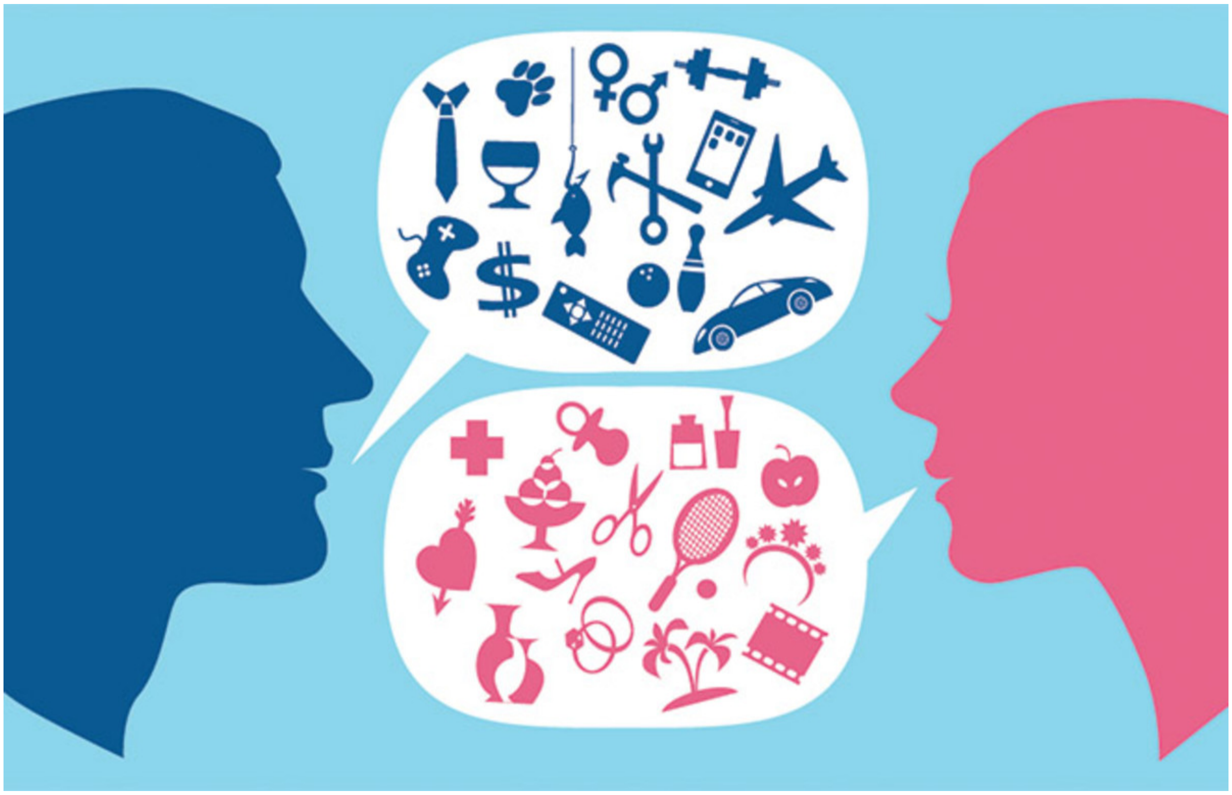
Stereotype formation



Vincent Huf

Layout: typeset by the author using LaTeX.
Cover illustration: Unknown artist

Stereotype formation

The influence of limited memory capacity on the formation of novel stereotypes

Vincent Huf
12382108

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Dr. K. Schulz

Institute for Logic, Language and Computation
Faculty of Science
University of Amsterdam
Science Park 107
1098 XG Amsterdam

Semester 2, 2022

Abstract

The origin of stereotypes has been widely discussed over the years. Multiple theories mark cognitive biases as reasons for the formation of stereotypes. However, it remains unclear to what extent humans' limited memory capacity causes the formation of novel stereotypes. This thesis aims to investigate to what extent experimental results on stereotype formation can already be explained by this. The current research provides a cognitive model that successfully implements an iterative learning method to simulate a laboratory-based experiment where information is repeatedly passed from person to person. By only implementing limited memory capacity, this model shows how novel stereotypes emerge even without other cultural and cognitive biases present.

Contents

Chapter 1

Introduction

Stereotypes are characterizations of social categories whereby group membership is associated with the possession of certain attributes [18]. While many stereotypes are based on some level of empirical truth, they may lead to discrimination against minorities and the incorrect categorization of individuals when these associations are exaggerated. For example, gender stereotypes ("Women are more emotional than men, so they are not suited to be a leader") can lead to sexism.

Stereotyping can be seen as a tool to organize and process the information around us [3, 10]. Since humans cannot perceive and recall all social information they encounter, categorization offers a mechanism to reduce this information flow [19]. Often, this happens intuitively in everyday life. Even if people reject the idea of stereotyping, simply knowing about the content of stereotypes is enough to create biased thoughts [6]. Other than its function to efficiently store and recall information, research has focused on explaining the origin of stereotypes through multiple cognitive biases. An example of this is a communication bias. When participants in an experiment discussed characteristics of an out-group, their conversations were rated as more stereotypic than their in-group discussions [9]. Moreover, people tend to favor communicating stereotype consistent information over stereotype inconsistent information [17].

In 2014, Martin et al. studied how stereotypes are formed *across* individuals, instead of within. In their experiment, information was transmitted from individual to individual. During this process, the information became increasingly structured and simplified, eventually leading to novel stereotypes [19]. Based on these results, they suggest that the formation of novel stereotypes across generations is caused by limited memory capacity and a shared bias towards, among others, sense-making, internal consistency, and categorical structure [18, 19]. These studies demonstrate how the origin of stereotypes has been attributed to many different factors. However, in this research, we want to investigate to what extent experimental results on stereotype formation can already be explained as due to limited memory capacity. If all other cultural and cognitive biases and assumptions are removed, and only the pressure of memorization is present, are novel stereotypes still formed? Therefore, the following research question is proposed:

To what extent can humans' limited memory capacity explain the spontaneous formation of novel stereotypes?

To answer this question, a cognitive model will simulate the laboratory-based experiment of Martin et al. (2014). This model does not implement cultural biases but instead only focuses on humans' memory capacity. The first chapters of this thesis are dedicated to the theoretical foundations of the experiment and the mechanisms of the cognitive model. Later on, the model's results are analyzed and compared to the original results, eventually providing an answer to the research question.

Chapter 2

Theoretical framework

2.1 The spontaneous formation of novel stereotypes - state of the art

As information is being passed from individual to individual, its content changes incrementally [11]. An example of this change is how stereotypes form when cultural information is transmitted across generations. Since early research mainly focused on understanding the nature and influence of stereotypes, Martin et al. (2014) decided to investigate *how* spontaneous and novel stereotypes form. They propose an experimental setup with serial reproduction chains to observe whether novel stereotypes emerge as information is passed down this chain.

The method of serial reproduction has long been used as a tool in psychology to examine how information transforms as it is transmitted serially from individual to individual in a communication chain [8, 12, 18]. Often, information changes drastically as a chain progresses, including unintended transformations [12]. A serial reproduction chain is similar to the 'whisper game' often played by children, where one person whispers a sentence to the next, who then whispers it to the next, and so on. The last person then reveals what they think the original sentence is. This method is useful for observing the dynamics of cultural evolution, such as the formation and transformation of stereotypes. One of the first researchers to investigate this topic is Bartlett (1932). In his study, participants are asked to remember and reproduce a mythical American Indian story. As the chain progresses, Bartlett observes how certain mythical and unfamiliar story elements turn into traditional western story elements that are more familiar to the participants (for instance, the word 'canoe' is transformed into the word 'boat'). Furthermore, the story becomes shorter and more coherent with each reproduction, as details that are not grouped about a central incident of the story are omitted [2]. Bartlett (1932) calls this phenomenon 'conventionalization', where information inconsistent with one's own beliefs and knowledge is transformed into familiar and stereotype-consistent information. This tendency to favor reproducing stereotype-consistent information has been established more often in social experiments [8, 16]. Even though these experiments focus on already existing stereotypes, the fundamental aspects of serial reproduction chains are also useful for studying the formation of *novel* stereotypes.

Martin et al. (2014) expect that during their experiment, information becomes continuously simpler and more structured through a process called cumulative cultural evolution (CCE). The concept of CCE was first brought up as a way to demonstrate how human culture differs from non-human culture [4]. Over time, CCE has been widely accepted as a theory stating that the evolution of human culture is analogous to biological evolution [18]. In other words, natural selection plays a central role in the evolution of human culture. Modifications of behavior or objects that are not an improvement will go extinct, and those that prove to be adaptive will remain. This leads to a long-lasting cultural change. A prime example where the theory of CCE can apply is the evolution of languages [7, 13, 14, 15]. As a fundamental part of human culture, language continuously transfers between individuals as a serial reproduction chain: people observe linguistic behavior, modify it to their advantage, and produce a new form for others to learn from. The application of CCE is not only suitable for studying the evolution of language. The formation and transformation of stereotypes have many similar aspects to language evolution. For instance, grammar rules allow one to correctly conjugate words of which the meaning is unknown [15, 22]. Similarly, stereotypes allow people to categorize other people they have not seen before based on established rules (e.g., all gamers are nerdy), making remembering them easier. So, both exhibit a structure which allows people to make rule-based inferences when encountering unknown information. For this reason, Martin et al. (2014) expect that through a process of CCE novel stereotypes will emerge in their experiment.

