**Doctoral thesis submitted to**
**the Faculty of Behavioural and Cultural Studies**
**Heidelberg University**
**in partial fulfillment of the requirements of the degree of**
**Doctor of Philosophy (Dr. phil.)**
**in Psychology**

Title of the publication-based thesis

*Understanding Cognitive Processes Underlying Belief Polarization and Function-Learning: Experimental and Modeling Approaches*

presented by
Nadia Said

year of submission
2019

Dean: Prof. Dr. Dirk Hagemann
Advisor: Prof. Dr. Dr. h.c. Joachim Funke

بِسمِ اللّٰهِ الرَّحمٰنِ الرَّحِمِ

## Acknowledgements

I would like to thank my supervisor Prof. Dr. Dr. h. c. Joachim Funke for his support throughout the years and especially for giving me the opportunity and freedom to follow my ideas.

My heartfelt thanks go to Prof. Dr. Dr. h. c. mult. Hans Georg Bock for his constant support and advice throughout the years. It was a great pleasure to be part of your research group.

I would like to express gratitude to my mentors Dr. Helen Fischer, Prof. Dr. Christian Kirches, Prof. Dr. Stefan Körkel, and PD Dr. Andreas Potschka. Their constant support and trust in my capabilities helped me to overcome any obstacles and inspired me to give my best. Furthermore, I would like to thank Dr. Johannes P. Schlöder for his support, advice, and the many interesting discussions.

My sincere thanks also go to my student assistants Debora Fieberg and Gerrit Anders who did an excellent job during our time working together.

It was a pleasure working at the Interdisciplinary Center for Scientific Computing and the Institute of Psychology in the past years. In particular, I would like to thank the members of Experimental and Theoretical Psychology, Simulation and Optimization, and Optimum Experimental Design Group, Dr. Dorothee Amelung, Dr. Carola Barth, Dr. Anja Bettendorf, Dr. Lilli Bergner, Dominik Cebulla, Dr. Michael Engelhart, Dr. Andreas Fischer, Dr. Jürgen Gutekunst, Dr. Christian Hoffmann, Dr. Dennis Janka, Johannes Herold, Dr. Florian Kehrle, Dr. Robert Kircheis, Daphne Pediaditakis-Kresse, Manuel Kudruss, Dr. Tom Kraus, Dr. Huu Chuong La, Dr. Conrad Leidereiter, Dr. Enrique Guerrero Merino, Christian Mittelstaedt, Andreas Meyer, Marta Sauter, Matthias Schlöder, Dr. Andreas Schmidt, Dr. Andreas Sommer, Robert Scholz, Alexander Wendt, and Dr. Leonard Wirsching. Thank you for the fun time, inspiring discussions, many cups of coffee, and words of encouragement.

For proofreading this manuscript I would like to thank Dr. Anja Bettendorf, Dr. Stephan Feder, Johannes Herold, Katharina Nief-Said, and Dr. Andreas Sommer.

For helping me through the jungle of administration, I would like to thank Sabine Falke, Annika Flämig, Silke Thiel, Jeannette Walsch, and Anastasia Walter. For maintaining the computing facilities I would like to thank Thomas Kloepfer, Marion Lammarsch, Dr. Hermann Lauer, and Martin Neisen.

Furthermore, I would like to thank the whole team of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences: Dr. Michael J. Winckler, Ria Hillenbrand-Lynott, Jan Keese, Oktavia Klassen, Anne Paulski, Christine Tomanek, and Sarah Steinbach.

I would like to express my gratitude to the DFG Graduate School 220 (HGS Math Comp), the Excellence Initiative Institutional Strategy ZUK 5.4., the Excellence Initiative FRONTIER (2017), and the Heidelberg Center for the Environment for providing funding for this work.

Finally, I would like to thank my parents Katharina and Ibrahim, my brothers Jonas and Osama, and my friends Anja, Anna, Fatma, Fattma, Nohma, Ruaa, and Sara. Without your constant love, support, encouragement and many many hugs, nothing of this would have been possible.

# Contents

# 1 List of Publications

- **Manuscript 1: Said N.** , Engelhart M., Kirches C., Körkel S., Holt D.V. (2016). Applying Mathematical Optimization Methods to an ACT-R Instance-Based Learning Model. *PLoS ONE* 11(7): e0158832. doi:10.1371/journal.pone.0158832

  **Author Contributions:**

  - Conceived & designed the experiments: **Said N.**, Engelhart M., Kirches C., Körkel S., Holt D.V.
  - Performed the experiments: **Said N.**, Engelhart M.
  - Analyzed the data: **Said N.**, Engelhart M.
  - Wrote the paper: **Said N.**, Engelhart M., Kirches C., Körkel S., Holt DV.

- **Manuscript 2:** Fischer H., Amelung D., **Said N.** (2019).The accuracy of German citizens' confidence in their climate change knowledge. *Nature Climate Change*. doi: 10.1038/s41558-019-0563-0

  **Author Contributions:**

  - Conceived & designed the experiments: Fischer H., **Said N.**, Amelung D.
  - Analyzed the data: Fischer H., **Said N.**
  - Wrote the paper: Fischer H., **Said N.**, Amelung D.

- **Manuscript 3: Said N.** , Fischer H. *Memory & Cognition (under review)*. Extrapolation performance underestimates rule learning: Evidence from the function-learning paradigm.

