on a task, providing more insight into human cognitive mechanisms through a quantitative measurement.

The general idea is to do a 'Cognitive Mirroring' experiment. Cognitive Mirroring is a concept and a framework suggested by Nagai [29] and in short states that human cognition can be quantified by imitating it with an artificial neural network through a human-robot interaction. During such an interaction between a human and a robot, human behaviour on a task can be observed in a controlled environment and the artificial neural network can learn to imitate the observed behaviour. To realize imitation by the network, parameters should be adjusted. From the achieved parameters and the internal representation in the network that best imitates the human on this task, conclusions can be drawn. This framework provides new opportunities to gain more insight into individual differences in cognition and differences between typical and abnormal development, such as ASD.

Regarding the experiment, I am interested in a theory by Pellicano and Burr [32] that suggests individuals with ASD have a so-called **hypo-prior**. This theory is based on the idea that our perceptual experience is based on the integration of sensory information with prior knowledge, or **prior** in short. Pellicano and Burr suggest that the difficulties faced by individuals with ASD originate in problems with this prior, either in its establishment or in the combination with sensory information. These problems yield reduced usage of internal predictions (hypo-prior), causing more reliance on sensory input. They state that due to this higher reliance on sensory input, such a hypo-prior would result in more accurate perception, problems with resolving ambiguity, and reduced capacity for generalization during learning. These symptoms have also been reported about individuals with ASD [32].

Making a first approach towards Cognitive Mirroring, I conduct an experiment to investigate whether typically developed (TD) participants generalize on the presented stimuli in a trajectory drawing task. A trajectory is a two-dimensional timeseries, in this case a drawing that exists of timesteps representing coordinates on a two-dimensional drawing plane. The research question is: Which shapes are replicable by participants, and which feature modifications elicit generalization behaviour in TD participants? The hypotheses are that participants will be good at replication of the Z and T shapes and worse at the H and X shapes, and that participants will show generalization on feature modifications that are less salient, and accurate replication on the salient feature modifications. If generalization behaviour is observed in the task, we have a baseline to which we can compare ASD in further experiments. **Generalization** of a trajectory in the experiment can be defined as simplification of the trajectory, thereby approximating the statistical average of that stimulus' class. In order to compare the behaviour of participants with ASD to the behaviour of TD participants, it is important that at least (some) TD participants show generalization (due to a normal prior) on at least some of the presented stimuli. In that case, if there is a difference in generalization between participants with ASD and with TD, it could be observed. If a difference in behaviour is found between TD individuals and those with ASD in an experiment including a human-robot interaction, the possibility that this difference is caused by the interaction instead of differences in the prior can not be ruled out. This is particularly relevant in research with participants with ASD, because they are generally known to have problems with social interaction and communication. Therefore, there was no human-robot interaction in the current experiment.

I first elaborate on the perceptual differences in ASD and the current theories, such as the widely known weak central coherence theory [21] and the more recent hypo-prior account by Pellicano and Burr [32], and I explain the importance of the observation of ASD as a continuous disorder for the hypo-prior account. I then discuss the Cognitive Mirroring Framework, which contains the basic principles for the study. Additionally, I introduce the neural network that was used, providing examples of the network behaviour using stimuli that were used in the conducted experiment. Next, I discuss the experiment I conducted. Based on an integration of the results from the experiment with results and conclusions from other studies, I end with recommendations for further experiments.

3 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a developmental disorder that is diagnosed based on problems with social communication and social interactions, as well as restricted, repetitive patterns of behaviour [2]. To provide some background, I summarize the perceptual differences in ASD, giving an illustration of the symptoms, I zoom in on generalization behaviour in ASD, and I summarize two theories that make up the foundation of my research.

Although there have been some studies into the perception of sensory modalities other than the auditory and visual senses, Baum and colleagues [7] find the results are far from conclusive. They present studies suggesting, for example, both decreased [9, 12] and increased/unchanged [36] tactile detection thresholds, or finding that individuals with ASD respond stronger to odors [24], but ending up unable to find a correlation between this response and ASD severity [14]. Therefore, most discussed studies will be focused on the auditory and visual modalities.

Many studies make use of the Autism-Spectrum Quotient (AQ). This instrument was developed by Baron-Cohen and colleagues [5] and intends to quantitatively measure the degree to which an adult possesses the traits associated with the autism spectrum. They validated the instrument by showing that participants diagnosed with Asperger Syndrome (AS) or high-functioning autism (HFA) scored significantly higher than controls and always above 32 out of 50. Therefore, whenever is referred to individuals with ASD, this means in the particular study participants were either diagnosed with ASD, or they had a high score on the AQ, as these are interpreted as equivalent.

3.1 ASD symptoms

Research into ASD indicated perceptual deviations from the typically developed population. People with ASD are shown to be faster and more accurate on tasks measuring visuo-spatial ability, such as the Embedded Figures Test and the Block Design Task [38], as well as the visuo-spatial items of the Raven's Advanced Progressive Matrices [18], they show superior visual search [13], and they suffer stronger interference at higher perceptual loads [8, 37]. All these findings indicate enhanced perceptual functioning.

Related to perceptual functioning is generalization. Training with a dataset that contains a lot of variation due to complex stimuli, as compared to a set that contains simple stimuli with little variation, requires less training for generalization to occur and increases the strength and flexibility of this generalization [10]. A review study by Baum, Stevenson and Wallace [7] that discusses sensory processing differences in individuals with ASD, however, states that more complex stimuli seem to cause more difficulty in ASD, in visual as well as auditory stimuli, suggesting problems with generalization in ASD. They note that with simple, low-level stimuli, participants with ASD often outperform their TD counterparts, but they seem to perform comparably worse when presented with more complex stimuli.

Generalization has an influence on how we categorize, for example into categories such as cats and dogs. We learn that a Dalmatian and a Beagle are dogs, and that a Bengal and a Persian are cats, without learning explicit features that define these categories. Gastgeb and Strauss [20] emphasize the importance of categorization:

"it reduces demands on memory and allows individuals to focus on important aspects of objects while ignoring irrelevant details" (p.1). Perceptual categories, as explained by Gastgeb and Strauss, are not strictly defined, with more typical examples at the center and atypical examples at boundaries of the category. They mention categorization of a new, unfamiliar stimulus is done by comparing it to the statistically averaged prototype. Therefore, categorizing a stimulus typical for the class and more similar to the prototype is easier and faster than categorizing an atypical stimulus.

Categorization provides a way to measure generalization behaviour in ASD. Gastgeb and Strauss [20] have shown that individuals with ASD have more problems than TD individuals categorizing atypical members of a class, both for faces and for objects, suggesting that indeed individuals with ASD have trouble with generalization when presented with more complex or atypical stimuli, but not as much with typical class members, at least from adolescence. Froehlich et al. [17] have shown that people with ASD have intact prototype formation, which indicates they are capable of generalizing stimuli into classes, but they were significantly worse at classifying stimuli that were less similar to the prototype compared to controls, which indicates that top-down generalization is impaired. In a similar trend, Nakano, Kato and Kitazawa [30] have shown intact integration of sensorimotor traces into a global visual shape in ASD, on a task that comprised feeling the contours of an object, and then visually identifying that object from a group of similar objects. Surprisingly, the ASD group performed significantly better on this task, which they then argue can be explained by an impairment in simplification generalization when identifying objects through touch, resulting in more accurate identifications compared to controls. These studies suggest that people with ASD have intact bottom-up generalization (i.e., they correctly develop an applicable prior or prototype), but they are impaired in application of this prior to new, more deviant stimuli.

The studies mentioned previously did not examine perception in a behavioural task, but all simply required participants to recognize the stimuli. With these results on perception, a next step can be taken: looking at behaviour. I am interested in generalization behaviour, because a deficiency in generalization behaviour may explain many if not all symptoms in ASD. Therefore, I conduct an experiment in which participants perform a drawing task. The task is inspired by the research by Froehlich et al [17] and Nakano et al [30], but instead of teaching the different classes, respectively letting participants choose which shape was just presented to them, a trajectory is provided one step at a time, and participants are instructed to replicate the trajectory by drawing it themselves.

A trajectory can be defined as a sequence of locations, in this case in two-dimensional space. Although this is a lot like the stimuli in the study by Nakano et al [30], replication may require different resources as compared to recognition, and therefore it may be the case that trajectory replication is more similar to learning to execute a new movement, even though movements are in three-dimensional space. Movements consist of a combination of variable and stable points [26], and in order to learn a new, goal-directed movement, variation in the variable parts is important to point out which parts of the movement can, and which cannot be changed. Looking at studies with babies, it becomes clear that this indeed is how humans learn new movements [19], with mothers including more variation when the baby does not seem to replicate the behaviour correctly, but keeping the critical parts stable.

In the experiment, participants are presented with stimuli with different shapes and different features. For each shape there was one **prototype**, which represents the approximate average of the shape as a class. Based on the findings by Nakano et al [30], one could expect for participants with ASD to generalize less, and therefore to draw a more accurate representation of the trajectory, specifically regarding the features of the stimulus. For TD participants, however, it can be expected that they generalize some of the features, just like they did when they were asked to identify one of three presented shapes as the one they just touched [30]. Because for the current experiment the trajectories from one shape class would have different features in the same position in time, the stable and variable points in the trajectories were easily identifiable. Therefore, it can be expected from all participants to draw the stable sections comparably, differing only on the variable parts.

Having formed an idea of the expected behaviour of both TD participants and those with ASD based on previous behavioural and perceptual findings, we are left with the question what then causes this behaviour. What is the origin of the problems people with ASD face regarding generalization? There are several theories that consider the cognitive mechanisms that may be deviant in ASD, and of these I will consider first the well-known weak central coherence theory by Frith, and Happé and Frith [16, 21], and then the hypo-prior account by Pellicano and Burr [32].