2.2 The alien experiment

For the experiment, Martin et al. (2014) create 27 novel 'aliens', each consisting of three different dimensions: color, shape and movement. Each dimension then has three possible features to assign to an alien (for example, the colors green, red and blue). Next, each alien is assigned six personality attributes which are randomly selected from a pool of 48 possible attributes (Figure 2.1).

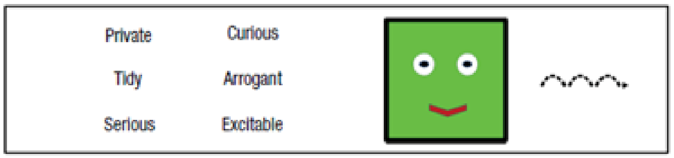
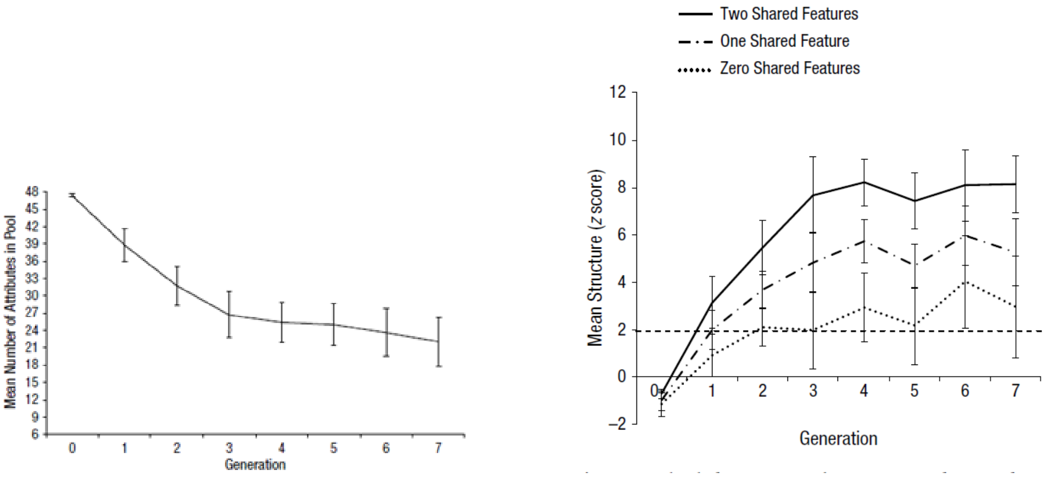


Figure 2.1: An example of an alien and its attributes (Martin et al. 2014)

These alien-attribute combinations are used as 'Generation 0' in the linear diffusion chain, with 13 combinations randomly selected as a training set while the other 14 are used as a test set. During the training phase, the first participant (Generation 1) is tasked to remember the attributes assigned to each alien from the training set. After this, the subsequent test phase requires the participant to correctly assign the original attributes to each of the 27 aliens, meaning that both seen and unseen aliens are presented. The results from this participant are then used as the data for the next participant (Generation 2). Once again, a random training set of 13 aliens is selected from the data, and the participant performs the same actions as the previous participant. This process is repeated for seven generations. The participants are unaware that their answers are used as data for the next participant [19].

Martin et al. (2014) find that cultural stereotypes form spontaneously when social information is passed down a chain. After each generation, correctly assigning attributes to the aliens becomes simpler as the total number of attributes decreases (figure 2.2a). More specifically, because of this decrease, later generations do not have to remember as many attributes as earlier ones. This makes it easier to to remember an alien and its attributes correctly. Besides the decrease in the total number of attributes, structure scores increase (figure 2.2b). Structure scores show the amount of overlap in attributes between two aliens. So, a high mean structure score for aliens with two shared features indicates much overlap in attributes between aliens who, for example, share the same color and shape. In other words, as the chain progresses, aliens who share features become more likely to share attributes as well, making it easier to recall and predict aliens' attributes correctly (figure 2.3). For instance, aliens sharing the color green were predominantly associated with being 'vulgar' [19]. From this, Martin et al. (2014) conclude that the process of cumulative cultural evolution can explain the formation of novel stereotypes that are of no obvious origin [19].



(a) Mean number of attributes ascribed to aliens at each generation

(b) Structure scores per number of shared features

Figure 2.2: Mean number of attributes and structure scores of the experiment by Martin et al. (2014)

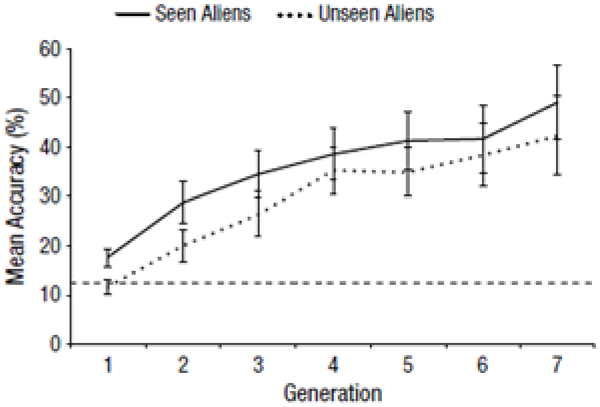


Figure 2.3: Mean accuracy of each generation. Accuracy is denoted as the percentage of attributes correctly assigned to the aliens based on the results of the previous generation (Martin et al. 2014)

Based on these results, Hutchison et al. (2018) perform further research to investigate how context and category salience influence the formation of novel stereotypes. For this, they replicate the original experiment and calculate the structure scores per category feature instead of per number of shared features. So, in this case, structure scores denote the amount of overlap between aliens who share a specified feature (color, shape or movement). They observe a bias towards forming stereotypes around the color dimension (Figure 2.4). In a follow-up study, they observed that when the color dimension was made less salient, this bias disappeared, and all three dimensions became equally important [11]. The importance of salience has been noted in earlier research, which centered around learning a new language through CCE [25]. These findings suggest that participants might remember aliens' attributes based on their different features (e.g., color or shape) instead of the alien as a whole. This is in line with research by Shephard et al. (1961) , who state that 'For stimuli varying along a given number of dimensions, the easiest classification is the one in which the value on a single dimension completely determines which of the two classificatory responses is appropriate'. In other words, when learning, people prefer to discriminate objects on as few dimensions as possible. This theory will later be implemented in one of the models for the current research.

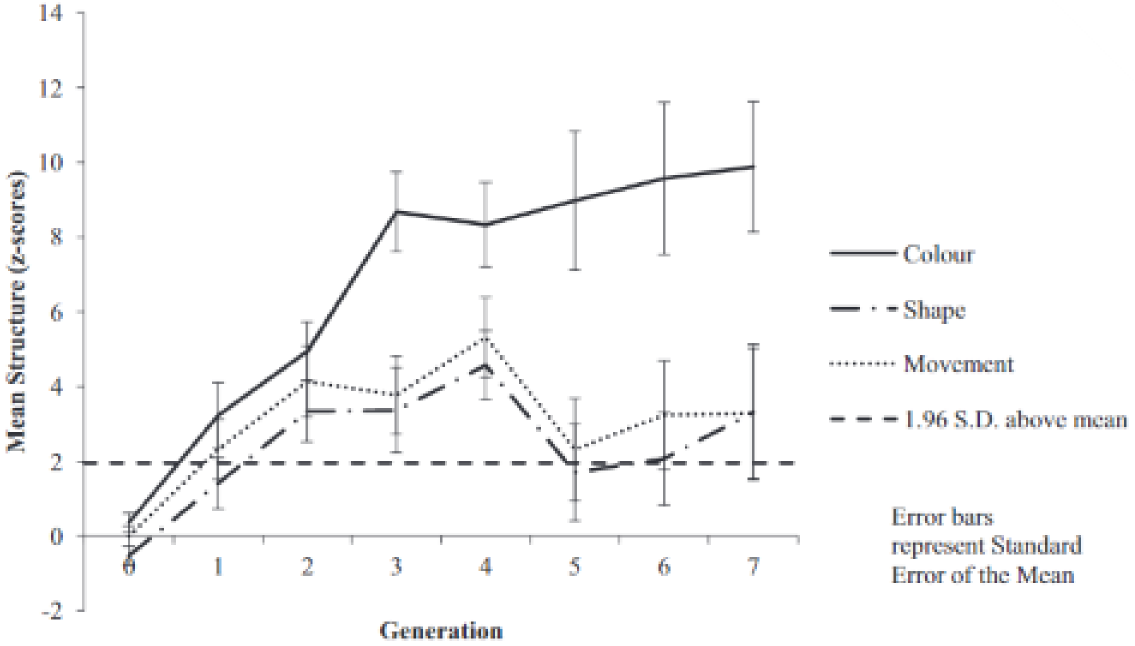


Figure 2.4: Mean level of structure at each generation by category feature. Increasing positive z-scores indicate increasing structure (Hutchison et al., 2018)