  **Author Contributions:**

  - Conceived & designed the experiments: **Said N.**, Fischer H.
  - Performed the experiments: **Said N.**
  - Analyzed the data: **Said N.**
  - Wrote the paper: **Said N.**, Fischer H.

# 2 | Introduction

The beliefs we hold not only influence how we seek out (Mynatt, Doherty, & Tweney, 1977; Nickerson, 1998) and perceive new information (Alloy & Tabachnik, 1984; Crocker, 1981) but also influence whether we take action (Tobler, Visschers, & Siegrist, 2012) and thus have far reaching impact on decision-making (Russo, Schoemaker, & Russo, 1989).

In recent years there has been a rise of *belief polarization* (Iyengar, Sood, & Lelkes, 2012; Pew Research Center, 2014; Webster, 2005; Zarkov, 2017). The term "belief polarization" refers to the intensification of disagreement over a topic due to the way new information is selected and processed. Some are arguing that this phenomenon is caused or at least accelerated by the Internet and social media (Gabler, 2016; Sunstein, 2018). Terms like "echo chambers" and "filter bubbles", the first referring to reinforcement of beliefs due to repetition inside a closed communication space and the latter to reinforcement of beliefs due to content tailored search algorithms, become increasingly popular in discussions about belief polarization.

Even though there is still an open debate about the contribution of social media to belief polarization, with some claiming it is overestimated (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015; Boxell, Gentzkow, & Shapiro, 2017), consequences of false beliefs can be severe, with for example near-record measles outbreaks in the U.S. during 2018 (CBSNews, 2019; Fox, 2019) due to the growing anti-vaccination movement (Dubé, Vivion, & MacDonald, 2015; Kata, 2012). Similarly, the causes of *climate change* have been subject of ongoing discussions. Even though the vast majority of scientific evidence clearly shows that climate change is a man-made issue (Cook et al., 2013), it still divides the population with, e.g., 32% of US citizens believing that the causes are natural in 2016 (Marlon, Howe, Mildenberger, & Leiserowitz, 2016). This is concerning, as fast implementations of counter-measures are becoming a more and more urgent matter (IPCC, 2018).

The main goal of the present research is twofold: First, to contribute to the understanding of propagation and polarization of beliefs from a *cognitive perspective* by integrating experimental findings regarding confidence in climate change knowledge (Manuscript 2) and cognitive modeling approaches (Manuscript 1) into an agent-based belief model. Second, to outline how an implementation of a function-learning model in a cognitive architecture, based on experiments conducted in Manuscript 3, can contribute to a better understanding of cognitive processes underlying the understanding of non-linear functions.

**Modeling belief polarization.** To model belief polarization we use an *agent-based modeling* (ABM) approach. Agent-based modeling allows for simulating individual behavior and study how it influences others as well as the environment. We implemented an agent-based belief model to explore the influence of *psychological parameters*, particularly the influence of *cognitive parameters*, on polarization of beliefs in an initially heterogeneous belief environment. The present model allows to investigate how (a)

differences in openness contribute to the occurrence of echo chambers, (b) belief polarization is influenced by cognitive parameters like working memory capacity and confirmation bias, and (c) individual differences in, for example, openness, working memory, or confidence influence belief polarization.

In sociology and social psychology, several agent-based models that examine the development of beliefs held by agents have been developed (e.g., Baumgaertner, Tyson, & Krone, 2016; Duggins, 2014; Mäs & Flache, 2013). One example are psychologically motivated implementations of Marsella, Pynadath, and Read (2004), who incorporated parameters like self-interest and consistency. Duggins (2014) investigated the influence of tolerance, conformity, and commitment to extreme beliefs on belief polarization. In contrast, the aim of the present research is to take a cognitive perspective by representing the memory capacity and temporal dynamics of agents' learning behavior as well as including parameters representing cognitive biases.

There are several models that focus on the role of confirmation bias on opinion formation (Fryer, Harms, & Jackson, 2018; Ngampruetikorn & Stephens, 2016; Sobkowicz, 2017). Fryer et al. (2018), for example, showed that ambiguous information can lead to belief polarization. As confirmation bias is an important driver regarding belief polarization, we included this parameter into our model. However, it is not the main focus of the presented work. Typically, agents' information processing is modeled using *Bayesian updating*, that is, agents update their beliefs based on Bayes' theorem (Bayes, 1763). Even though this is a widely used approach (e.g. Dixit & Weibull, 2007; Sobkowicz, 2017), one could argue that assuming fully rational agents does not reflect people's naturally flawed reasoning processes (McKelvey & Page, 1990). Pilditch (2017), for example, equipped agents with a different learning mechanism using a reinforcement learning model. However, as Bayesian belief updating provides a feasible starting point (Acemoglu & Ozdaglar, 2011; Moore & Healy, 2008), it is used in our first model set up. To account for humans' limited memory, we introduce a working memory capacity parameter. As will be outlined in Section 3.4, we will then even go one step further, discussing how agents' memory can be modeled in a more complex way by equipping agents with our full parameterized declarative memory module developed in Manuscript 1.