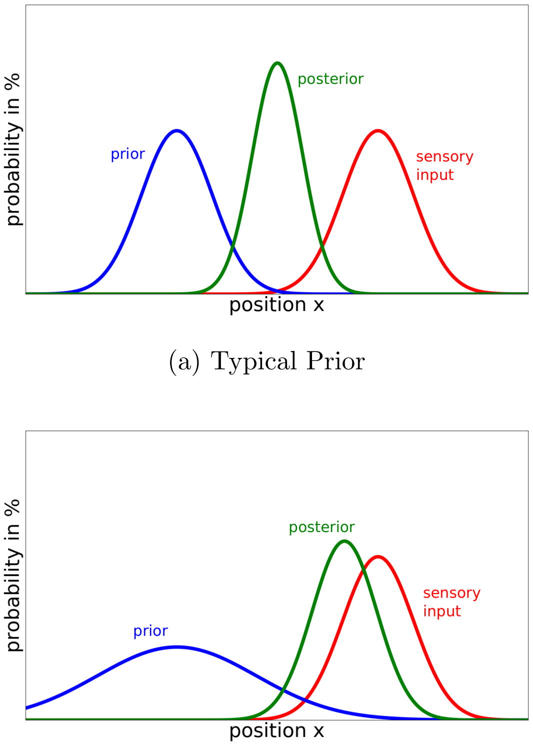
3.2 Theories on the cause of ASD symptoms

In 1989, Frith conceptualized a theory to explain the (perceptual) symptoms in ASD: the **weak central coherence theory** [16]. Frith suggests that individuals with ASD do not have the tendency, as shown by TD individuals, to combine information into a coherent whole, to form a Gestalt, or to process it for meaning. This suggests that people with ASD show more focus on the local features as compared to the whole. This theory was slightly challenged by Plaisted in 1998 [35], when she suggested that weak central coherence might only be a cluster of effects that can be explained better by alternative mechanisms, instead of actually causing the symptoms. She suggests that the weak central coherence effects are due to problems with generalization. Since the conceptualization of the theory in 1989, the theory has been revised, taking into account the suggestion by Plaisted, and the major changes are summarized as follows [21]:

- There is more emphasis on improved to superior local processing, as compared to the formerly main idea of a deficit in global processing.
- The symptoms may not be caused by a deficit, but by more of a cognitive style or processing bias that can be overcome when a task requires global processing.
- Weak central coherence may not be the central cause in ASD, but may exist alongside deficits in social cognition.

After this revision, the theory still does not explain how or why individuals with ASD show enhanced local processing. The **hypo-prior account** by Pellicano and Burr [32], however, attempts to. This account is based on the principles of predictive coding. The main idea of predictive coding is that our perception, which is called the posterior, is inferred from combining sensory information (bottom-up) with prior knowledge (top-down), see Figure 1a (and see Section 4.1 for more details). For example, when reading the word ‘dove’, we may think either of the bird, or of the past tense of ‘dive’. When reading the word in a sentence that provides context, we apply top-down information to derive the meaning of this word, inferring the correct posterior. The hypo-prior account suggests that the perceptual deviations observed in ASD are caused by attenuated priors or **hypo-priors**, emphasizing that this does not suggest individuals with ASD do not have a prior, but instead that their prior is more broad. Pellicano and Burr suggest that the prior in individuals with ASD is too broad, causing them to rely more on sensory information, see Figure 1b. If the hypo-prior account is correct, we would expect for individuals with ASD not to profit from the contextual information where TD individuals would, which is exactly what has been found when people with ASD were asked to read homographs (words of which the pronunciation depends on the context, e.g. dove) out loud [22].

Prior specificity (or prior broadness) can be assessed as a continuous variable, with small changes having small impact on behaviour and large changes having large impact on behaviour. This fits the recent trend of ASD being assessed as a disorder on a spectrum. Until DSM-IV-TR, autism was assessed as a categorical disorder with different categories under the name Pervasive Developmental Disorders (PDD) [1]. This changed when DSM-5 was published in 2013, in which Autistic Disorder, Asperger’s Syndrome (AS), Childhood Disintegrative Disorder and Pervasive Developmental Disorder - Not Otherwise Specified (PDD-NOS) are consolidated into Autism Spectrum Disorder, acknowledging that individuals with autism fall on a spectrum. This change in assessment is based on and supported by findings supporting a ‘broad autism phenotype’, such as higher deficiency rates in social interaction, communication and stereotyped behaviours in families of people with ASD [34], indications that relatives of people with ASD are more likely to express mild autistic traits, regardless of diagnosis [3], and the finding that people with a more mathematical and scientific background score higher on the AQ than those with a more social background [5] combined with the findings that direct family members of people with ASD are employed more often in the field of engineering [6] and that in families of students in the fields of physics, engineering, and mathematics, ASD occurred significantly more often [4], suggesting families of those with ASD possess autistic traits, but not to such an extent that they are



(a) Typical Prior

(b) Hypo-Prior

Figure 1: Integration of a typical prior with sensory input results in a typical posterior (a), but integration of a hypo-prior with sensory input results in a posterior closer to the sensory input (b).

diagnosed with ASD. These findings can be explained from the perspective of the hypo-prior account. If indeed a hypo-prior causes the symptoms of ASD, autistic traits as found in family members and engineers may be caused by a prior specificity that is somewhere in between a typical prior and a hypo-prior.

Relating the hypo-prior account to generalization and the experimental task, it can be expected that people with ASD do not generalize as much as people with TD. In the trajectory drawing task this implies that it can be expected that participants with TD, who are expected to have a typical prior, show generalization on at least some of the stimuli, thereby drawing more towards the class average or prototype, whereas participants with ASD, who are expected to have a hypo-prior, may show generalization on some stimuli as well, but not as much, relying more on sensory information, and thereby drawing more local features of the stimuli.

In this thesis, a first attempt is made at the design of an experiment in which individuals with TD show generalization, drawing a trajectory close to the class average. When TD participants show generalization, we have a baseline to compare ASD to in further experiments. By presenting participants with trajectories of different shapes and with different features, I hope to identify both shapes that are not too difficult, such that participants are able to replicate the general outline of the shape, and features that are not too salient, but obviously there. In subsequent experiments, the trajectories drawn by the participants will be compared to the outputs by a computational network with a variable prior to decide to which prior the human drawing matches and to assess whether the hypo-prior theory is suitable for assessing some of the characteristics of ASD.

4 Cognitive Mirroring

To investigate cognitive mechanisms, Nagai [29] suggested a framework she called Cognitive Mirroring. In the Cognitive Mirroring framework, a human interacts with a robot, and during this interaction the robot will learn to imitate the behaviour of the human. More specifically, an artificial neural network that is implemented in the robot learns with which parameters it is able to imitate the behaviour of the human. An example of such an interaction would be a child with ASD interacting with a robot in play.

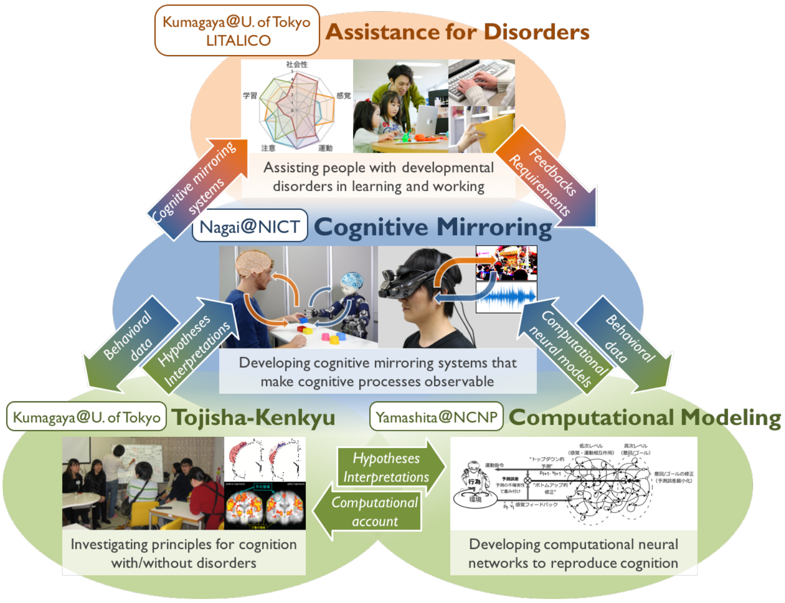

Kumagaya@U. of Tokyo
LITALICO
Assistance for Disorders
Assisting people with developmental disorders in learning and working
Cognitive mirroring systems
Feedbacks
Requirements
Nagai@NICT
Cognitive Mirroring
Developing cognitive mirroring systems that make cognitive processes observable
Behavioral data
Hypotheses Interpretations
Computational neural models
Behavioral data
Kumagaya@U. of Tokyo
Tojisha-Kenkyu
Yamashita@NCNP
Computational Modeling
Hypotheses Interpretations
Computational account
Investigating principles for cognition with/without disorders
Developing computational neural networks to reproduce cognition

Figure 2: The Cognitive Mirroring Framework by Nagai is part of an interdisciplinary collaboration initiative. Taken from [29].

The Cognitive Mirroring framework by Nagai (Figure 2) is part of a collaboration project between multiple labs. The aim of the cognitive mirroring project is to better understand ASD and other mental and developmental disorders, and to use the knowledge from the disorders to get a better understanding of cognitive mechanisms in TD individuals. Using computational models, Cognitive Mirroring systems are created that attempt to imitate behaviour typical for (develop-)mental disorders, making underlying cognitive processes more observable. These systems are then verified and tested from a first-person perspective, providing feedback to improve the Cognitive Mirroring system. Nagai, who now has her lab at the University of Tokyo, collaborates with Kumagaya, who works on Tojisha-Kenkyu: self-support research. Tojisha-Kenkyu is an initiative unique to Japan, and the main idea is that people with a mental or developmental disorder actively participate in the research into their own disorder. As Kumagaya states it: “Those experiencing similar difficulties carefully and compassionately watch over each other as they work to formulate hypotheses about themselves, and then test these hypotheses experientially in their daily live” [25]. In summary, although it is a relatively new approach, the Cognitive Mirroring framework is expected to have a large impact on scientific approaches towards understanding developmental disorders, as well as the psychiatric population suffering from developmental disorders.

The role of the human-robot interaction in the Cognitive Mirroring system is to

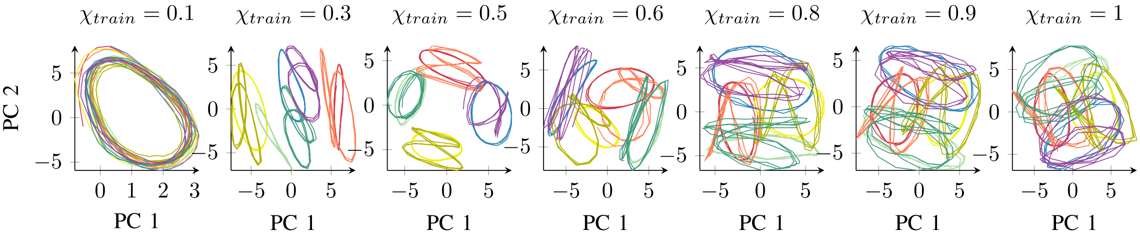


Figure 3: Results from Philippsen and Nagai [33] displaying two-dimensional PCA of network activations of one trial (using the same initial network weights) for proactively following the target trajectories for different values of the parameter *external contribution* (χ_{train}), which reflects how precisely the network tries to replicate the external signal. The figure shows large differences between the internal network representations for different values of the parameter, but Philippsen and Nagai found no significant effect of internal representation on network behaviour. Therefore, they suggest different cognitive mechanisms might not (always) be observable in behaviour.

exert natural behaviours from the human, while keeping a direct way of observing this behaviour in a controlled environment. A robot is very consistent in its behaviour and therefore the results of the interaction are not influenced by differences in the experimental setting, making the results more reliable. Furthermore, the deterministic appearance of the robot may make interaction easier for children and even adults with ASD [11]. In addition, experiments with children with severe ASD may be complicated since they may not understand instructions due to language deficiency. Playing with a robot may not need explicit instructions, creating a more natural and comfortable environment for the child, while preserving the ability to analyze their behaviour.