2.3 Research gap

Martin et al. (2014) suggest that the emergence of novel stereotypes in their experiment is due to three elements: the limited memory capacity of the participants, a bias towards categorical structure, and a bias for within-category consistency. The first bias assumes that humans tend to further strengthen patterns of structure found in data. So if a participant senses that some aliens with the same category feature also share the same attribute, they tend to assign this attribute to other aliens with that feature. The second bias means that people overestimate the number of shared attributes for aliens sharing multiple features, thus incorrectly assigning attributes based on an alien's feature. In the experiment of Martin et al. (2014), the biases become apparent in both recalling aliens' attributes from the training set and predicting unseen aliens' attributes in the test set. For seen aliens from the training set, attributes are categorized around shared features. Categorized information is easier to remember since fewer unique attributes are needed [5, 16, 19, 24]. Because there are fewer attributes through the process of categorization, accuracy increases for correctly recalling the training set as the chain progresses. Moreover, associations between specific attributes and aliens' features become so strong over time, that participants can make correct predictions about unseen aliens. This is caused by participants' overestimation of the within-category similarity of the aliens [19, 21]. To summarize, Martin et al. (2014) suggest a strong bias towards favoring the categorization of aliens with shared features, causing the emergence of novel stereotypes.

However, the role of people's limited memory capacity is only briefly touched upon, as Martin et al. (2014) primarily focus on the categorization bias. Multiple studies, nevertheless, indicate that limited recall influences the amount and content of information when passed from individual to individual through serial reproduction [5, 16, 18, 23, 24]. In further research from Martin et al. (2017), there is a more considerable emphasis on people's cognitive limitations by depicting how they create 'a bottleneck of social transmission'. The role of memory is briefly discussed, as the study states that people are constrained in the quantity of information that they can transmit to others, thus making them highly selective in the information they choose to transmit [18]. Therefore, they often decide to transmit stereotype consistent, simplified, and categorically structured information using easily learnable systems.

Thus, it has not yet been demonstrated *to what extent* limited memory capacity can explain the spontaneous formation of novel stereotypes via social transmission. This research, therefore, aims to examine whether limited memory capacity alone leads to categorization. This objective has been specified in the following research question:

To what extent can humans' limited memory capacity explain the spontaneous formation of novel stereotypes?

To answer the question, two cognitive models are developed to simulate the alien experiment, after which their results are compared to those of Martin et al. (2014). Iterated learning models provide a suitable mechanism for simulating processes involving CCE, like the formation of stereotypes. For example, Griffiths and Kalish (2007) use an iterated learning model which includes the principles of Bayesian inference to examine the consequences of iterated learning on learning algorithms. The foundation of this model is introduced by Kirby (2002) to examine the evolution of language, a process previously noted for having similarities to the evolution of stereotypes. The basic idea behind Bayesian inference is that agents have certain biases encoded as a prior probability distribution which is influenced by observing data [7, 13]. When applying this method to the alien experiment, agents have a prior probability of assigning particular attributes to an alien before actually seeing it. This prior can be seen as a stereotype held by the agent (e.g., red aliens are kind). After seeing the alien, the agent generates a posterior probability, which puts forward the probability of assigning an attribute to the observed alien. Thus, an agent constantly updates their belief after seeing data. Bayesian iterative learning models have been successfully used in models centered around cognitive studies other than language learning, making it a suitable theoretical foundation for recreating the serial reproduction chains (Girffiths & Kalish, 2007). While the cognitive models of this research do not explicitly use Bayesian inference, they will be based on the assumptions of prior probabilities for assigning attributes to aliens. The following chapter will explain the design behind the models and their implementation in more detail.

Chapter 3

Method

3.1 Introduction

For this research, two different models are proposed, each implementing a different method of learning and reproducing the alien-attribute combinations from Martin et al. (2014). Their main difference is that the first model uses the premise of learning and reproducing an alien and its attributes as a whole (e.g., all three alien features together are associated with the attributes). In contrast, the second model assumes that the aliens' features are decomposed and only one feature is focused on when remembering the attributes. The expectation is that the second model has a stronger tendency towards more simplicity and categorization, as its learning strategy requires less memory capacity and generalizes the aliens. From now on, the first model will be referred to as M1, and the second model as M2. The research goal is to examine whether either of the models is able to reproduce the results from the experiment by Martin et al. (2014), thereby confirming the following hypothesis:

> **Hypothesis 1:** *The spontaneous formation of novel cultural stereotypes can be explained solely by human's limited memory capacity.*

Based on this hypothesis and the fact that limited memory capacity requires a method focusing on simplicity, a second hypothesis is formed which focuses on the models' performance.

> **Hypothesis 2:** *M2 generates results that are more similar to results of the experiment by Martin et al. (2014) than M1 does.*

Similarly to Martin et al. (2014), accuracy scores, the total number of attributes and structure scores are calculated to review the models' performance. Furthermore, both the internal working of the model and its results compared to those of Martin et al. (2014) are statistically analyzed. Finally, structure scores per category feature are compared to those found by Hutchison et al. (2018). The following section explains both models and their implementation in more detail. First, the creation of Generation 0 is demonstrated, which is identical for M1 and M2. Then, for each model both the training and the reproduction phase are explained. These phases repeat for each generation in the chain. In the training phase, the models simulate the cognitive process of remembering the aliens and their attributes. The reproduction phase consists of assigning new attributes to the aliens, based on the remembered attributes from the training phase. Similar to the original experiment, each chain has seven generations. To prevent outliers, the mean results of 40 chains are used as final results.

3.2 Creating the starting data

Before starting the serial reproduction chain, Generation 0 is created. All 27 aliens are built based on the possible category features, and six random attributes are ascribed to each of them. This is done in a Python Class for efficiently creating a serial reproduction chain. All aliens and their attributes are then stored in a dictionary that looks as follows:

```
{(color, shape, movement): [att.1, att.2, att. 3, att. 4, att. 5,
    att. 6]}
```

Listing 3.1: Alien storage in a dictionary

Next, this dictionary is split into a training set of 13 aliens and a test set of the remaining 14 aliens. The central idea for both models is that during the following training phase, an alien from the training set is presented three times, and each time N amount of attributes are remembered and stored in a new dictionary. This mirrors the experiment from Martin et al. (2014), where each alien was shown three times to the participant. Finally, in the reproduction phase, all aliens are presented one by one. Six new attributes are taken from the dictionary with the attributes from the training phase, each one sampled under specific conditions depending on the model. After this process, a new generation of aliens and attributes is created, which is then used as input for the next iteration.

3.3 Model 1

As stated in the introduction of this chapter, M1 assumes that participants review an alien in its entirety; color, shape and movement are all taken into account. This means that three different dimensions are remembered in combination with the associated attributes. Thus, a unique representation of each alien from the training set is remembered. Since this learning strategy has proven to be more difficult since it requires maximum memory capacity (Medin & Smith, 1981; Shepard, Hovland & Jenkins, 1961) , the model stores a low number of attributes associated with these aliens to compensate for this. In the following two subsections, the training and reproduction phase for M1 will be explained.

3.3.1 The training phase

Since there is no proof for the number of attributes the participants remember during the original experiment, M1 runs simulations for two scenarios: remembering one attribute per alien and remembering two attributes per alien. This value is indicated by N. Besides this variable, both simulations are exactly the same. Thus, for each alien in the training set, N random attributes are remembered and added to a dictionary containing all possible aliens as keys. Before the learning starts, this dictionary has no values, but as the training phase progresses, remembered attributes are added as values to the corresponding keys. As in the original experiment, this process repeats itself three times per alien (Listing 3.2).

```
N = 2
for key in formed_alien_train.keys():
    attributes = formed_alien_train[key]
    for _ in range(3):
        self.remember[key].append(sample(attributes, N))
```

Listing 3.2: The training phase of model 1 with N

3.3.2 The reproduction phase

After the training phase, there now exists a new dictionary containing each alien seen in the training phase, together with six remembered attributes. Since the attributes were sampled at random, there may be duplicates. In the reproduction phase that follows, the algorithm iterates over each alien. First, if the alien was seen in the training phase, attributes from that alien are sampled. Next, the alien is compared to all aliens from the training phase that share two category features (e.g., color and movement). Further attributes are sampled from those aliens. If there are still not enough attributes sampled, the same process is applied to aliens from the training phase sharing only one category feature. Last of all, if necessary, random attributes are assigned to the alien to achieve a total of six different attributes. As can be seen in figure 3.1, this method uses a hierarchical structure. When recalling attributes, the first samples are of the alien as seen during the training phase. If the alien is from the test set, this phase is omitted. Following this are samples of aliens from the training set who share two features, one feature and random attributes, respectively. This method is meant to model how humans find it easier to classify when the number of differences in their category dimensions is lower [20, 21, 24].