**Understanding non-linear processes.** We do not only encounter non-linear behavior of processes on a small scale in daily life (for example fuel consumption) but also on a larger scale with impact on society as a whole (economical and population growth Hajamini, 2015, climate change Schneider, 2004). Thus, a correct understanding of this type of behavior is highly relevant in terms of decision-making (Newell, McDonald, Brewer, & Hayes, 2014). In Manuscript 3, we conducted two experiments and showed that while participants demonstrated accurate understanding of the function-rule, they were not able to apply this understanding in the standard function-learning paradigm. In Section 4.1, I will briefly present the results of our working memory assessment that we conducted alongside the two function-learning experiments. In Section 4.2, I will outline a first approach to model cognitive processes underlying function-learning based on our experimental results.

This thesis is structured as follows: First, I will describe the setup of the agent-based belief model and present selected simulation results on how (a) differences in *openness* contribute to the occurrence of echo chambers, (b) belief polarization is influenced by *cognitive parameters* like working memory capacity and confirmation bias, (c) *individual differences* in openness and working memory influence belief polarization. Second, I will discuss how the model can be extended by incorporating the mathematical

formalization of the declarative memory module of the *Adaptive Control of Thought-Rational* (ACT-R) architecture developed in Manuscript 1. This will allow to simulate agents' cognitive processes in a more complex way as well as to introduce other cognitive parameters like memory decay.

Third, I will outline how to include another cognitive parameter (*confidence*) into the model based on the experimental results of Manuscript 2 and show simulation results regarding the influence of different values of (over-)confidence on belief polarization. Fourth, I will discuss how the results of the function-learning experiment conducted in Manuscript 3 can be integrated into a cognitive model that allows for (a) modeling cognitive processes underlying rule-based prediction and application failure and (b) investigating the influence of individual differences in working memory capacity on prediction performance.

# 3 An Agent-Based Model of Belief Polarization

## 3.1 Introduction to Agent-Based Modeling

*Agent-based modeling* is a method to computationally study individual behavior and how it affects others as well as the environment. Smith and Conrey (2007) define ABM as "a tool to conceptually bridge between the micro level of assumptions regarding individual agent behaviors, interagent interactions, and so forth and the macro level of the overall patterns that result in the agent population." An agent-based model typically consists of three elements: 1. agents to which specific attributes and behaviors are assigned, 2. agents' relationships and methods of interaction, and 3. an environment in which agents live in and interact with each other. An agent is defined as an autonomous, discrete entity equipped with specific attributes and a set of dynamic states (e.g., position, behaviors) that vary over time (Macal & North, 2011).

## 3.2 Modeling Belief Polarization

To explore the influence of psychological parameters on polarization of beliefs, we implemented an agent-based model modifying and extending Schelling's Segregation Model (Schelling, 1971), i.a., by equipping agents with a belief. This allows us to analyze how (a) differences in *openness* contribute to the occurrence of echo chambers, (b) belief polarization is in general influenced by *cognitive parameters* like confirmation bias and working memory capacity, and (c) *individual differences* in, e.g., openness, working memory capacity, and *confidence* influence belief polarization.

### 3.2.1 Attributes of Agents

The model consists of multiple agents $a_i$ with $i \in \{1, ..., N\}$, each with dynamic states varying over time $t$ (e.g. round $R$). These states are an agent's position $(x_i(t), y_i(t)) \in [0,1] \times [0,1]$ and belief $b_i(t)$, see Figure 3.1. An agent's *belief* can take values between 0 and 1, $b_i(t) \in [0,1]$.

$$a_i : T \rightarrow [0,1]^3, T = \{1, ..., n_T\} \tag{3.1}$$

$$t \mapsto (x_i(t), y_i(t), b_i(t)). \tag{3.2}$$

Furthermore, the agents have explicit goals (finding an x-y-position where they are "happy" ($h(a_i) = 1$, with $h(a_i) \in \{0,1\}$) and the ability to learn and adapt their behaviors based on experience (which requires a memory $\gamma_{a_i}$) (see Tab. 3.1). Interaction between

agents is restricted to a limited number of agents at any given time. This is achieved by defining a local neighborhood. In our case the maximum number each agent can interact with is set to $n_{NN} = 15$[1]. The nearest agents $a_{l(j)}, ..., a_{l(n_{NN})}$, with $l(j) = j$-next neighbor with which an agent $a_i$ can interact with, are calculated as follows:

$$\left\| a_i - a_{l(j)} \right\|_{\omega_i} := \sqrt{(x_i(t) - x_{l(j)}(t))^2 + (y_i(t) - y_{l(j)}(t))^2 + \omega_i \cdot (b_i(t) - b_{l(j)}(t))^2}, \quad (3.3)$$

with $a_i, a_{l(j)} \in \mathscr{A}$, $\mathscr{A}$ the set of all agent points, and $\omega_i \in \mathbb{R}_{\geq 0}$ the *confirmation bias*. The belief dimension is included into the nearest neighbors calculation to introduce the psychological construct *confirmation bias* into our model. The weighting parameter $\omega_i$ reflects agents' tendency of seeking out confirming information of their belief. If $\omega_i = 0.0$, the calculation of the nearest neighbors is solely based on their position on the x-y-plane[2].