In Cognitive Mirroring experiments, the implemented computational network aims to provide more insight into cognitive mechanisms. In ASD, for example, participants only show deviant behaviour in some, but not all cases. Therefore, it is hard to draw conclusions based on behaviour only. When imitating this behaviour, that sometimes does and other times does not deviate, with a neural network, it becomes possible to determine the parameter values necessary for imitation and to visualize internal representations. This framework thereby provides a new way to quantitatively measure cognitive mechanisms.

In order to identify a computational model and parameters that can simulate differences in ASD, Philippsen and Nagai [33] let a recurrent neural network reproduce a two-dimensional trajectory with different parameters. They found that the network was able to achieve good performance on the behavioural level, even with less well structured internal representations. Figure 3 shows the influence of a parameter on the internal representations of the network, indicating both small and high parameter values yield unstructured internal representations, whereas medium values yield the most structured internal representations. However, Philippsen and Nagai found no differences in behaviour (not shown here), from which they conclude that differences in internal representations may not always lead to aberrant behaviour, but maybe only on one end of the spectrum. The study by Philippsen and Nagai only included computational modeling, so the next step in the Cognitive Mirroring framework would be to use such a network to compare to human behaviour, and to investigate which parameters can explain deviant behaviour. In this thesis, I attempt to design an experiment that can be used in a Cognitive Mirroring experiment to get more insight into cognitive mechanisms in both ASD and typical development. Therefore I use a similar approach with a continuous parameter, but as opposed to the study by Philippsen and Nagai, I use a parameter that directly changes the behaviour so the subject's behaviour can be imitated.

5 Computational Network

As explained in the previous chapter, a Cognitive Mirroring experiment requires a computational network that can replicate the behaviour of the human participant. It is important that such a network has an adjustable parameter that affects the behaviour of the network on the experimental task. It must be possible to explain this behavioural change by current hypotheses on the underlying cognitive mechanisms, such that the network behaviour is representative of how a human is expected to behave if the hypothesis is correct.

The hypothesis of interest in this study is the hypo-prior account by Pellicano and Burr [32], which suggests that individuals with ASD have a hypo-prior, or a weaker prior. This theory is based on the principles of predictive coding, which is shortly explained next. Afterwards, the computational network that is suggested for the experiment and that incorporates the principles of predictive coding is discussed.

5.1 Predictive Coding

The predictive coding formulation of perception [15, 27] states that perception is an integration of bottom-up sensory input with a top-down prediction based on experience, forming a posterior. As illustrated in Figure 4, if our prediction of the next sensory input is not in line with the actual sensory input, a prediction error occurs, and an attempt will be made to minimize this prediction error by updating the internal model that forms the prediction, so the prediction next time may be closer to the sensory input.

The result of the integration of sensory input with the prediction from the internal model depends on the confidence in both signals. More confidence in the sensory input implies the sensory input has more influence on the posterior. To decide on the levels of confidence, an estimation of the accuracy of both the sensory input and the internal prediction is made, which we call the precision. High sensory precision, thus, indicates more confidence in the sensory input, relative to the internal prediction. By suggesting a broader prior, the hypo-prior account [32] proposes people with ASD have less confidence in the internal prediction, which increases the relative confidence in sensory information. This would explain the focus of individuals with ASD on local features, as compared to the global whole.

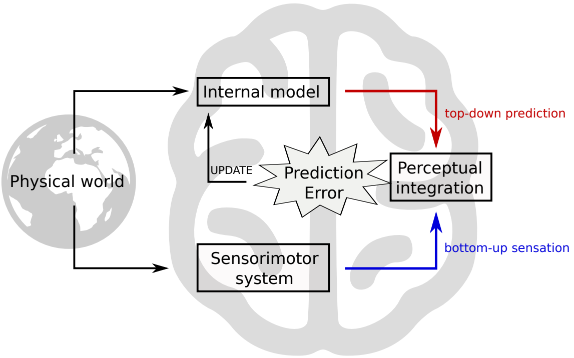


Figure 4: Schematic illustration of predictive coding. Sensory input is integrated with the prediction from the internal model. If they don't match, a prediction error occurs, causing for the internal model to be updated. Figure is courtesy of A. Philippsen.

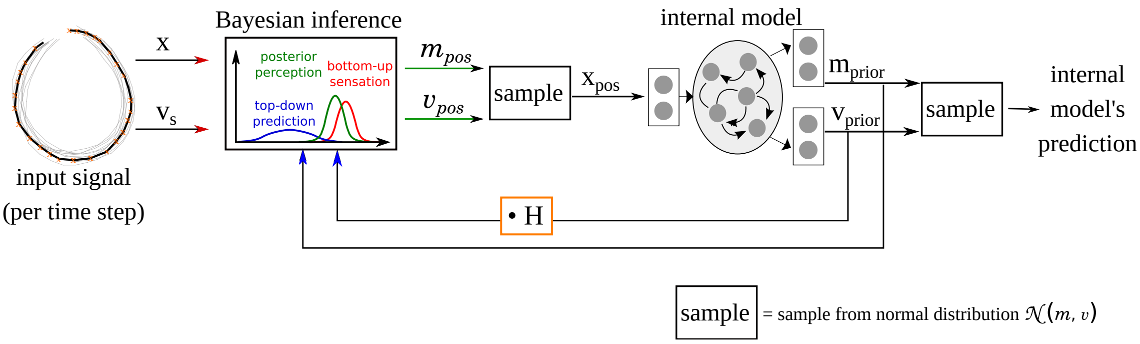


Figure 5: The implementation of the computational network, displayed here for a two-dimensional drawing task, with on the left the input signal (signal x with variance v_s) that is given to the Bayesian inference layer where it is combined with the prediction (signal m_{prior} with variance v_{prior}) from the S-CTRNN (internal model in the figure). This results in a posterior (x_{pos}) that is then given to the S-CTRNN, that makes a prediction for the next timestep. Parameter H can be adjusted to change the influence of the prediction from the internal model on the posterior that results from Bayesian inference, which mimics the hypo-prior theory. (Figure is courtesy of A. Philippsen)

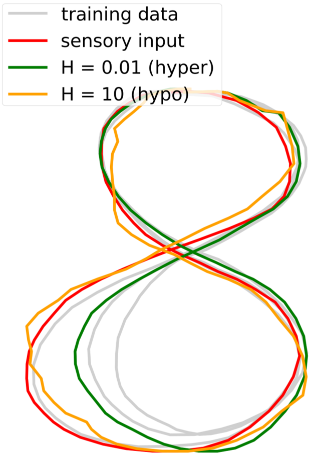


Figure 6: When tested with a hypo-prior (orange), the network implementation produces a result that is closer to the input signal (red), whereas with a hyper-prior (green), the result is closer to the average of the training data (grey), which can be interpreted as the prior.

5.2 Stochastic Continuous-Time Recurrent Neural Network

The artificial neural network that is suggested for the Cognitive Mirroring experiment this study works towards is a Stochastic Continuous-Time Recurrent Neural Network (S-CTRNN), as presented by Murata and colleagues [28], combined with a Bayesian inference layer. Murata and colleagues showed this dynamic network can learn multidimensional time-series, and to extract stochastic or fluctuating structures that are hidden in time-series data, that is, it can learn to recognize unpredictability in data by minimizing the prediction error using backpropagation. Figure 5 shows a schematic representation of the implementation of the network with the Bayesian inference layer. As shown in the figure, the S-CTRNN does not only output a prediction, but it also provides an estimation of the variance of the prediction, representing the confidence of the network in the prediction. The network prediction is integrated with the input signal in the Bayesian inference layer, resulting in a posterior with corresponding variance, which is then given to the S-CTRNN as input for the next prediction. Please note that the input signal also has a variance estimation that represents the amount of noise that is present in the trajectory due to measurement error and trajectory variance, adding up to a value of roughly 0.05 (i.e. v_s in Figure 5 is set to a value of 0.05).

To make the network implementation mimic the influence of a hypo-prior as suggested by Pellicano and Burr [32], the relative confidence in the prediction should be reduced, thereby increasing the relative confidence in the input signal. To change the confidence in the prediction made by the network, the variance estimation is multiplied with value H. A large value for H thus increases the estimated variance, thereby decreasing the confidence in the prediction and mimicking the influence of the hypo-prior.

The behaviour that the network will replicate in the experiment is drawing trajectories. Tests with the network have shown that adjusting H indeed mimics the expected behaviour (Figure 6), drawing a trajectory more like the input signal with a hypo-prior, and more like the average (or prior) when tested with a hyper-prior.

6 Experiment

An exploratory experiment was conducted to investigate generalization behaviour of TD subjects. For this, a new task was designed that makes use of two-dimensional trajectories. The aim of the experiment was to assess how accurate participants could replicate these shapes and the influence of features added to these shapes (modifications) on generalization behaviour in drawing the trajectories. It is expected that participants generalize towards the prototype of the shape on some trajectories with less obvious features, such as slope, and not to generalize towards the prototype of the shape on trajectories with more salient features, such as size.

6.1 Methods

Participants

Eight typically developed individuals participated in the study (4 males, 4 females, $M_{age} = 26.75$, $SD_{age} = 3.28$). All participants declared they were not diagnosed with any mental disorder, including ASD, at the moment of testing.

General Display

Stimuli were displayed on a Dell 55 4K Interactive Touch Monitor model C5518QT with a 55 inch screen diameter. The monitor was placed flat on a table, and the participant stood in front of it such that the longer side (the bottom) of the screen was nearest to the participant. The screen responds to both a provided touchscreen pen and touch. Participants got specific instructions which to use during the experiment. The experiment was programmed in Python and presented using the Matplotlib package.

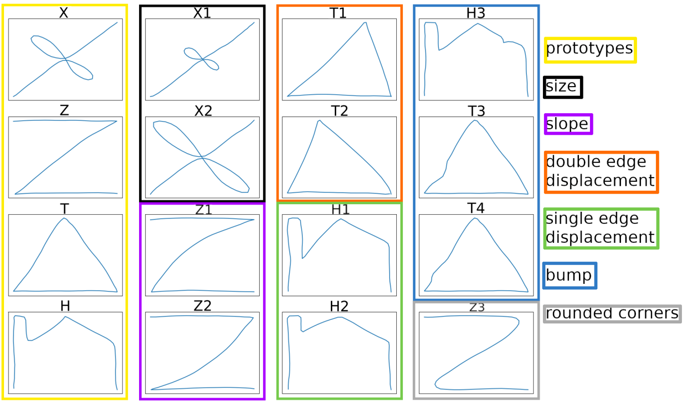


Figure 7: Stimuli per modification with different shapes. Prototypes per shape are shown in the yellow frame. Features are: size (black), slope (purple), double edge displacement (orange), single edge displacement (green), bump (blue), and rounded corners (grey).