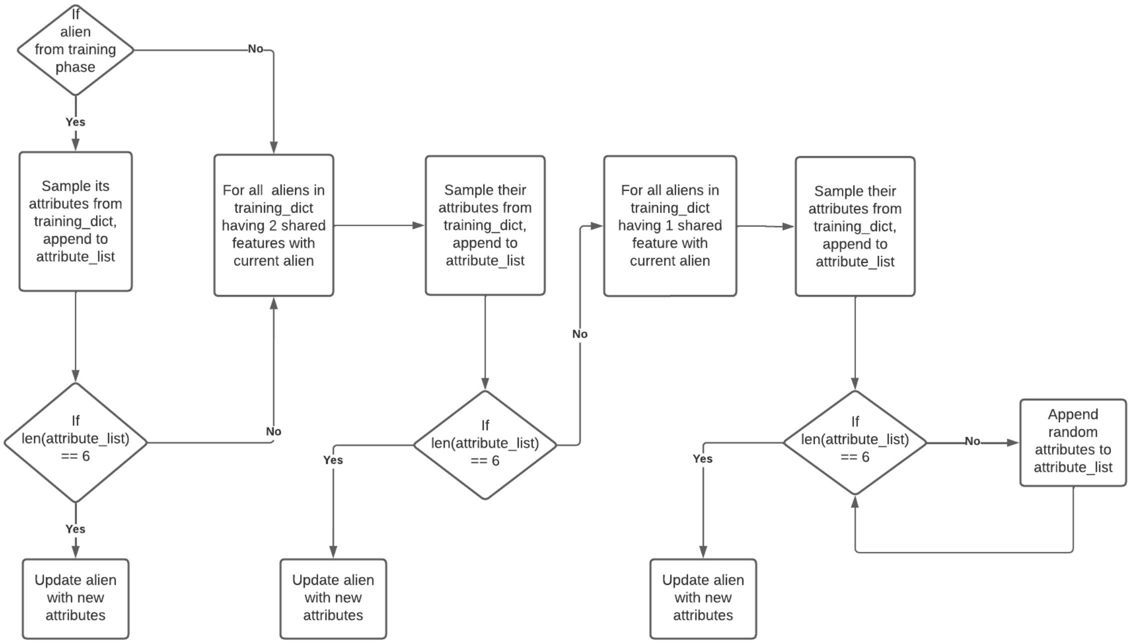

If alien from training phase
No
Yes
Sample its attributes from training_dict, append to attribute_list
For all aliens in training_dict having 2 shared features with current alien
Sample their attributes from training_dict, append to attribute_list
For all aliens in training_dict having 1 shared feature with current alien
Sample their attributes from training_dict, append to attribute_list
No
If len(attribute_list) == 6
No
Yes
If len(attribute_list) == 6
Yes
If len(attribute_list) == 6
Yes
No
Append random attributes to attribute_list
Yes
Update alien with new attributes
Update alien with new attributes
Update alien with new attributes

Figure 3.1: Training phase of M1 per alien, simplified

3.4 Model 2

The second model applies the learning strategy of remembering an alien's attributes based on one category feature. This strategy is based on the findings from

Shepard et al. (1961) and Hutchison et al. (2018). As described in chapter 2, the former found that classification is easiest when it is determined by the value of a single dimension, and the latter suggest that participants in the alien experiment form categories around the color dimension. Based on these findings, M2 is expected to perform more accurate than M1. The following two sections will explain the training and reproduction phase for M2 in more detail.

3.4.1 The training phase

Once again, this model starts with a training set and a test set. As in the previous model, N random attributes are remembered per alien from the training set and added to a new dictionary with the corresponding key. However, instead of the alien that is presented, the key in this dictionary is now one of the category features from that alien. This way, the learning strategy of learning per category feature is recreated. Since this method requires less memory capacity, the value of N is raised to 3. The results from Hutchison et al. (2018) touching on the influence of category salience on the formation of stereotypes are used to decide which category feature is chosen as a key. For instance, Hutchison et al. (2018) found that people are more likely to categorize around color than around shape or movement, so it is most likely that an aliens' color is used as a key. This is implemented by normalizing the mean structure scores of the category dimensions from their research, and then using them as probability scores per feature. This results in the following probabilities:

[color: 0.553, shape: 0.197, movement: 0.250]

The bias towards forming stereotypes around the color dimension is clearly represented in these probabilities. Now, each time an alien is presented, a category feature is sampled based on these probabilities, after which N attributes are sampled and stored with that specific feature as a key (Listing 3.3).

```
N = 3
for key in formed_alien_train.keys():
    attributes = formed_alien_train[key]
    for _ in range(3):
        choice = np.random.choice(keyset, 1, p = key_norm)
        self.remember[key[keuze]].append(sample(attributes, N))
```

Listing 3.3: The training phase of model 2. 'keyset' represents the possible category features and 'key_norm' the probabilities per feature.

An example of how this dictionary may look like is given in Listing 3.4. Since not every category feature always appears in the training set, there is the possibility of having empty keys in the dictionary. If this is the case, six random attributes are added as a value. This is to make sure that every category feature has attributes to sample in the reproduction phase.

```
self.remember = {(blue): ['Funny', 'Arrogant', 'Tidy'], (square):
    ['Friendly', 'Shy', 'Cautious'], (red): []}
```

Listing 3.4: Dictionary containing the remembered attributes per feature.

3.4.2 The reproduction phase

During the reproduction phase, the model uses the probability scores of the category features once more. Similar to how a feature is sampled in the training phase, it is now sampled for retrieving attributes. In other words, a feature is first selected from the current alien, based on the probabilities per category as defined in the previous section. Next, an attribute is sampled from that specific feature. This happens six times so as to form a complete alien-attribute combination. To avoid duplicates, the model keeps on sampling attributes from new category features if all attributes from the sampled feature are already chosen. If all attributes from the three category features of the current alien are chosen and a total of six different attributes is not yet sampled, random attributes are sampled until it is reached. Figure 3.2 shows this iterative sampling process.

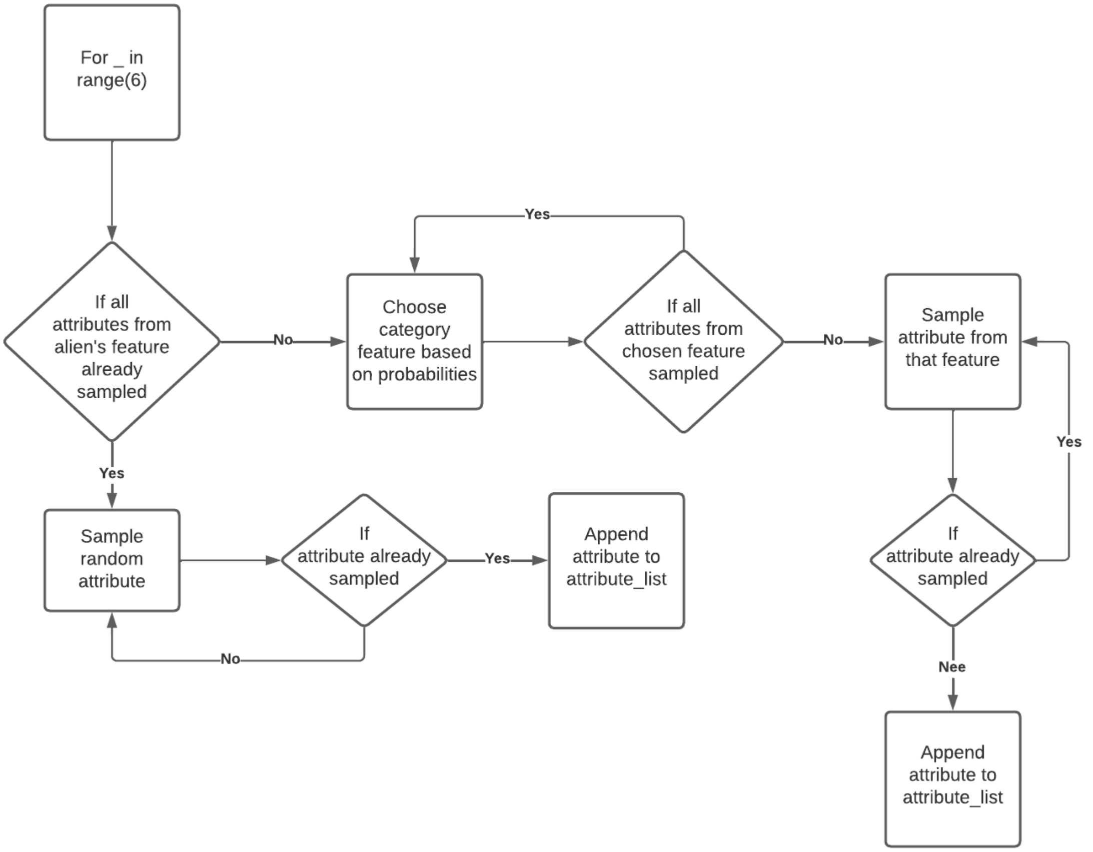

For _ in range(6)
If all attributes from alien's feature already sampled
No
Yes
Choose category feature based on probabilities
If all attributes from chosen feature sampled
Yes
No
Sample attribute from that feature
Yes
Sample random attribute
If attribute already sampled
Yes
No
Append attribute to attribute_list
If attribute already sampled
Nee
Append attribute to attribute_list