(a) (b)

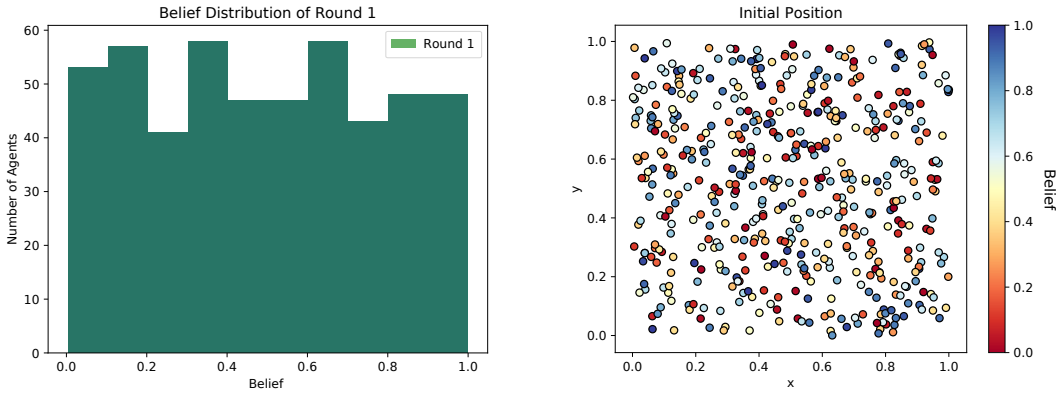



**Figure 3.1.** Initial beliefs and positions of agents. Figure (a) shows the initial belief distribution and figure (b) the initial position of $n = 500$ agents.

**Table 3.1.** Model properties of agent-based belief model.

| | | |
|---|---|---|
| $R$ | $\in \mathbb{N}$ | number of rounds |
| $\delta_{\max}$ | $\in \mathbb{N}$ | number of loops |
| $n_a$ | $\in \mathbb{N}$ | number of agents |
| $n_{\text{NN}}$ | $\in \mathbb{N}$ | number of neighbors with whom information is shared |
| $h$ | $\in \{0,1\}$ | happiness (sufficient number of nearest neighbors hold same belief) |
| $\omega$ | $\in [0,1]$ | weight for belief dimension of distance (confirmation bias) |
| $\theta$ | $\in [0,1]$ | belief deviation threshold (openness) |
| $b$ | $\in [0,1]$ | belief |
| $\gamma$ | $\in \mathbb{N}$ | working memory capacity |

Algorithm 1 displays how the simulations are generated. Each agent starts out with a position and belief, both randomly drawn from a uniform distribution $(x, y, b \sim U([0,1]))$.

[1]Note that even though setting $n_{NN}$ to 15 interaction neighbors is a reasonable choice, further simulation runs varying the values for $n_{\text{NN}}$ are required.

[2]Note that for $\omega \in (0,1]$ a norm is defined, for $\omega_i = 0.0$ a norm is not defined.

We chose a uniform distribution as this allows us to assess the impact of the different cognitive parameters on belief polarization. Goal of each agent is to find a position where it is "happy" ($h(a_i) = 1$). The happiness $h$ of an agent $a_i$ is defined as being close to agents who hold similar beliefs,

$$h : \mathscr{A} \to \{0,1\}, \tag{3.4}$$

$$a_i \mapsto h(a_i),\ h(a_i) := \begin{cases} 1 & \text{if } \text{mean}_{j=1,\dots,n_{NN}} \left\| b_{l(j)}(t) - b_i(t) \right\|_{\omega_i} \leq \theta_i \\ 0 & \text{else} \end{cases} \tag{3.5}$$

with $j = 1,\dots,n_{NN}$ the nearest neighbors, and $\theta_i \in \mathbb{R}^+$ the *belief deviation threshold*. How much deviation from their own belief is tolerated by an agent is set by $\theta_i$, the belief deviation threshold. This can be interpreted as *openness* in that openness is associated with curiosity as well as a tendency to be liberal and having a higher tolerance of diversity (Butrus & Witenberg, 2013; Jost, 2006; McCrae, 1996; Peterson, Seligman, et al., 2004). Put differently, agents with high openness values are more likely to be close to and interact with agents who hold different beliefs.

**Algorithm 1:** Model simulation

```
Input: x_i, y_i, b_i, δ_max;
Output: h(a_i);
for each round R = 1,...,n_T do
    for each agent i = 1,...,n_a do
        for j = 1,..., δ_max do
            if h(a_i) = 1 then
                Bayesian belief update;
                proceed to next agent;
            end
            select (x_i, y_i) randomly, x, y ~ U([0,1]);
        end
    end
end
```

During each round, agents randomly switch their position until they find a place where the beliefs of their nearest neighbors are equal or below an agent's belief deviation threshold. The parameter $\delta_{\max} \in \mathbb{N}$ defines how often an agent can switch its position. If $h(a_i) = 1$, the agent will update its belief based on the mean value of the beliefs of its neighbors,

$$b_i(t+1) = \frac{\text{mean}_{j=1,\dots,n_{NN}} b_{l(j)}(t) \cdot b_i(t)}{\text{mean}_{j=1,\dots,n_{NN}} b_{l(j)}(t) \cdot b_i(t) + (1 - \text{mean}_{j=1,\dots,n_{NN}} b_{l(j)}(t)) \cdot (1 - b_i(t))}. \tag{3.6}$$

If the agent cannot find a "happy" position ($h(a_i) = 0$), it will not update its belief. In this first set up, our model agents' beliefs are updated by *Bayesian belief updating*. In the following sections however, we will successively extend the model by equipping agents with a (limited) memory. In Section 3.3.3, we will introduce a memory parameter that allows to model individual differences in working memory capacity and show how those influence belief updating. In Section 3.4, we will outline how agents beliefs can be updated by integrating our mathematical formalization of the declarative memory module of the cognitive architecture ACT-R into the agent-based belief model (Manuscript 1).