Stimuli

The stimuli in the experiment consisted of 16 hand-drawn trajectories with four general shapes (called X, Z, T, and H) and modifications that were created by adding features to the general shapes, illustrated in Figure 7. The X shape is based

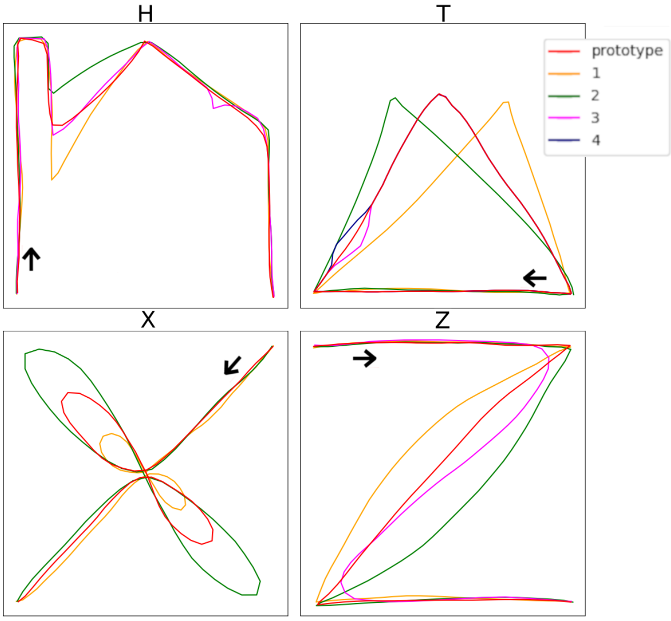


Figure 8: For each shape, the prototype (red) and all modifications of that shape corresponding to the different features. The prototype is a hand-drawn approximation of the average over all modifications of a shape. Arrows indicate the direction in which the shape was drawn.

on the handwritten letter *X*, but rotated 90 degrees. The Z shape is drawn like the letter *Z*. The T in the T shape stands for *triangle*. And the H in H shape stands for *house*. Modifications were:

- **size**, which was applied to the X shape, affecting the curls of the shape, comprising a shape with smaller curls (X1) and a shape with larger curls (X2).

- **slope**, which was applied to the Z shape, affecting the slope of the diagonal, comprising a shape with a convex slope (Z1) and a shape with a concave slope (Z2).

- **double edge displacement**, which was applied to the T shape, affecting the two upper edges, comprising a shape with the two upper edges displaced to the right (T1) and a shape with the two upper edges displaced to the left (T2).

- **single edge displacement**, which was applied to the H shape, affecting two edges representing 'the right side of the chimney' and 'the left side of the roof' when assessing the H shape as representing a house, comprising a shape with a shorter 'chimney'-edge and displaced left side of the 'roof'-edge, and a shape with a longer 'chimney'-edge and displaced left side of the 'roof'-edge.

- **bump**, which was applied to the H and T shapes, affecting one of the diagonal edges, comprising an H shape with a concave bump at approximately 2/3 of the right side of the 'roof'-edge (when assessing the H shape as representing a house), a T shape with a concave bump at approximately 1/3 of the left upper edge, and a T shape with a convex bump at approximately 1/3 of the left upper edge.

- **rounded corners**, which was applied to the Z shape, affecting both corners, comprising a shape with rounded instead of sharp corners.

All trajectories of the same general shape are drawn in the same direction, start at the same location, and end at the same location, for which the first and last location of the trajectory are not necessarily the same. Each of the four shapes starts in another corner of the screen to make a clear and balanced distinction between the different shapes. All stimuli are drawn by hand and interpolated such that all stimuli consist of 70 timesteps, each timestep defined by a location for that point in time in the trajectory. Because the general aim of this experiment is to exert generalizing behaviour from the participants, there was a **prototype** for each shape towards which generalization is expected. The prototype is defined as the approximate average over all trajectories of the same shape. Figure 8 shows for each shape the prototype of the shape, the modifications for each shape with the different features, and the direction in which the trajectory was oriented.

Procedure

During the entire experiment, which in total took about 30 minutes, participants stood in front of the screen and were instructed to use the provided touchscreen pen for drawing the trajectories. For other interactions with the touchscreen they were free to use their hands or the provided touchscreen pen as they pleased.

Participants were instructed to thoroughly investigate the trajectories and to replicate them to their best abilities. They were presented with two rectangle-shaped windows, one on the left and one on the right side of the screen (Figures 9 and 10), which both contained a light grey dot in each corner to indicate the starting and ending points of the trajectories. Only one of the windows was active at a time. When beginning the experiment, the left window was active, in which participants could investigate the trajectory by pressing the buttons underneath this window. For investigation, participants could choose to show the next timestep, the previous timestep, or to automatically show (or stop showing) subsequent timesteps with 0.3 second intervals. Maximally one timestep was visible at a time. Participants were instructed to press *draw* whenever they felt ready to replicate the trajectory.

Participants were then instructed to replicate the trajectory in the window on the right, starting and ending in the same corners as the example trajectory. If the drawing was ready, they could press *ready*, but if they made a mistake they could clear the screen and start over by pressing *clear*. Participants were requested to draw the trajectory in one single movement, starting over whenever they relieved the pen from the monitor before finishing the full trajectory. The trajectory drawn by the participant would be directly shown on the monitor, represented by dots, as

	Group 1		**Group 2**	
Phase 1	All simplified trajectories in fixed order	4	All simplified trajectories in fixed order	4
	All other trajectories in random order	12	3x all simplified trajectories in random order	3*4=12
Phase 2	All trajectories in random order	16	All trajectories in random order	16

Table 1: Procedure per group per phase. The total number of stimuli per phase was constant at a total of 16 stimuli. Phase 2 was equal for both groups. Instructions were equal for both groups and both phases.

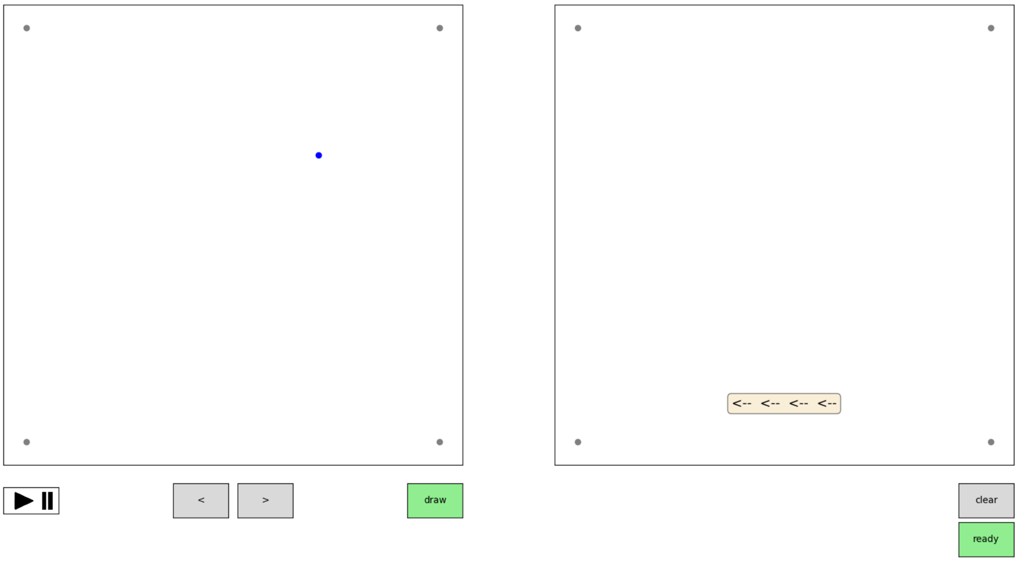


Figure 9: Experiment display during investigation of the trajectory on the left (window on the right is inactive). Only one timestep is visible at a time. Buttons below the screen can be used to navigate through the trajectory: pressing the arrow to the right to show the next timestep, the arrow to the left to show the previous timestep, and the *play/pause* button to automatically show subsequent timesteps with intervals of 0.3 seconds. Pressing the *draw* button activates the window on the right and deactivates the window on the left.

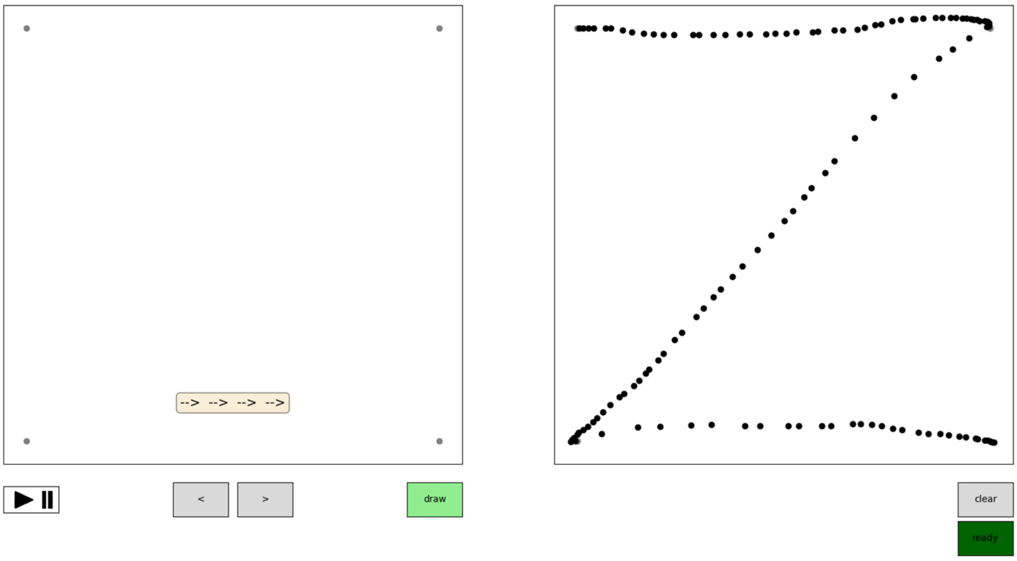


Figure 10: Experiment display during drawing of the trajectory. Here, the subject has made an attempt to replicate the trajectory as was investigated in the window on the left (now inactive). The subject can choose to clear the drawing and start over, or to submit the drawing and continue the experiment. Please note the *ready*-button is darkened because it is already being pressed.

shown in Figure 10.

The experiment consisted of two phases: first a familiarization phase, followed by a testing phase. Participants were divided into two groups. In the familiarization phase, group 1 saw first for each shape the simplified trajectory, and then all other trajectories in random order, whereas group 2 saw first for each shape the simplified trajectory, and then for each shape three times the simplified trajectory in random order. In the testing phase, all participants were presented with all trajectories in random order. Thus, per phase participants were presented with 16 stimuli, adding up to a total of 32 stimuli for the complete experiment, see Table 1. Instructions were the same for all conditions and phases and were displayed on the monitor. The familiarization phase was introduced as a practice phase.