Figure 3.2: The reproduction phase of M2

Chapter 4

Results

The main aim of this research is to examine to which extent the proposed models give results similar to those of Martin et al. (2014) and Hutchison et al. (2018). To accurately compare results, the accuracy score and the mean number of attributes per generation are calculated for M1 and M2, since these results are also discussed in the original studies. Paired samples t-tests are then performed on both scores to see whether they have a significant difference at the end of the chain relative to the start. These tests will assess the internal validity of the models. Next, the standard error of estimates (SEOE) between the models' results and those of the original experiment are calculated to see which model performs best. The SEOE represents the difference between an estimated value (the models' results) and the true value (the original results). Finally, structure scores per number of shared features and structure scores by category feature are calculated and compared with the initial experiments through descriptive statistics. Because these results determine both the internal validity of the models and their similarity to the original experiment, the proposed hypotheses from chapter 3 can be accepted or rejected. As mentioned before, the first model has an implementation for both N = 1 (M1N1) and N = 2 (M1N2), with N standing for the number attributes that are remembered whenever an alien is presented.

4.1 Mean number of attributes

For M1N1, there is a significant reduction in the mean number of attributes ascribed to the aliens at Generation 7, compared to Generation 1. The mean number of attributes used in Generation 7 (M = 21) is lower than at Generation 1 (M =

28; $t(39) = 9.9104$, $p < 0.001$). When running M1N2, the model also shows a significant reduction in the mean number of attributes, as the number of attributes at Generation 7 ($M = 31$) is lower than at Generation 1 ($M = 18$; $t(39) = 24.743$, $p < 0.001$). Compared to the previous implementation, this version has an even more significant reduction. The second model (M2) shows a significant reduction as well, with the mean number of attributes used at Generation 7 ($M = 33$) lower than at Generation 1 ($M = 18$; $t(39) = 22.776$, $p < 0.001$). These findings are visualized in Figure 4.1 together with the results from the original experiment from Martin et al. (2014).

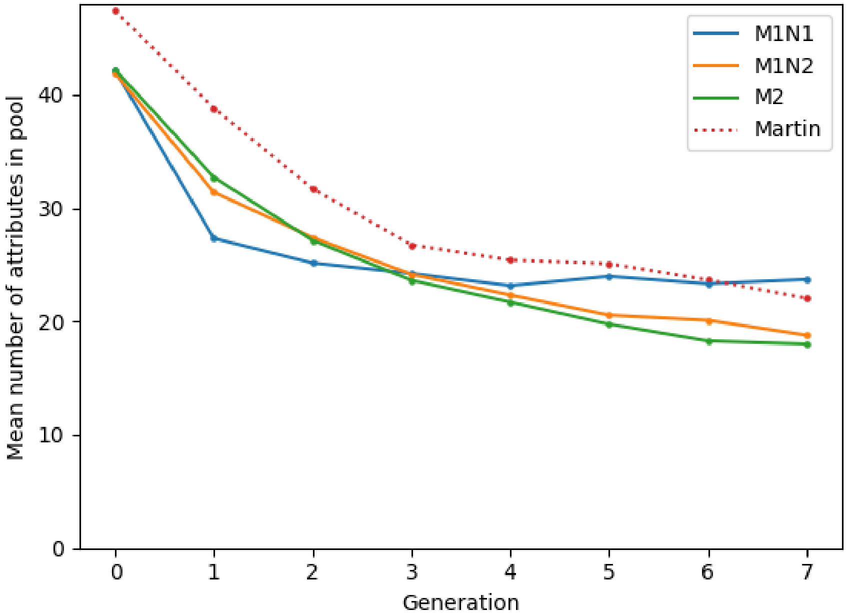


Figure 4.1: The mean number of attributes used per model compared to the results of the original experiment

To compare the performance of the models to the original experiment, the SEOE is calculated. This calculation means that for each generation, the difference between the original data and the model's data is taken, which is then squared. Next, all squared errors are added and then divided by the number of generations minus two.

Finally, to normalize the error, the square root is taken which results in the final error. The SEOE of all three models is displayed in table 4.1. The errors for all three models are relatively similar, but figure 4.1 shows that M1N2 and M2 follow the course of Martin's graph more closely than M1N1.

Model	**Error**
M1N1	5.81
M1N2	5.59
M2	5.88

Table 4.1: SEOE for mean number of attributes per model, compared to the results of Martin et al. (2014)

4.2 Accuracy

The data is split into seen and unseen aliens to analyze the mean accuracy, following Martin et al. (2014). For seen aliens, accuracy is defined as 'the percentage of attributes correctly assigned to aliens that appeared during the training phase' [19]. For unseen aliens, it is defined as 'the percentage of responses to aliens that did not appear during the training phase that matched those of the previous generation' [19]. Model M1N1 shows a significant increase in mean accuracy for seen aliens from Generation 1 (M = 46.8%) to Generation 7 (M = 62.7%; $t(39) = 17.968$, $p < 0.001$). More significant is the increase in mean accuracy for unseen aliens. A paired samples t-test show greater mean accuracy at Generation 7 (M = 45.1%) relative to Generation 1 (M = 16.1%; $t(39) = 25.413$, $p < 0.001$). However, these accuracy scores are remarkably lower than for seen aliens.

Model M1N2 also shows higher mean accuracy scores for seen aliens than for unseen aliens. Starting from Generation 1 (M = 70.6%), the score has increased significantly once Generation 7 is reached (M = 78.4%; $t(39) = 13.035$, $p < 0.001$). For unseen aliens, the mean accuracy score increases from Generation 1 (M = 14.4%) relative to Generation 7 (M = 78.4%; $t(39) = 36.095$, $p < 0.001$) as well, mirroring the effect seen in the previous implementation of the model.

Last, Model M2 also reveals greater mean accuracy for both seen aliens at Generation 7 (M = 54.2%) relative to Generation 1 (M = 32.9%; $t(39) = 17.968$, $p <$ .001) and unseen aliens at Generation 7 (M = 42.1%) relative to Generation 1 (M = 14.9%; t(11) = 25.413, p < .001). The results of the models compared to the original experiment are visualized in figure 4.2. In order to keep it comprehensible, the average of seen and unseen aliens is shown. As seen, all models give higher accuracy scores than the actual experiment. This is explained by a relatively high accuracy score for the seen aliens, as can be seen in figure 4.3. All models follow the same pattern for unseen aliens, with mean accuracy scores at Generation 1 around 15%, and at 40% at Generation 7, similar to the original experiment. However, the scores for seen aliens for M1N1 and M1N2 are deviating from the original, whereas M2 mostly captures the correct progress of increasing accuracy.