## 3.3 Simulation Results

In this section, we will present exemplary simulation results for different parameter settings. As the main goal is to outline the setup of the model as well as to show how the findings from Manuscript 1 and 2 can be integrated into the model, we restrict ourselves to the presentation of some selected simulation results. In all the simulation results presented here agents start with a uniform belief distribution.

### 3.3.1 Emergence of Echo Chambers

Figure 3.2 displays the clustering of agents over time (after $R = 10$ rounds) depending on their openness $\theta$.

Simulation results show that higher values of openness lead to less clustering than lower values. However, as Figure 3.3 shows, agents did not only display clustering but also belief polarization. Thus, in the next section we will have a closer look at how psychological parameters, particularly cognitive parameters, influence belief polarization.

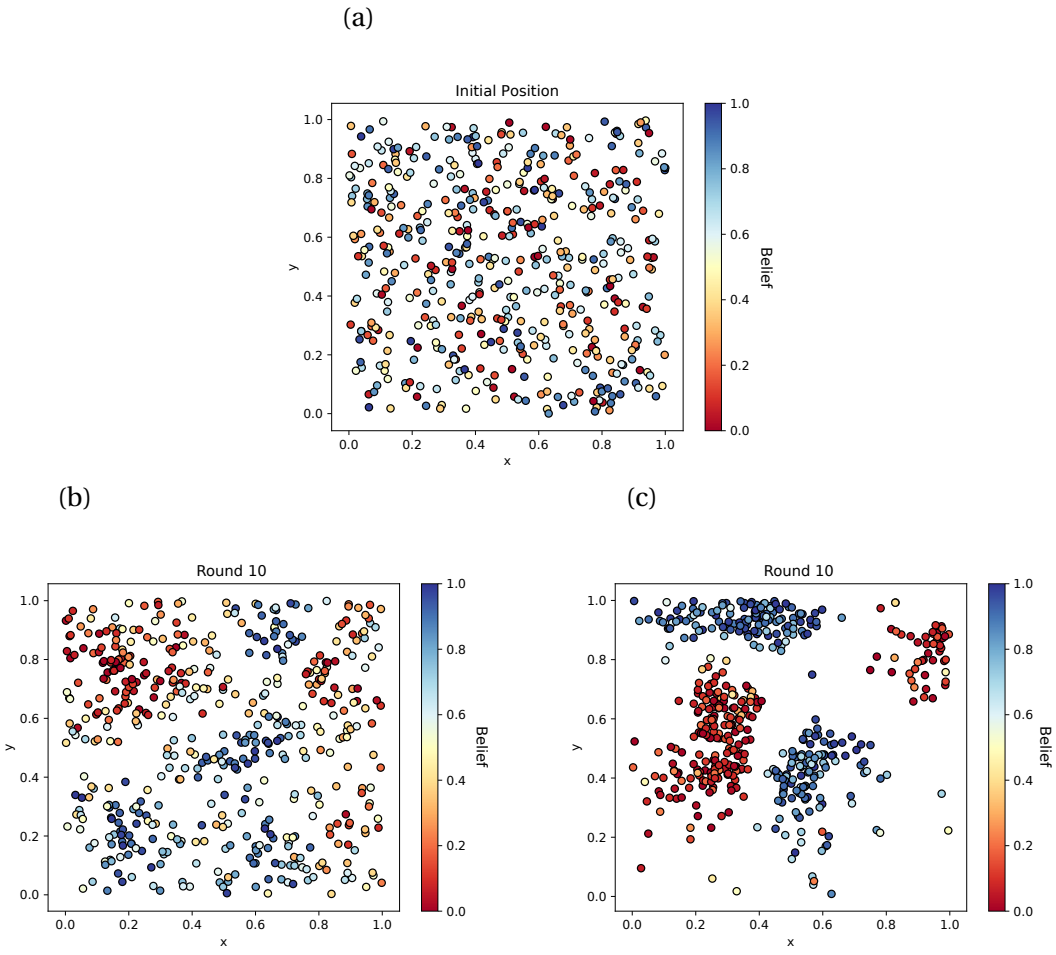



**Figure 3.2.** Influence of *openness* $\theta$ on clustering. The figure displays the position of $n = 500$ agents, with $n_{\mathrm{NN}} = 15$ interaction neighbors and confirmation bias $\omega = 0.0$. (a) shows the initial position of agents. (b) shows the position of agents after $R = 10$ rounds, for openness $\theta = 0.3$. (c) shows the position of agents after $R = 10$ rounds, for openness $\theta = 0.1$.