Analyses

Since I am interested in generalization behaviour, I analyzed the drawings by the participants by comparing them to both the target trajectory (the trajectory presented during investigation in the left window) and to the prototype of the shape of the target trajectory. Whether a participant showed generalization behaviour depended on the proximity of the participant's drawing relative to both the target trajectory and the prototype, which is explained below. Please note only modified trajectories were analyzed, because there can be no generalization on the prototypes. For the trajectories to be mathematically comparable they were preprocessed. The trajectories drawn by participants were interpolated such that they all consisted of 70 timesteps (the same as the target and prototype). Next, for all trajectories the distance between the locations (coordinates) of two timesteps was equalized, so target, prototype and participant's drawing had comparable stepsizes. Analyses were performed on the testing phase only.

Hausdorff distance is the maximum of the shortest distances between each point in A and any point in B

$$h(A,B) = \underset{a \in A}{max}\{\underset{b \in B}{min}\{d(a,b)\}\} \tag{1}$$

where h is the Hausdorff distance, A and B are both sequences of locations, d is the distance between location a and location b. To make the measurement symmetrical, the maximum Hausdorff distance was calculated:

$$H(A,B) = max(h(A,B), h(B,A)) \tag{2}$$

where H is the maximum Hausdorff distance between sequences A and B.

To obtain a quantitative measure of generalization behaviour a *critical area* was identified for each of the modified trajectories. The critical area is the area in which the feature is visible, and is calculated per trajectory by identifying the region where target and prototype differ from each other above a threshold that is manually determined for each feature modification. This difference was calculated using the

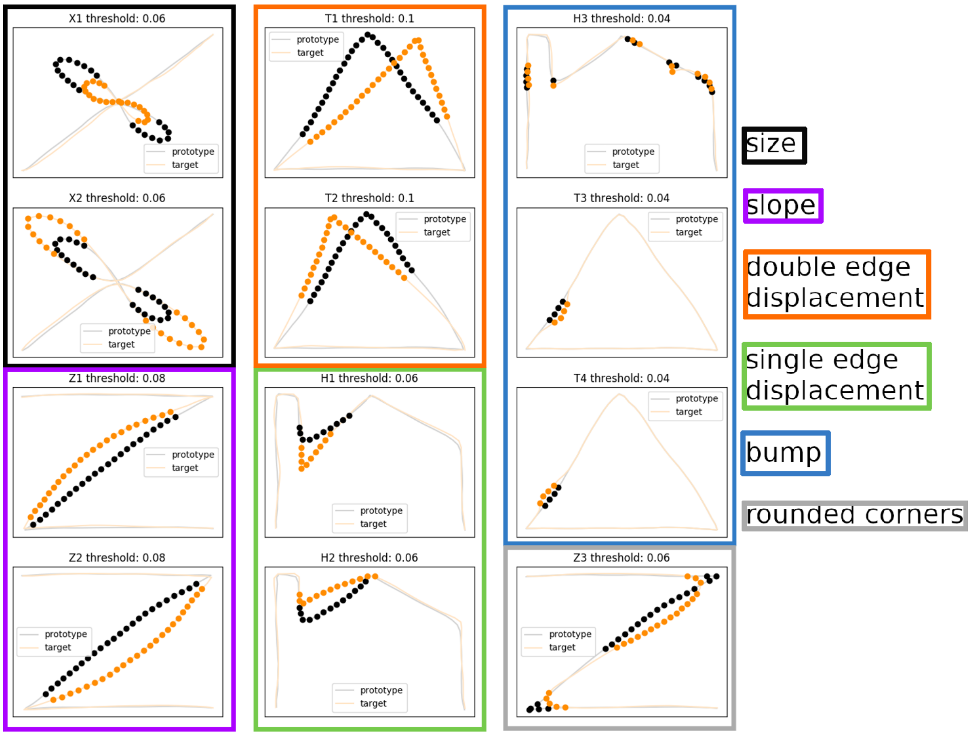


Figure 11: Critical areas (darkened dots) visualized on the prototype (black dots, grey line) and target (orange dots, light orange line) for each trajectory. Each feature has a fixed threshold, displayed above each trajectory. Features are: size (black), slope (purple), double edge displacement (orange), single edge displacement (green), bump (blue), and rounded corners (grey).

Euclidean distance between corresponding timesteps[1] in the target and prototype trajectories. Identified critical areas were then applied to participants' drawings for comparisons. Figure 11 displays the identified critical areas for the different trajectories per feature.

Based on the identified critical areas a score was calculated (see Algorithm 1) making use of the maximum Hausdorff distances (see textbox) between the three (participant's drawing, target and prototype) trajectories. This score indicates the behaviour shown by the participant for this specific target. Participants could show different types of behaviour:

- They could show generalization, which means they drew a trajectory that was exactly like the prototype (full generalization), or a trajectory that was in between the target and the prototype (partial generalization).
- They could show accurate replication, which means they drew approximately the target.
- They could show exaggeration, which means they drew an exaggeration of the target.
- They could show unexpected behaviour, which could mean they drew a trajectory that was closer to the prototype than to the target but not in between them, or it could mean they did something unexpected, such as alternating between aforementioned behaviours.

[1] All trajectories consist of 70 timesteps, so for example timestep 5 in the target trajectory corresponds to timestep 5 in the prototype trajectory.

These different behaviours are shown in Figure 12, with the colour of each behaviour representing the corresponding score that results from the calculation in Algorithm 1. In short, a negative score indicates a drawing closer to the target trajectory than to the prototype, with a negative score that is smaller than -1 indicating exaggeration behaviour and scores bigger than or around -1 indicating accurate replication. A positive score indicates a drawing closer to the prototype than to the target trajectory, with a score bigger than 1 indicating unexpected behaviour and scores smaller than or around 1 indicating generalization behaviour. A score of 0 indicates a generalization tendency, with a drawing that is exactly in the middle of the target trajectory and the prototype.

```
if drawing closer to target than to prototype then
    Score = -1 * (H(drawing, prototype) / H(target, prototype))
end
if drawing as close to target as to prototype then
    Score = 0
end
if drawing closer to prototype than to target then
    Score = 1 * (H(drawing, target) / H(target, prototype))
end
return Score
```

Algorithm 1: How to calculate the behavioural score. A positive score can be interpreted as either generalization, or unexpected behaviour. A negative score can be interpreted as accurate replication behaviour, or exaggeration behaviour.

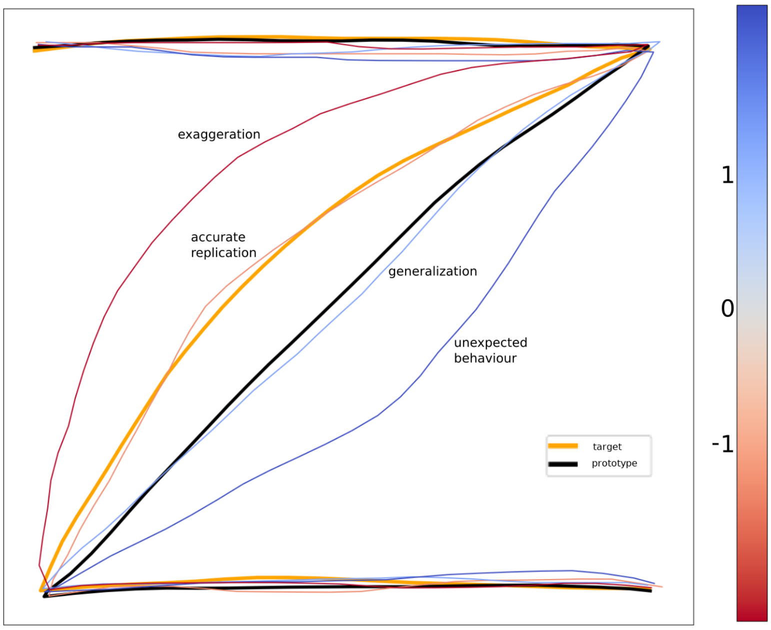


Figure 12: Visualization of score in terms of shown behaviour. On top of the target (orange) and prototype (black) trajectories, four example behaviours are displayed in the colour representing the ascribed score (see colourbar). Next to each example trajectory is the interpretation of the behaviour.

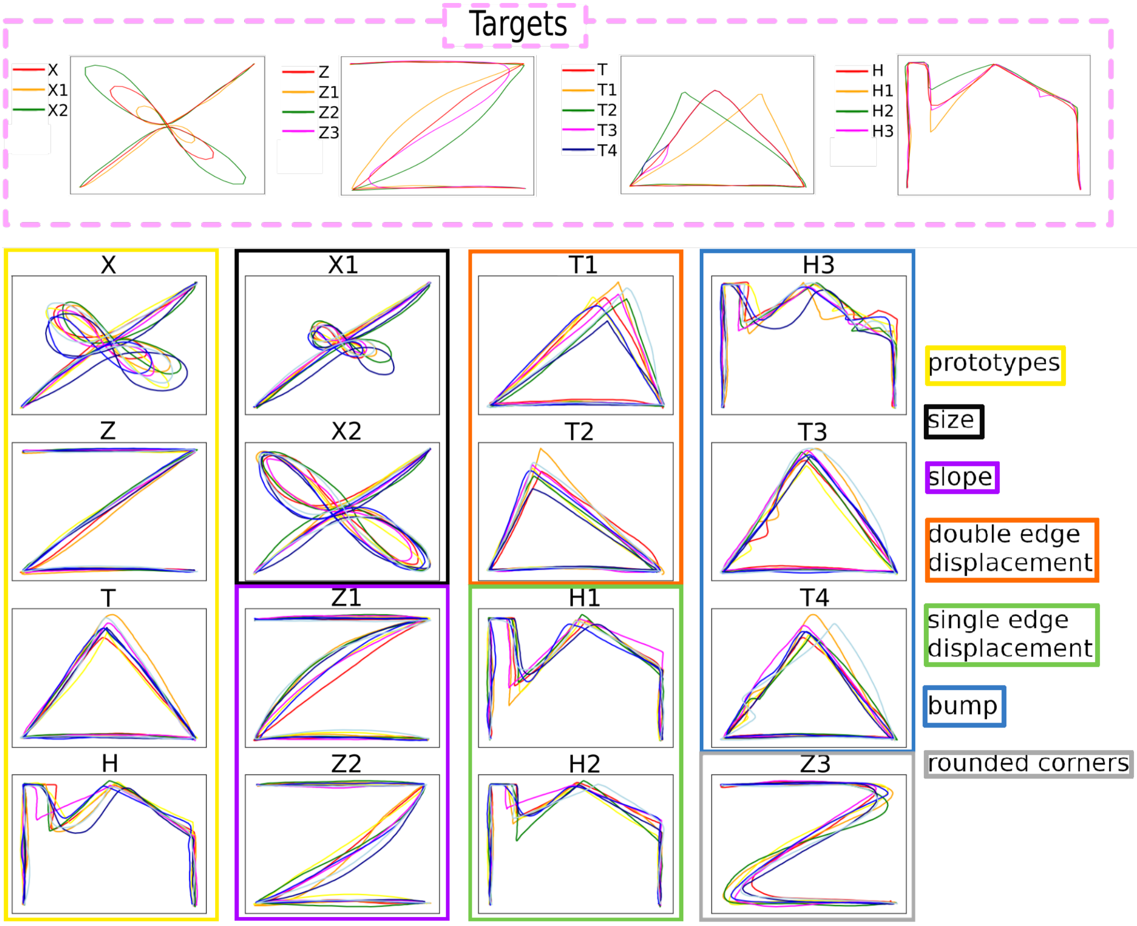


Figure 13: Results from the testing phase on each of the target trajectories sorted by feature. For clarity, the targets are also shown, plotted per shape (in the pink striped border). Each individual participant is represented by a coloured line. Prototypes for each shape are shown in the yellow frame. Features are: size (black frame), slope (purple frame), double edge displacement (orange frame), single edge displacement (green frame), bump (blue frame), and rounded corners (grey frame).