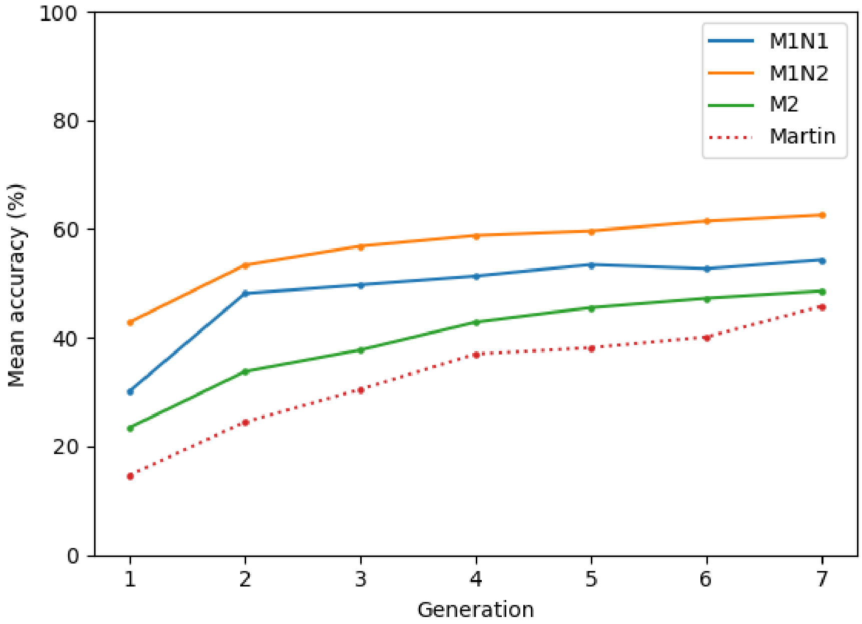
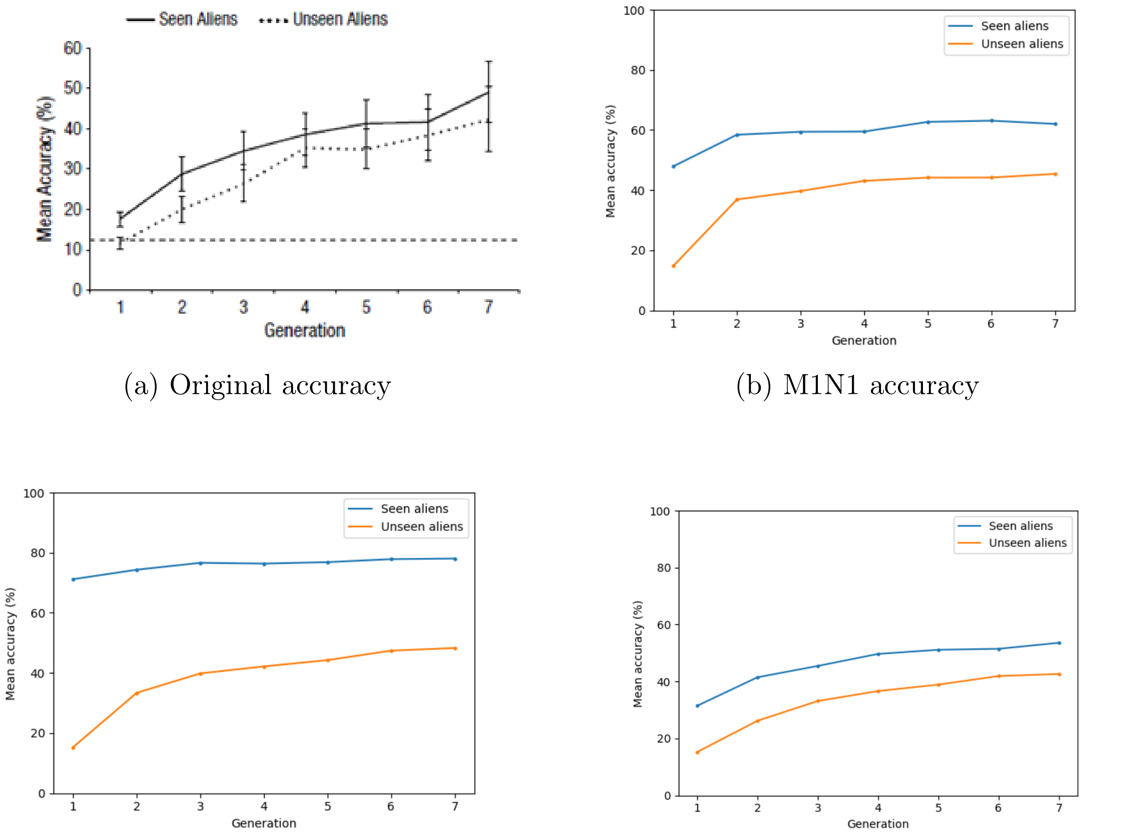


Figure 4.2: The mean accuracy per model compared to the results of the original experiment



(a) Original accuracy

(b) M1N1 accuracy

(c) M1N2 accuracy

(d) M2 accuracy

Figure 4.3: Accuracy scores per model and the original experiment, split into seen and unseen aliens

Once again, the SEOE is calculated. These results are displayed in table 4.2. The absolute values of the errors are much lower than in table 4.1. This is because the values of the mean accuracy are set between 0 and 1 for the calculation. Similar to the mean number of attributes SEOE, the SEOE of M2 is noticeably lower than that of M1N1 and M1N2.

Model	**Error**
M1N1	0.188
M1N2	0.288
M2	0.0864

Table 4.2: SEOE for mean accuracy (%) per model, compared to the results of Martin et al. (2014)

4.3 Structure

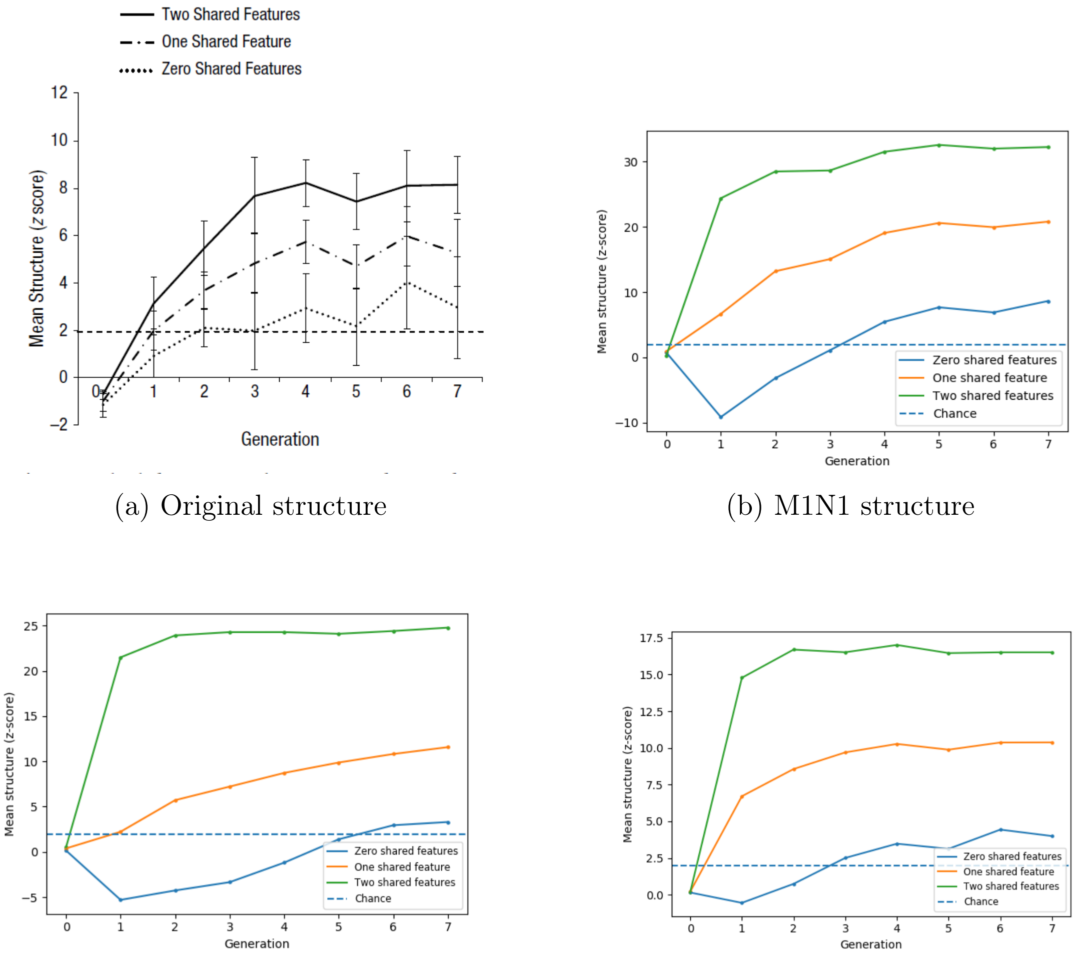
Besides accuracy and the mean number of attributes, Martin et al. (2014) and Hutchison et al. (2018) calculate structure scores to examine the overlap in attributes between aliens. In the case of Martin et al. (2014), structure scores are used for comparing aliens with zero, one or two shared category features. Meanwhile, Hutchison et al. (2018) use structure scores to investigate the amount of structure between aliens who share the same category feature (e.g., aliens who share color or shape). The structure scores per number of shared features are calculated for all three models. The structure score by category feature is only calculated for M2 since only this model implements the bias found by Hutchison et al. (2018). So, modeling this structure score for M1 will not give insightful results. For calculating structure scores, the raw structure scores are first measured. These are defined as the overlap in attributes between each individual alien and the other 26 aliens in the same generational pool [19]. If two aliens share three attributes, they are given a score of three, and so on. Next, the mean of these scores is taken for all aliens with zero, one or two shared features, resulting in the final raw structure scores per number of shared features. To avoid an increase or decrease in raw structure scores (if the initial number of attributes for Generation 0 increases, the structure decreases, and vice versa), 12.000 Monte Carlo simulations are run. These simulations randomly assigns the attributes to aliens to determine the mean structure between aliens, depending on the total number of used attributes. For example, it is now known what the average structure score is when 30 unique attributes are assigned to the aliens. These simulated results are then used as a comparison data set to calculate z-scores for the raw structure scores, forming the final structure scores.