(a) (b)

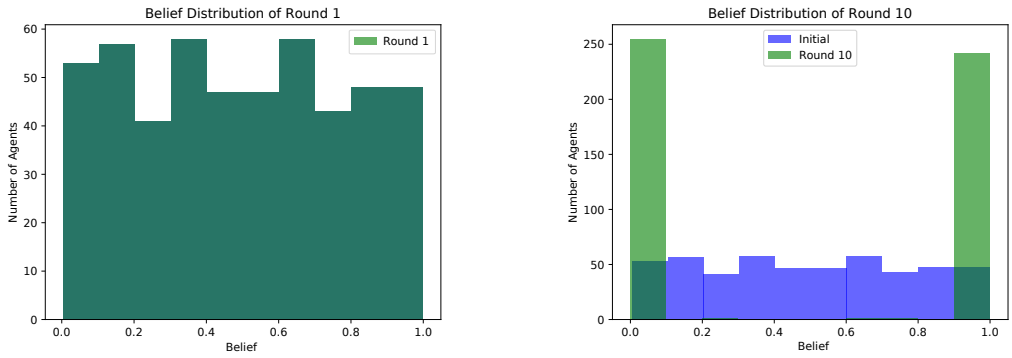



**Figure 3.3.** Belief distribution of $n = 500$ agents, with $n_{NN} = 15$ interaction neighbors and confirmation bias $\omega = 0.0$. (a) shows the initial belief distribution and (b) the belief distribution after $R = 10$ rounds, for openness $\theta = 0.1$.

### 3.3.2 Belief Polarization

In the following, *belief polarization* is defined as the percentage of agents who hold either 0 or 1 as belief after $R$ rounds. For the parameter values investigated in this thesis we performed simulation runs in which $R$ was successively increased from 5 to 20. Results showed that $R = 10$ is a feasible cut off to display the effect of different parameter values on belief polarization. Thus, $R$ was set to 10 for all results presented here. Furthermore, we will address the question of how *quickly* the system converges to a state where the absolute majority of agents hold either 0 or 1 as belief. Exemplary simulation results are presented for the confidence parameter in Section 3.5.

#### Influence of Confirmation Bias and Openness

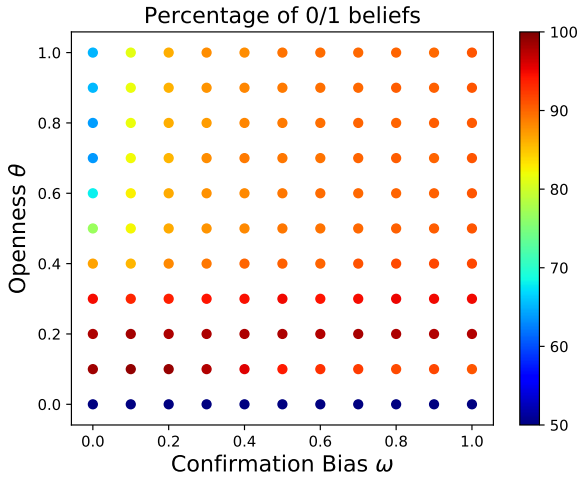



**Figure 3.4.** Influence of openness $\theta$ and confirmation bias $\omega$ on belief polarization. The figure displays the percentage of 0/1 beliefs for $n = 500$ agents, with $n_{\mathrm{NN}} = 15$ interaction neighbors, $\omega \in \{0.0, 0.1, ..., 1.0\}$, $\theta \in \{0.0, 0.1, ..., 1.0\}$ over $R = 10$ rounds. To account for the randomness, results are averaged over 50 simulation runs.

Figure 3.4 displays belief polarization depending on different values of openness $\theta$

and confirmation bias $\omega$. Results show that the percentage of agents who hold 0/1 beliefs increases with higher values of confirmation bias and lower values of openness. Note that in case of $\theta = 0$ there is no belief polarization (polarization values in the last row are about 1%). This is due to the set up of our model, as with $\theta = 0$ the agents happiness will always be $h(a_i) = 0$. Therefore, no updating takes place and thus the belief of each agent remains the same throughout the simulation. This also impacts belief polarization in case of $\theta = 0.1$, in that belief polarization is slightly decreasing for increasing values of $\omega$. Put differently, for very low values of openness and an increasing tendency to seek out confirming information the number of agents that are not updating their beliefs because their nearest neighbors beliefs deviate too much from their own is increasing as well. For $\theta \geq 0.3$ values simulation results show that both, confirmation bias and openness, contribute to belief polarization, with openness having a counteracting effect in case of low confirmation bias values $\omega = [0.0, ..., 0.3]$.

### 3.3.3 Simulating Individual Differences

So far, openness $\theta$ and confirmation bias $\omega$ were set to the same value for all agents. In this section, we will outline how individual differences in cognitive parameters can influence belief polarization.

#### Influence of Individual Differences in Openness

There is a large body of literature on individual differences in openness and their relationship to a variety of concepts, for example, tolerance (Butrus & Witenberg, 2013) or curiosity (Kashdan, Rose, & Fincham, 2004). Curiosity can be defined as a desire for new information. Curiosity induces exploratory behavior to acquire new knowledge (Berlyne, 1954; Litman, Hutchins, & Russon, 2005) and is closely linked to openness (Peterson et al., 2004).