6.2 Results

The aim of the experiment was first to identify general shapes that are replicable by participants, and second to identify modifications (i.e. prototype trajectories with added features) that may elicit generalization behaviour in TD participants. First, it was expected that participants would be good at drawing the Z and T prototypes, but worse at the H and X shapes, that have a more difficult construction. Second, it was expected that participants would show accurate replication behaviour on the salient features size and rounded corners. The other features (i.e. slope, double edge displacement, single edge displacement, and bump) are less salient, so some generalization behaviour is expected for these features.

Since no difference has been found between the two groups, results will be shown for both groups as a whole. Results for each trajectory are shown in Figure 13 and show that participants were accurate at drawing the Z and T prototypes, less accurate at the H prototype and least accurate at drawing the X prototype. Accurate, here, indicates that the drawn trajectories correspond to the target trajectory, and that there was not much variability between participants. The results for the modified trajectories are assessed quantitatively with the scores from Algorithm 1 and shown in Figure 14a. In general, this figure shows that on most modified trajectories, most participants showed accurate replication (score around -1), but on some trajectories participants showed exaggeration behaviour (score smaller than -1), generalization behaviour (score around 1) or unexpected behaviour (score larger than 1; as illustrated in Figure 12).

Participants' behaviour differed between and sometimes within the different features (Figures 13 and 14b):

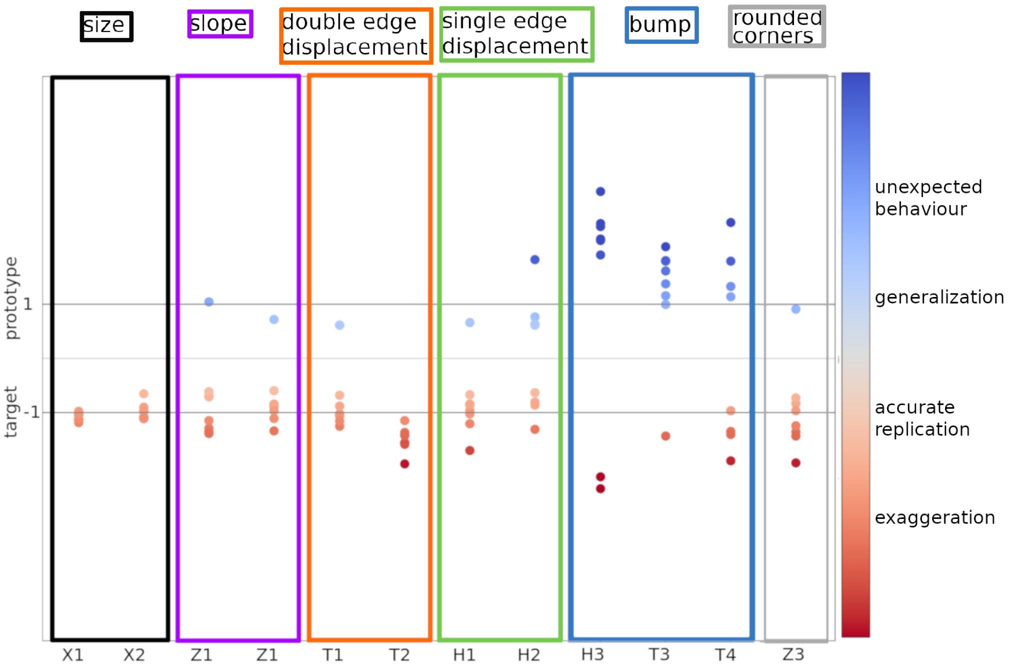
- **Size** appeared to be a very salient feature. Participants were very accurate at reproducing trajectories that differed from the prototype by size (X1 and X2), so no generalization behaviour occurred.
- **Slope** elicited different types of generalization behaviour. Most participants reproduced the target trajectories (Z1 and Z2) correctly, but some drew the prototype or something in between.
- **Double edge displacement**, for some participants, elicited partial or even complete generalization, but for most participants it elicited exaggeration behaviour. Exaggeration behaviour was especially present in the T2 shape, which has a smaller displacement than the other trajectory with this feature (T1). Additionally, participants drew trajectory T2 with less variation than trajectory T1.
- **Single edge displacement** caused some generalization behaviour, some exaggerations, but mostly correct replications of the modified trajectories (H1 and H2). Interestingly, from Figure 13 it seems that participants were better at drawing the modified trajectories than the prototype.
- **Bump** elicited unexpected behaviour from most participants on all three trajectories with this feature (H3, T3 and T4). There was also some, though few, generalization behaviour, and accurate replication.
- **Rounded corners** covered trajectory Z3 and showed a lot of variation in how the trajectory was drawn, but barely any generalization behaviour.

As can be seen in Figure 14b, participants that showed generalization behaviour in one of the trajectories were not more likely to show generalization behaviour in another trajectory.

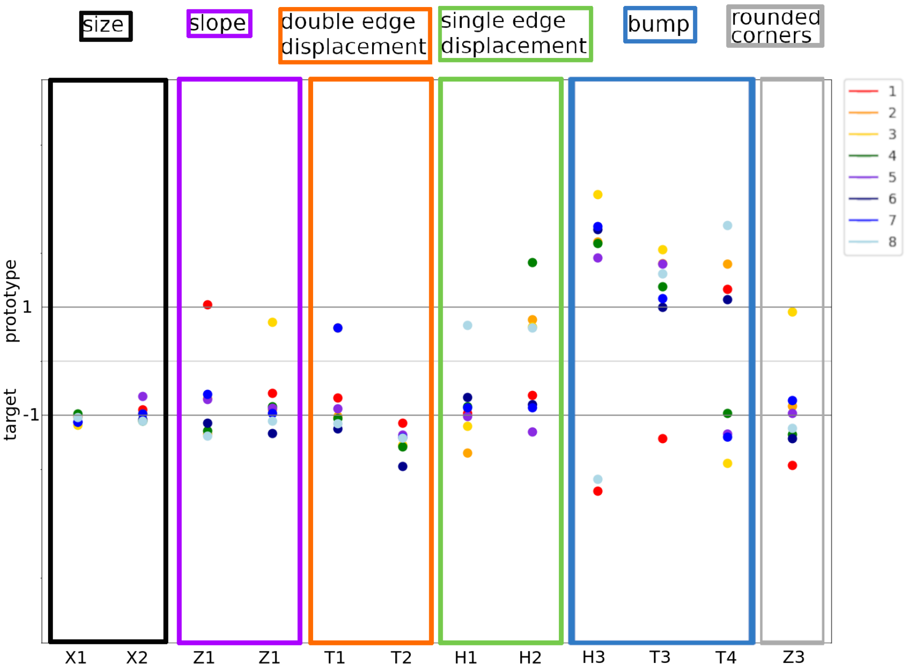
6.3 Discussion

In this study I made an attempt to show the ability of TD individuals to replicate trajectories in specific shapes, and the influence of adding features (modifications) to these shapes on drawing behaviour. If the combination of the experimental design with the selected stimuli yields some generalization behaviour from TD individuals, this can be used as a baseline to compare to behaviour from individuals with ASD in a later experiment that tests both groups. The aim of that experiment is to test the hypo-prior theory by Pellicano and Burr [32] and get more insight into cognitive mechanisms, both in TD individuals and in those with ASD.

In summary, following expectations, participants were good at accurate replication of the Z and T prototypes and worse at the H and X prototypes. In terms



(a) Scores heatmap



(b) Scores per participant

Figure 14: Behaviour-indication scores per trajectory sorted by indicated features. Each circle represents one trajectory drawn by one participant in the testing phase. **a.** Scores with the type of behaviour represented by colour. Gradient on the right indicates type of behaviour indicated by the score. **b.** Same scores as in (a.), now with each participant represented by a colour equal to the colour in Figure 13

of behaviour, participants showed approximately accurate replication on the size feature, and at least some generalization behaviour on the slope feature and the single edge displacement feature. Unexpectedly, participants showed a combination of accurate replication behaviour with exaggeration behaviour on the double edge displacement feature, unexpected behaviour on the bump feature, and a mix of behaviours on the rounded corners feature.

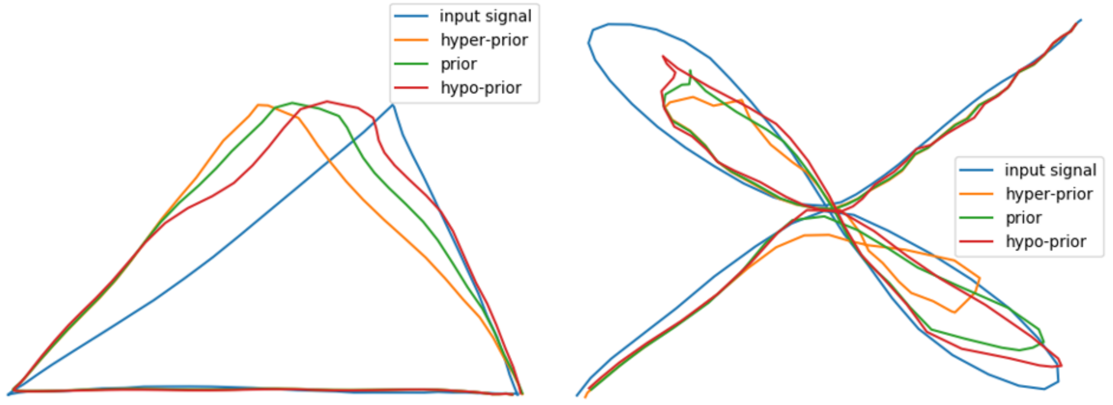
The overall level of generalization behaviour was very low, which may be explained by the setup of the experiment. Generalization can only occur after a prior has been formed out of several stimuli with a certain level of variability. In the familiarization phase participants were expected to form a prior based on the stable points in the shape, which would be represented by the prototype. The experimental setup, however, required for participants to directly draw each trajectory after its first presentation, which may have caused for participants to assess each of the trajectories as a separate shape, blocking generalization tendencies based on common general shape.