4.3.1 Structure scores per number of shared features

Figure 4.4 shows the four different structure scores per number of shared features for the original experiment, M1N1, M1N2 and M2, respectively. All three models follow the same pattern of an increasing structure as the chain progresses, with the score being the highest for aliens with two shared features. Mean structure at Generation 1 was lower for M1N1 (mean z = 7.7), M1N2 (mean z = 5.9) and M2 (mean z = 6.4) than at Generation 7 (M1N1: mean z = 19.9, M1N2: mean z = 13.0, M2: mean z = 11.2). Moreover, M1N1, M1N2 and M2 all show lower mean structure scores for aliens with zero shared features (M1N1: mean z = 1.9, M1N2: mean z = -1.4, M2: mean z = 3.0) compared to aliens with one shared feature (M1N1: mean z = 14.2, M1N2: mean z = 6.7, M2: mean z = 8.9) and aliens with two shared features (M1N1: mean z = 26.1, M1N2: mean z = 21.1, M2: mean z = 15.0). Aliens with zero shared features all have a mean structure score close to what would be expected by chance, but this is no longer the case when the number of shared features increases. However, the scores of each model are noticeably higher than the original experiment, with structure scores of M1N1 three times as high at the end of the chain compared to 4.4a. Moreover, all three models show a big increase in structure for aliens with 2 shared features in Generation 1 compared to Generation 0, after which the score reaches a plateau. Another apparent observation is the decrease in structure for aliens with zero shared features in all models. These results are summarized in table 4.3 and figure 4.4.

Variable	Martin	M1N1	M1N2	M2
Gen 1	2.0	7.7	5.9	6.4
Gen 7	5.5	19.9	13.0	11.2
Zero SF	2.4	1.9	-1.4	3.0
One SF	4.6	14.2	6.7	8.9
Two SF	6.9	26.1	21.1	15.0

Table 4.3: Mean structure scores for shared feature measurement for the data of Martin et al. (2014) and all three models. SF stands for shared feature



(a) Original structure

(b) M1N1 structure

(c) M1N2 structure

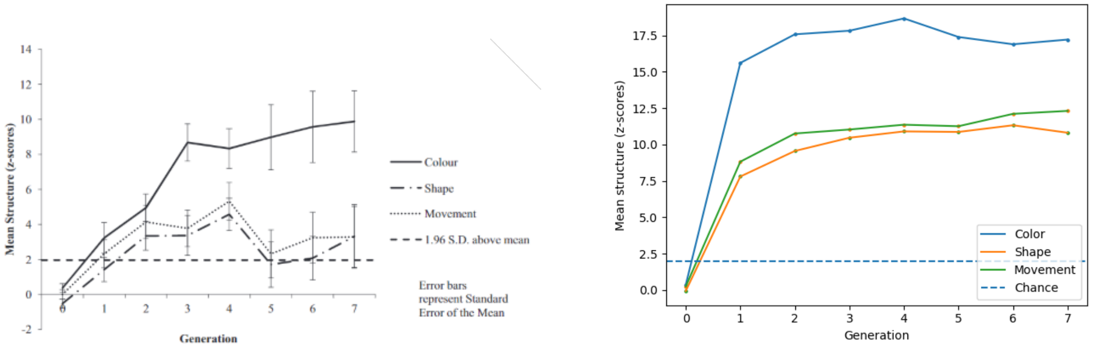
(d) M2 structure

Figure 4.4: Structure scores per number of shared features for the original experiment and the three models

From table 4.3 and figure 4.4, it follows that M2 has the overall best performance in terms of similarity to the original structure scores. Compared to the other two models, M2 overfits the least, especially for aliens with two shared features and the mean structure score at the end of the chain.

4.3.2 Structure scores per category feature

Structure scores per category feature are calculated in the same way as the previous structure scores. In figure 4.5, the results from Hutchison et al. (2018) are shown next to the results from M2.



(a) Original category structure

(b) M2 category structure

Figure 4.5: Structure scores per category feature for the original experiment and M2

Similar to the findings of Hutchison et al. (2018), structure for the color feature is higher than for the other two features. A noticeable difference however, is the faster increase in structure in the model than in the original experiment. The same effect was also seen in all models for the structure scores per number of shared features. Moreover, the values of the z-scores are approximately twice as high in M2 than in the original experiment.

4.4 Summary

As stated before, this research tries to answer the following question: *To what extent can the spontaneous formation of novel stereotypes be explained by human's limited memory capacity?* To do so, two hypothesis were proposed in Chapter 3:

> **Hypothesis 1:** *The spontaneous formation of novel cultural stereotypes can be explained solely by human's limited memory capacity.*
>
> **Hypothesis 2:** *M2 generates results that are more similar to results of the experiment by Martin et al. (2014) than M1 does.*

For mean accuracy and structure scores, M2 performs best out of all models. The mean number of attributes gave fitting results for both M1N2 and M2. Moreover, M2 gives significant increases in accuracy and structure, and has a significant decrease in the mean number of attributes as the chain progresses. These significant results give M2 a high internal validity. Since M2 gives results that are the most similar to the original results, hypothesis 2 is also accepted. However, while the mean number of attributes mirrors the original results, mean accuracy and structure scores of M2 both show some deviation from it. Thus, the current results don't confirm hypothesis 1. Nevertheless, the model captures the important components of memory loss. Its results indicate that by tweaking the model, the hypothesis could be accepted. In the following chapter, the limitations and possible improvements of the model are discussed.

Chapter 5

Discussion

This chapter will focus on the limitations of the models, and how further research could improve them. First, the lack of confirmation bias will be discussed. Next, possible explanations for the observed results that were unexpected are considered. Since M2 turns out to be the best fitting model, the following section will mostly focus on the limitations of and possible improvements for this model. However, M1N1 and M1N2 will be mentioned whenever relevant.

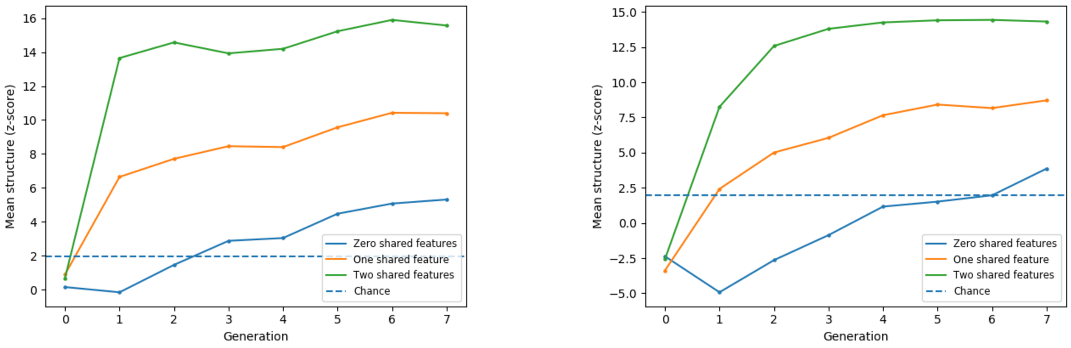
One of the main limitations of the proposed model is that it fails to simulate the updated beliefs of participants in the training phase. For example, if the attributes 'happy' and 'funny' had already been remembered at some point in the training phase, people are more likely to remember these attributes again if they are presented with a new alien. This is a form of confirmation bias; people tend to memorize things which are consistent with their beliefs and dismiss information which is not. This is relevant for the current model because it tries to simulate limited memory capacity, and remembering the same attribute more often requires less capacity. In the current implementation, each alien which is presented during learning is treated as if the participant has zero prior knowledge. Further research could look into the effect of the confirmation bias on the working of the proposed model. One way to do this is to assign a prior probability to each attribute at the beginning of the training phase, which is constantly updated based on the observation of new attributes. So, if an attribute is remembered, the prior probability of remembering this attribute again is higher than for attributes that are not remembered. A version of this method based on Bayesian inference is already used in M2 for determining which category feature (color, shape or movement) to use as a referent for the alien's attributes. Since this method provides a more complete overview of memorization, its implementation could give a more conclusive answer to the current research questions. The models' limitations that may be prevented by using a confirmation bias will be mentioned in the following part of the discussion.