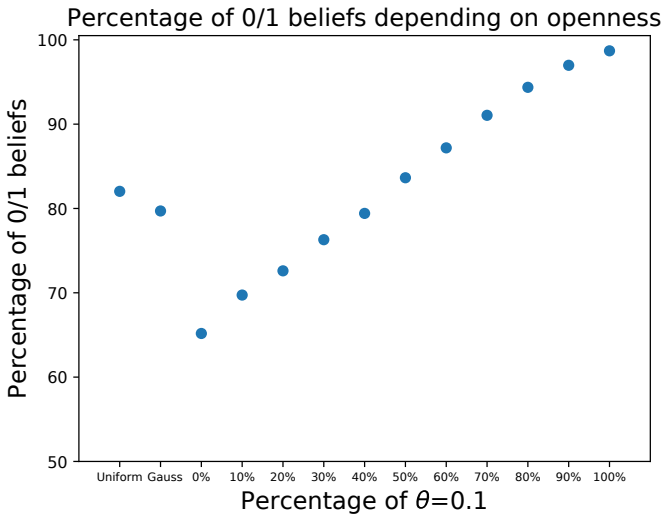



**Figure 3.5.** Influence of individual differences in openness. The figure shows the percentage of 0/1 beliefs hold by $n = 500$ agents, with confirmation bias $\omega = 0$, $R = 10$ rounds, $n_{\text{NN}} = 15$ interaction neighbors. The x-axis displays the proportion of agents with openness $\theta = 0.1$ (low openness) in a population with $\theta = 1.0$ (maximum openness). To account for the randomness, results are averaged over 50 simulation runs.

Figure 3.5 shows that with an increasing proportion of agents who have low openness ($\theta = 0.1$) the percentage of polarized beliefs increases as well. As shown in Figure 3.2, lower values of openness lead to the emergence of echo chambers. Thus, agents place themselves in an environment in which their own beliefs are reinforced and amplified throughout the simulation. Consequently, belief polarization is accelerated. These results are in concordance with the assumption that the formation of echo chambers in social networks contributes to belief polarization in the population (Quattrociocchi, Scala, & Sunstein, 2016; Sunstein, 2018).

As openness approximately follows a Gaussian distribution in populations (McCrae & John, 1992), we included the percentages of polarized beliefs for approximately Gaussian and uniform distributed openness values. In case of uniform distributed openness values, results were slightly lower than those when the population of agents consisted of one half of agents with $\theta = 0.1$ and the other half of agents with $\theta = 1.0$. In case of the Gaussian distributed openness values, percentages of polarized beliefs were about the same as a population consisting of 40% of agents with openness $\theta = 0.1$ and 60% of agents with openness $\theta = 1.0$. However, setting openness for all agents to the maximum ($\theta = 1.0$) yielded by far the lowest percentage of polarized beliefs. These results suggest, that high values of openness decelerate belief polarization.

### Influence of Individual Differences in Working Memory Capacity (WMC)

In all simulation runs so far $n_{\text{NN}}$ was set to 15, i.e., agents updated their beliefs based on the beliefs of all 15 interactions. In other words, agents "remembered" all $n_{\text{NN}}$ encounters. In the following, working memory capacity is integrated into the model by equipping each agent with a memory parameter. Now, only the last $\gamma$ encounters with $\gamma$ = WMC are taken into account. Figure 3.6 displays the influence of WMC on belief polarization.

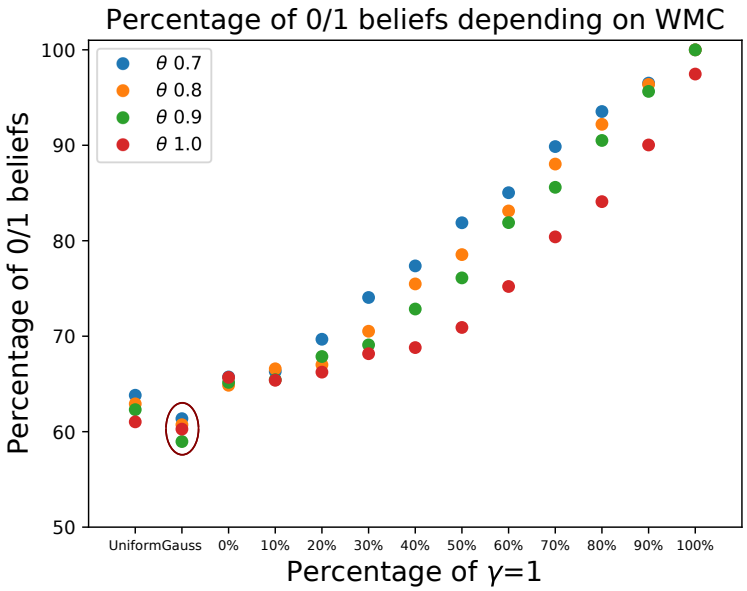



**Figure 3.6.** Influence of WMC on belief polarization. The figure displays the influence of WMC on belief polarization. $n = 500$ agents, with confirmation bias $\omega = 0$, $R = 10$ rounds for four different values of openness $\theta = \{0.7, 0.8, 0.9, 1.0\}$. The x-axis displays the proportion of agents with WMC $\gamma = 1$ (low WMC) in a population with $\gamma_{\text{default}} = 15$. To account for the randomness, results are averaged over 50 simulation runs.

Simulation results show that increasing the proportion of agents with the lowest possible value for WMC ($\gamma = 1$) leads to an acceleration of belief polarization. However, in contrast to individual differences in openness, results show that polarization is *lowest* for uniform and approximately Gaussian ($\gamma \sim \mathcal{N}(7,3)$) distributed working memory capacities. This is rather surprising, as one would expect that the percentage of polarized beliefs would be lowest if all agents are equipped with maximum WMC ($\gamma = 15$). Please note that those are preliminary simulation results for only four values of openness $\theta = \{0.7, 0.8, 0.9, 1.0\}$. Nonetheless, these first results seem to indicate that *Gaussian distributed* working memory capacities decelerate belief polarization.