Participants that showed generalization behaviour on one stimulus were not more likely to show generalization behaviour on another stimulus. Although we would expect for participants to show a consistent tendency to incorporate priors in terms of their drawing behaviour, the different shapes may have yielded different prior-incorporation tendencies. Moreover, the influence of the features may have been variable, stressing the importance of identification of a feature that yields generalization behaviour in most participants.

Participants weren't very accurate at replicating the H prototype. This is not surprising, since the general shape is quite complicated with several sharp corners and may not directly represent something familiar (the shape looks a bit like the contours of a house with a chimney, but this may not be clear to the participant who sees only one point at a time). What is surprising, however, is the finding that participants were good at replicating the H1 and H2 trajectories (the single edge displacement feature modifications), drawing the corners sharply and the lines straight, whereas the corners in the prototype and the H3 trajectory were drawn rounded. One would expect for all trajectories of the same general shape to at least be drawn similarly, since trajectory presentation was randomized and all participants had drawn the prototype at least once already in the familiarization phase. The difference between the H prototype and the H3 modification, and the H1 and H2 modifications may be caused by participants not assessing these trajectories as belonging to the same general shape (i.e. the H prototype shape), but as separate trajectories. In other words, participants may not have placed the H1 and H2 trajectories in the same implicit category as the H prototype, but in their own separate categories.

There were high levels of unexpected behaviour with the bump feature, that may have been caused by the characteristics of the feature. Most participants noticed, for all three trajectories (H3, T3 and T4), that the trajectory was not the prototype. Nevertheless, they misplaced the bump, changed the size of the bump or its direction, or they were even unable to correctly replicate the general shape with this feature, even though in other drawings they showed they were able to do this (participants were able to correctly replicate the general H shape in the H1 and H2 trajectories, as mentioned above, but not in the H3 trajectory). The high levels of unexpected behaviour may be caused by the size of the feature, as the bump is very small and not enough points were included.

As the goal of the experiment was to identify stimuli that later on can be used in a cognitive mirroring experiment, the stimuli should be replicable by the S-CTRNN, the network that will be used in the cognitive mirroring experiment. To find out if the network can replicate the designed stimuli, I trained several networks using backpropagation with the dataset as shown in Figure 7 with added variance of 0.001 for a minimum of 500 epochs and a maximum of 30000 epochs, and a stopping condition tracking the likelihood that ends training when there is stagnation in improvement. With this small dataset, the network did not show the expected behaviour when tested with different values for the influence of the prior relative to the input signal (see Figure 5, value H). Figure 15a shows that the network is able to differentiate between different prior levels, but the behaviour when tested with a hypo-prior only approaches the input signal, and is still closer to the prototype than the input signal. Therefore the network would not be able to predict accurate replication of the input signal for these stimuli. Figure 15b shows that the network differentiates between the different prior levels from halfway through the trajectory, in the second curl, but it is not able to replicate the input signal correctly, as it should differentiate starting from the first curl. Also note that the H-value for the hypo-prior was set to 1000 (H-values for the hyper-prior and the prior were 0.01 and 1.0, respectively), whereas a value of 5 already showed significant differences when testing the behaviour of the network with other stimuli. This emphasizes the importance of identification of a suitable dataset, both for the network and for the participants.



(a) T1

(b) X2

Figure 15: Examples of network behaviour with different prior levels. Hyper-prior (orange) is expected to closely approximate the prototype, and hypo-prior (red) is expected to closely approximate the input signal (blue). Prior (green) is expected to be somewhere in between the hyper- and hypo-prior and represents what we believe might be a typical prior. **a.** Network behaviour on T1 trajectory shows the hypo-prior approximates the input signal. **b.** Network behaviour on X2 trajectory shows the network does not respond correctly on the first half of the trajectory, but it approximates expected behaviour on the second half.

The difficulty the network has with showing expected behaviour is probably due to an indistinctness of the stable and variable parts of the shapes. This may be caused by a lack of training stimuli, for there was only one stimulus per trajectory in the dataset, and only a small level of variance was added each trial. The network may need multiple stimuli for each of the target trajectories, providing more natural variation in space and, more importantly, in time. The time dimension causes differing dynamics that may obstruct learning for the network. Each of the trajectories has a certain length (the sum of the distances between all consecutive points), but adding a feature may increase or reduce this length. A point in the modified trajectory may now be slightly moved relative to the point representing the same location in the prototype. The network can only learn the differences and commonalities between the target trajectories when there is variation in these dynamics within each trajectory (so over multiple stimuli that all represent one trajectory). Differing dynamics in the time dimension may be the cause of the differentiation between the X shape trajectories only starting from the second (right) curl, instead of the first (left). Finally, another cause for the indistinctness of the stable and variable parts of the shapes may be that all trajectories were drawn by hand. It might have been better if they would have been created computationally. If a computationally created prototype would be the base for each of the modified trajectories, then the features could be added systematically, ruling out factors such as the size of the feature within a feature category, for example the difference in edge displacement between the T1 and T2 trajectories. All stimuli would then have straight lines and equal lengths between timesteps, ruling out the influence of irregularities in the target trajectories and creating high stability in and extra emphasis on the recurring parts of the general shape. However, when testing with trajectories drawn by humans, a network trained with suggested computationally created stimuli may still not be able to accurately replicate them due to differences in dynamics between the human drawings and the computationally created trajectories. Therefore, simply creating more hand-drawn stimuli per trajectory may be a better solution.

7 Conclusion & Recommendations

The goal of the targeted Cognitive Mirroring experiment is to measure the influence of the prior relative to the sensory information in human participants by comparing their behaviour to that of a computational neural network. Because the aim is to investigate whether individuals with ASD have a hypo-prior, it is important that at least TD individuals show generalization behaviour on the stimuli in the experimental task, so it can be used as a baseline to compare behaviour by individuals with ASD to. In this study, the aim was to explore human behaviour in response to different shapes and features, and if possible to identify shapes and features that (may) elicit generalization behaviour from TD individuals.

The research question of the conducted experiment is: Which shapes are replicable by participants, and which feature modifications elicit generalization behaviour in TD participants? The hypotheses are that participants will be good at replication of the Z and T shapes and worse at the H and X shapes, and that participants will show generalization on feature modifications that are less salient (i.e. slope, double edge displacement, single edge displacement, and bump), and accurate replication on the salient feature modifications (i.e. size and rounded corners). The first hypothesis, that participants were better at replicating the Z and T shapes than the H and X shapes, was confirmed. The second hypothesis could be partially confirmed, as participants indeed showed less generalization behaviour on the size feature modification, and some generalization behaviour on the slope, double edge displacement and single edge displacement feature modifications. However, since there was barely any generalization and a lot of unexpected behaviour overall, this hypothesis could not be confirmed in full.

In summary, results indicated that participants showed different types of behaviour that differed widely between the shapes and features, but overall a low level of generalization behaviour was shown, which may be due to the fact that participants had to draw directly after the first stimulus presentation. Moreover, generalization behaviour on one stimulus did not guarantee generalization behaviour on other stimuli, which may improve when the experimental design and stimuli yield a higher level of shown generalization behaviour. Also, participants seem to be worse at drawing the H and H3 shapes than the H1 and H2 shapes, which may be caused by participants assessing the trajectories as belonging to separate categories, instead of placing them in one category. Furthermore, the bump feature yielded very inconsistent results and often unexpected behaviours, which the network won't be able to replicate. And, finally, the network was not able to replicate all the different trajectories, even when tested with an extremely high H-value. This is probably caused by an indistinctness of the variable and stable parts of each of the shapes, which may be caused by a lack of training stimuli, especially with differences in the time dimension, or by having hand-drawn instead of computationally created stimuli.

I would like to conclude by making some recommendations for further experiments based on the results of the experiment.

To start, I recommend **making use of the Z and T shapes, and not the X and H shapes**. Participants clearly showed difficulty with the H and X shapes, as there was high variability between participants and participants drew the prototypes inaccurately. Therefore, it would be impossible to tell whether finding differences, or not being able to find any differences, would be caused by the difficulty of the shape, or by the influence of the prior (a possible hypo-prior) relative to the sensory input. This is particularly important because motor capabilities in ASD might not be fully intact [23].

Second, I recommend **not using the bump feature or the rounded corners feature**. Participants often showed unexpected behaviours on these features, which cannot be replicated by the network.

As discussed in Section 6.3, the low level of generalization behaviour shown by participants may be due to directly drawing after the first presentation of a stimulus. In order to stimulate categorization tendencies, I highly recommend **adding a short practice phase and adjusting the familiarization phase**. The aim of the practice phase would be to make sure participants understand the task and can practice drawing on the screen. The practice phase would be the same as the testing phase, except for a lower number of presented stimuli, and that it would have different target stimuli than the testing phase to ensure practice does not influence experimental results. Thus, the participant would first investigate the trajectory in the left window, and then draw the trajectory in the right window, equal to the testing phase. This should be repeated until it is made sure that the participant understands the task and knows how to draw on the touchscreen monitor, which can be assessed for example by testing similarity between the target trajectory and the drawing by the participant, continuing to the next phase when the similarity is above a certain threshold. Since practice with drawing on the monitor was the main aim of the drawing step in the familiarization phase, and this is also the aim of the practice phase, no drawing is necessary in the familiarization phase when a practice phase is included. This creates the possibility to show more stimuli and thus more variation per target stimulus in the same time span in the familiarization phase, thereby creating more opportunity for participants to implicitly form a prior. Also, not drawing directly after the first presentation of a stimulus may increase generalization behaviour, because there will be less emphasis on individual stimuli.

Moreover, participants were given the opportunity to investigate the trajectories themselves based on a consideration mentioned by Nakano et al [30]: in their study wherein participants saw a visual stimulus behind a slit at predefined speed [31], they did not find better performance for participants with ASD as compared to those with TD, but in the study in which participants could investigate the stimulus at their own pace [30] they did find better performance for participants with ASD as compared to those with TD, even though there was no significant difference between the total time spent on investigation per stimulus for the two groups. It was argued that this difference may be caused by the difference in the participants' own influence on investigation of the stimulus. In the current experiment, however, participants always chose the option to 'play' the sequence at a predefined speed, instead of investigating the trajectories step by step by themselves. This may be caused by the length of the trajectories, which could be shortened, but I would suggest to rather **adjust or remove the 'play'-function** instead. Adjustment of the function could mean adjusting the set speed of the sequence presentation, or it could mean letting participants change the speed themselves while watching the sequences, so they can investigate specific parts more thoroughly if necessary. It would then also be interesting to compare TD individuals to those with ASD on whether and, if so, on which parts of the trajectories they increase or decrease the presentation speed. Removing the 'play'-function would force participants to decide on the presentation speed by hand, because they have to press a button for every next time step in the sequence. However, when removing the 'play'-function, I highly suggest shortening the trajectories to at least half the length, because otherwise the task becomes tedious (pressing a button 70 times to investigate one trajectory once can be tiring) and participants may just try to get through the trajectories as fast as possible, paying attention to getting through the experiment faster more than to the trajectories.