In Chapter 4, some noticeable deviations from the models compared to the actual experiment were already noted. First of all, the models show a higher mean accuracy score for seen aliens than would be expected. This indicates that the models' implementation of memory during the training phase is too ideal. The difference in mean accuracy between M1N1 and M1N2, where the latter has the higher score, can be explained by the fact that more attributes had been remembered in the training phase. For instance, the implementation of M1N2 led to a total of six remembered attributes per alien. Because of this, it is very likely that four or more of the previous attributes are recalled in the reproduction phase, almost recreating the associations from the previous generation. Even when only a total of three attributes is remembered (which is the case in M1N1), chances are that 50% of the previous attributes are recalled during reproduction. In M2 this effect is less present, since its method for remembering attributes per category feature, instead of per alien, decreases the chance of recreating the alien from the previous generation. While M2 has the best fit, its mean accuracy for seen aliens at the start of the chain is still around 15% higher than the original mean accuracy. Changing the value of N, which stands for the number of attributes that are remembered per alien, does not influence the accuracy score in a significant way. Nonetheless, the high accuracy score could be attributed to the fact that M2 does not capture the confirmation bias described earlier. Mistakes in remembering the attributes are never made by the model, which could be happening in real life because of the confirmation bias. So, the models could be tweaked by including a function in the training phase that stores random attributes that are not originally assigned to the alien. Incorrect attributes may now be assigned to seen aliens in the reproduction phase, possibly lowering the accuracy score.

None of the models are able to accurately capture the effect on structure found by Martin et al. (2014) and Hutchison et al. (2018). Even though all of them display the difference in structure scores depending on the number of shared features or the type of shared feature correctly, the scores are higher than expected. For aliens sharing two features, these higher scores are already achieved at Generation 1, instead of the slower increase seen in the original results. Moreover, structure scores for these aliens stop increasing significantly after Generation 1. The high scores for M2 may be caused by the low number of stored attributes, since different N values cause different structure scores. Remarkably, a high N value produces relatively low structure scores, whereas a low N value leads to higher structure scores. A possible explanation is that whenever a small amount of attributes is stored during the training phase, similar aliens are more likely to end up with the same attributes. On the other hand, if there are many attributes to pick from, the chance of overlap between aliens decreases. So, it appears that modeling larger memory capacity leads to lower structure scores. However, the aim of this model is to simulate how limited memory capacity leads to more structure. Further research could investigate how to lower the absolute values of the structure scores. Moreover, the implementation of the confirmation bias might prevent the effect of low structure scores for large memory capacity.

Currently, the models make a clear distinction between recalling attributes (from aliens that were seen during the training phase) and predicting attributes (for aliens that were not seen). To see whether the element of prediction has a significant effect on the results of M2, an extra model is build which reduces the set of aliens from 27 to the 13 aliens from the training set. Now, unseen aliens are no longer observed during the reproduction phase. Since participants did not notice that they also assigned attributes to unseen aliens in the original experiment, this new implementation will reveal whether they truly treat seen and unseen aliens the same.

In this updated model, only aliens that were seen during the training phase are presented during the reproduction phase, thereby removing the element of prediction. Since this method may cause certain category features to disappear (for instance, none of the aliens in the training set may have the color blue), the average of 80 chains instead of 40 is taken to analyze its performance. As expected from the results of M2, the mean accuracy at the start is higher without the prediction element. Most interesting is the better performance in structure for the adapted model. Figure 5.1 shows the mean structure scores for M2 and for the adapted model featuring only seen aliens side by side.



(a) M2 structure

(b) Adapted model structure

Figure 5.1: Structure scores for M2 and for the adapted model only featuring seen aliens

This new model shows the slower increase as seen in the original research, instead of the steep incline in M3. This indicates that the prediction element might partly cause the deviations from the original results as found in M2. Further research could investigate this effect more extensively.

To summarize, further research can explore the role of the confirmation bias in memorizing, specifically for this experiment. This would provide a more complete implementation of human memory in the model, and thus a more inclusive answer to the research question. And secondly, the difference between predicting and recalling aliens and their attributes can be further investigated.

Chapter 6

Conclusion

The goal of this thesis was to offer a better understanding to how novel cultural stereotypes form. More specifically, the role of memory capacity in regards to this process is studied. How are these stereotypes formed without previous beliefs or biases regarding the subjects? By using an iterative learning model to simulate a psychological experiment focusing on stereotype formation from Martin et al. (2014), this thesis tries to recreate the original results by assuming that human's limited memory capacity influences the formation of novel stereotypes. To do so, three different models were proposed, each using a unique learning strategy. Other cognitive biases were not implemented in order to isolate the effect of memory, and ignore any (cultural) assumptions people might have. In the end, the learning strategy most efficient for one's memory turned out to provide the best results. This already indicates the importance of memory on stereotype formation. This strategy involves dissecting an alien into its three category features, and using these features as a referent for memorizing the assigned attributes. During the reproduction phase, attributes are assigned based on an alien's category feature and the included attributes from the learning phase. While the model provides some accurate results and suggests that categorization takes place even without other cognitive biases, its limitations cannot be ignored. Some fundamental aspects of the memorization are left out, such as a confirmation bias. Moreover, the results show accuracy and structure scores way higher than expected. To form a better understanding of the role of limited memory capacity, further research should look into how a confirmation bias might affect the models.

Bibliography

[1] Gordon W Allport and Leo Postman. The psychology of rumor. 1947.

[2] Frederic Charles Bartlett and Frederic C Bartlett. *Remembering: A study in experimental and social psychology*. Cambridge university press, 1995.

[3] Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794, 2016.

[4] Robert Boyd, Peter J Richerson, et al. Why culture is common, but cultural evolution is rare. In *Proceedings-British Academy*, volume 88, pages 77–94. Oxford University Press Inc., 1996.

[5] Bruno S Frey. "just forget it."memory distortions as bounded rationality. *Mind & Society*, 4(1):13–25, 2005.

[6] Anthony G Greenwald and Mahzarin R Banaji. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27, 1995.

[7] Thomas L Griffiths and Michael L Kalish. Language Evolution by Iterated Learning With Bayesian Agents. Technical report, 2007.

[8] Abdul Haque and Mohammed Sabir. The image of the indian army and its effects on social remembering. *Pakistan Journal of Psychology*, 8(1):55, 1975.

[9] Amy S Harasty. The interpersonal nature of social stereotypes: Differential discussion patterns about in-groups and out-groups. *Personality and Social Psychology Bulletin*, 23(3):270–284, 1997.

[10] James L Hilton and William Von Hippel. Stereotypes. *Annual review of psychology*, 47(1):237–271, 1996.

[11] Jacqui Hutchison, Sheila J. Cunningham, Gillian Slessor, James Urquhart, Kenny Smith, and Douglas Martin. Context and Perceptual Salience Influence the Formation of Novel Stereotypes via Cumulative Cultural Evolution. *Cognitive Science*, 42:186–212, 5 2018.

[12] Yoshihisa Kashima and Victoria Wai-Lan Yeung. Serial Reproduction: An Experimental Simulation of Cultural Dynamics. *Acta Psychologica Sinica*, 42(1):56–71, 2 2010.

[13] Simon Kirby, Mike Dowman, and Thomas L. Griffiths. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 104(12):5241–5245, 3 2007.

[14] Simon Kirby and James R Hurford. The emergence of linguistic structure: An overview of the iterated learning model. *Simulating the evolution of language*, pages 121–147, 2002.

[15] Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 8 2015.

[16] Anthony Lyons and Yoshihisa Kashima. How are stereotypes maintained through communication? the influence of stereotype sharedness. *Journal of personality and social psychology*, 85(6):989, 2003.

[17] Anthony Lyons and Yoshihisa Kashima. Maintaining stereotypes in communication: Investigating memory biases and coherence-seeking in storytelling. *Asian Journal of Social Psychology*, 9(1):59–71, 2006.

[18] Douglas Martin, Sheila J. Cunningham, Jacqui Hutchison, Gillian Slessor, and Kenny Smith. How societal stereotypes might form and evolve via cumulative cultural evolution. *Social and Personality Psychology Compass*, 11(9), 9 2017.

[19] Douglas Martin, Jacqui Hutchison, Gillian Slessor, James Urquhart, Sheila J. Cunningham, and Kenny Smith. The Spontaneous Formation of Stereotypes via Cumulative Cultural Evolution. *Psychological Science*, 25(9):1777–1786, 9 2014.

[20] Douglas L Medin and Marguerite M Schaffer. Context Theory of Classification Learning. Technical Report 3, 1978.

[21] Douglas L Medin and Edward E Smith. Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, 7(4):241, 1981.

[22] Alex Mesoudi. How Cultural Evolutionary Theory Can Inform Social Psychology and Vice Versa. *Psychological Review*, 116(4):929–952, 10 2009.

[23] Sendhil Mullainathan. A memory-based model of bounded rationality. *The Quarterly Journal of Economics*, 117(3):735–774, 2002.

[24] Roger N Shepard, Carl I Hovland, and Herbert M Jenkins. Learning and memorization of classifications. *Psychological monographs: General and applied*, 75(13):1, 1961.

[25] Catriona Silvey, Simon Kirby, and Kenny Smith. Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39(1):212–226, 1 2015.