Additionally, I would like to give a suggestion for stimulus creation. As mentioned in Section 6.3, there was not enough variability in the dataset for the network to learn to replicate the modified trajectories, but there may also not have been enough stability in the dataset to learn the stable parts for each of the general shapes, which both may have been caused by a shortage in stimuli. Therefore, I suggest **increasing the number of stimuli for each target trajectory in the dataset** by letting several people draw the trajectories, both the prototypes and the target trajectories, by hand on the monitor. This would increase variability in the dataset, emphasize the stable parts of the general shapes, and reduce the influence of the time dimension by increasing variability in terms of dynamics. A contrasting suggestion would be to create this variability computationally, keeping the stable parts perfectly stable and systematically varying the intensity of the features and dynamics related to the time-dimension. Hand-created stimuli may be preferable, however, since they are more likely to have comparable dynamics to the drawings by the participants.

Next, I suggest **adding a recognition phase** at the end of the experiment to confirm that potentially found effects in behaviour are also present in recognition, and that they are not due to problems with motor skills. This phase would be similar to the experiment by Nakano et al [30], but instead of haptic information, only visual information is used, namely by letting participants visually investigate trajectories in the left window, equal to the testing phase, but instead of drawing the trajectory themselves, participants are instructed to choose the investigated trajectory from three slightly different trajectories. This way, influence of drawing skills can be ruled out, by comparing results from the testing phase for a specific stimulus to the results of the recognition phase for this same stimulus. For example, a participant may have shown exaggeration behaviour in the testing phase, but correct target recognition in the recognition phase, which suggests the exaggeration behaviour is caused by poor drawing skills. The recognition phase would be used to assess the validity of the drawing experiment: if the drawings by participants are in line with the trajectories chosen in the recognition phase, it can be assumed the drawings represent how participants perceive the stimuli.

Finally, to increase generalization tendencies, I recommend considering including a simple but time-consuming distraction in between the investigation of the trajectory and the drawing, including a task with high cognitive load during or after the investigation of the trajectory, or during drawing, or masking discrete features by more salient features.

References

[1] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (4th ed., Text Revision).* Washington, DC: Author, 2000.

[2] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (5th ed.).* Arlington, VA: Author, 2013.

[3] Bailey, A., Palferman, S., Heavey, L. & Le Couteur, A. “Autism: the phenotype in relatives”. In: *Journal of Autism and Developmental Disorders* 28 (1998), pp. 369–392.

[4] Baron-Cohen, S., Bolton, P., Wheelwright, S., Scahill, V., Short, L., Mead, G. & Smith, A. “Autism occurs more often in families of physicists, engineers, and mathematicians.” In: *Autism* 2 (1998), pp. 296–301.

[5] Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J. & Clubley, E. “The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians”. In: *Journal of Autism and Developmental Disorders* 31.1 (2001), pp. 5–17.

[6] Baron-Cohen, S., Wheelwright, S., Stott, C., Bolton, P. & Goodyer, I. “Is there a link between engineering and autism?” In: *Autism* 1 (1997), pp. 153–163.

[7] Baum, S.H., Stevenson, R.A. & Wallace, M.T. “Behavioral, Perceptual, and Neural Alterations in Sensory and Multisensory Function in Autism Spectrum Disorder”. In: *Progress in Neurobiology* 134 (2015), pp. 140–160.

[8] Bayliss, A. P. & Kritikos, A. “Brief report: Perceptual load and the autism spectrum in typically developed individuals.” In: *Journal of Autism and Developmental Disorders* 41.11 (2011), pp. 1573–1578.

[9] Blakemore, S. J., Tavassoli, T., Calò, S., Thomas, R. M., Catmur, C., Frith, U., & Haggard, P. “Tactile sensitivity in Asperger syndrome”. In: *Brain and cognition* 61.1 (2006), pp. 5–13.

[10] Brown, D.J. & Proulx, M.J. “Increased Signal Complexity Improves the Breadth of Generalization in Auditory Perceptual Learning”. In: *Neural Plasticity* 2013 (2013).

[11] Cabibihan, J.J., Javed, H., Ang, M. & Aljunied, S.M. “Why Robots? A Survey on the Roles and Benefits of Social Robots in the Therapy of Children with Autism”. In: *Interational journal of social robotics* 5.4 (2013), pp. 593–618.

[12] Cascio, C., McGlone, F., Folger, S., Tannan, V., Baranek, G., Pelphrey, K. A., & Essick, G. “Tactile perception in adults with autism: a multidimensional psychophysical study.” In: *Journal of autism and developmental disorders* 38.1 (2008), pp. 127–137.

[13] Cribb, S. J., Olaithe, M., Di Lorenzo, R., Dunlop, P. D. & Maybery, M. T. “Embedded figures test performance in the broader autism phenotype: a meta-analysis.” In: *Journal of Autism and Developmental Disorders* 46.9 (2016), pp. 2924–2939.

[14] Dudova, I., & Hrdlicka, M. “Olfactory functions are not associated with autism severity in autism spectrum disorders.” In: *Neuropsychiatric disease and treatment* 9 (2013), p. 1847.

[15] Friston, K. "The free-energy principle: a unified brain theory?" In: *Nature reviews neuroscience* 11.2 (2010), p. 127.

[16] Frith, U. *Autism: Explaining the enigma.* Oxford: Blackwell Publishing, 1989.

[17] Froehlich, A. L., Anderson, J. S., Bigler, E. D., Miller, J. S., Lange, N. T., DuBray, M. B., Cooperrider, J. R., Cariello, A., Nielsen, J. A. & Lainhart, J. E. "Intact prototype formation but impaired generalization in autism." In: *Research in autism spectrum disorders* 6.2 (2012), pp. 921–930.

[18] Fugard, A.J., Steward, M.E. & Stenning, K. "Visual/verbal-analytic reasoning bias as a function of self-reported autistic-like traits: A study of typically developing individuals solving Raven's Advanced Progressive Matrices." In: *Autism* 15.3 (2011), pp. 327–340.

[19] Fukuyama, H., Qin, S., Kanakogi, Y., Nagai, Y., Asada, M., & Myowa-Yamakoshi, M. "Infant's action skill dynamically modulates parental action demonstration in the dyadic interaction." In: *Developmental science* 18.6 (2015), pp. 1006–1013.

[20] Gastgeb, H. Z., & Strauss, M. S. "Categorization in ASD: The role of typicality and development." In: *Perspectives on language learning and education* 19.2 (2012), pp. 66–74.

[21] Happé, F. & Frith, U. "The weak coherence account: detail-focused cognitive style in autism spectrum disorders." In: *Journal of Autism Developmental Disorders* 36 (2006), pp. 5–25.

[22] Happé, F.G. "Central coherence and theory of mind in autism: reading homographs in context". In: *British Journal of Developmental Psychology* 15.1 (1997), pp. 1–12.

[23] Haswell, C. C., Izawa, J., Dowell, L. R., Mostofsky, S. H., & Shadmehr, R. "Representation of internal models of action in the autistic brain." In: *Nature neuroscience* 12.8 (2009), pp. 970–972.

[24] Hrdlicka, M., Vodicka, J., Havlovicova, M., Urbanek, T., Blatny, M., & Dudova, I. "Brief report: significant differences in perceived odor pleasantness found in children with ASD." In: *Journal of Autism and Developmental Disorders* 41.4 (2011), pp. 524–527.

[25] Kumagaya, S. *An invitation to Tojisha-Kenkyu.* URL: `https://www.u-tokyo.ac.jp/en/whyutokyo/indpt_tojisha_018.html`. (accessed: 18.07.2019).

[26] Kuniyoshi, Y., Ohmura, Y., Terada, K., Nagakubo, A., Eitoku, S., & Yamamoto, T. "Embodied basis of invariant features in execution and perception of whole-body dynamic actions—knacks and focuses of Roll-and-Rise motion." In: *Robotics and Autonomous Systems* 48.4 (2004), pp. 189–201.

[27] Lawson, R. P., Rees, G., & Friston, K. J. "An aberrant precision account of autism." In: *Frontiers in human neuroscience* 8 (2014), p. 302.

[28] Murata, S., Namikawa, J., Arie, H., Sugano, S., & Tani, J. "Learning to reproduce fluctuating time series by inferring their time-dependent stochastic properties: Application in robot learning via tutoring." In: *IEEE Transactions on Autonomous Mental Development* 5.4 (2013), pp. 298–310.

[29] Nagai, Y. *Self-Understanding of Cognition.* URL: `http://cognitive-mirroring.org/en/`. (accessed: 09.07.2019).

[30] Nakano, T., Kato, N., & Kitazawa, S. "Superior haptic-to-visual shape matching in autism spectrum disorders." In: *Neuropsychologia* 50.5 (2012), pp. 696–703.

[31] Nakano, T., Ota, H., Kato, N., & Kitazawa, S. "Deficit in visual temporal integration in autism spectrum disorders." In: *Proceedings of the Royal Society B: Biological Sciences* 277.1684 (2010), pp. 1027–1030.

[32] Pellicano, E., & Burr, D. "When the world becomes 'too real': a Bayesian explanation of autistic perception." In: *Trends in cognitive sciences* 16.10 (2012), pp. 504–510.

[33] Philippsen, A., & Nagai, Y. *Understanding the cognitive mechanisms underlying autistic behavior: a recurrent neural network study.* in Proceedings of the 8th IEEE International Conference on Development, Learning, and on Epigenetic Robotics, 2018, September, pp. 84–90.

[34] Piven, J., Palmer, P., Jacobi, D., Childress, D. & Arndt, S. "Broader Autism Phenotype: Evidence From a Family History Study of Multiple-Incidence Autism Families." In: *American Journal of Psychiatry* 154.2 (1997), pp. 185–190.

[35] Plaisted, K., O'Riordan, M., & Baron-Cohen, S. "Enhanced discrimination of novel, highly similar stimuli by adults with autism during a perceptual learning task." In: *The Journal of Child Psychology and Psychiatry and Allied Disciplines* 39.5 (1998), pp. 765–775.

[36] Puts, N. A., Wodka, E. L., Tommerdahl, M., Mostofsky, S. H., & Edden, R. A. "Impaired tactile processing in children with autism spectrum disorder". In: *Journal of neurophysiology* 111.9 (2014), pp. 1803–1811.

[37] Remington, A., Swettenham, J., Campbell, R. & Coleman, M. "Selective attention and perceptual load in autism spectrum disorder". In: *Psychological Science* 20.11 (2009), pp. 1388–1393.

[38] Scheeren, A.M. & Stauder, J.E. "Broader autism phenotype in parents of autistic children: reality or myth?" In: *Journal of Autism and Developmental Disorders* 38 (2008), pp. 276–287.