In the next section, we will outline the integration of our fully parameterized mathematical model of the ACT-R declarative memory module (Manuscript 1). This will enable us to model the cognitive process of each agent in a more realistic way.

## 3.4 Merging the Agent-Based Belief Model with ACT-R

So far, agents' memories were simply modeled by varying the number of beliefs that could be remembered. In order to (a) introduce a more realistic way to model agents' memory and (b) include several cognitive parameters (working memory capacity $W$, memory decay $d$, and retrieval threshold $\tau$), we will outline how our in Manuscript 1 developed fully parameterized mathematical model of the ACT-R declarative memory module can be integrated into the agent-based belief model.

ACT-R is a commonly used cognitive architecture which allows to model cognitive processes (Anderson, 2009). The ACT-R architecture consists of three main components: modules, buffers, and a pattern matcher. A central production system coordinates the interaction between different modules (Anderson, 2009; Borst & Anderson, 2017). Each module is associated with distinct cortical regions, supporting the structure of the architecture (Anderson, 2009) (see Figure 3.7, for an illustration).

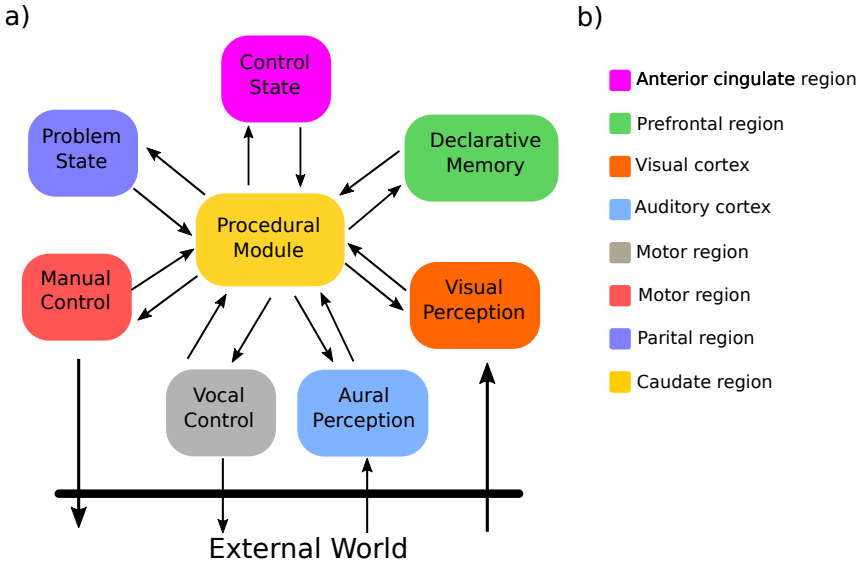



**Figure 3.7.** Sketch of the different modules of the ACT-R architecture. The figure displays the connection of modules in ACT-R 7.0 (a) and their mapping to brain regions (b). For the precise locations, i.e., Talairach coordinates, and fMRI images see Anderson (2009); Borst and Anderson (2017).

In the following section, we will give a short description of the setup of the ACT-R model and how it can be used to investigate the influence of multiple cognitive parameters (working memory capacity $W$, memory decay $d$, retrieval threshold $\tau$) on belief polarization.

### 3.4.1 Mathematical Description of the ACT-R Model

In this section, we will outline the setup of the declarative memory module that describes a more complex simulation of agents' memory. The current implementation simulates the cognitive process of each agent before belief updating (see Algorithm 1, line 7).

**The declarative memory module.** Instances of knowledge are stored in the declarative memory module of the ACT-R architecture. A single element of declarative knowledge is called a chunk $c_{\iota,k}$, with $\iota, k \in \mathbb{N}$. Each chunk contains a number of slots $k$ that hold information. Furthermore, with each chunk a so-called activation value $A_i$ is associated that reflects the usefulness of the stored information for the current task to be solved. In our case, an agent's chunk contains its own belief $b_i(t)$ as well as the beliefs of other agents $b_{l(j)}(t)$ it interacted with, with $j = 1, ...n_{NN}$.

**Production rules used in this model.** Procedural knowledge is modeled in ACT-R by so called production rules (short: *productions*) that allow the interaction between different modules. In our case, the ACT-R model starts with an agent's current belief $b_i(t)$ in round $R$ and the beliefs $b_{l(j)}(t)$ of the surrounding neighbors it interacted with. For an agent to update its belief, encounters are successively recalled:

- **First step:** Request retrieval of chunk necessary for updating belief.
- **Second step:** If there is such a chunk and the activation of this chunk is above the threshold $\tau$, include belief of this agent into belief updating. If there is no such chunk or the activation of the chunk is lower than the threshold $\tau$, interaction with this agent is counted as "forgotten".
- **Third step:** Update own belief.

Whether or not information can be retrieved from the declarative memory depends on the cognitive parameters that are part of the ACT-R architecture. These are working memory capacity $W$, memory decay $d$, and the retrieval threshold $\tau$. In order to investigate the influence of working memory capacity on belief polarization, we extended our mathematical model from Manuscript 1 by including the spreading activation component from the ACT-R architecture (R. Anderson, 1983). Spreading activation reflects the associative nature of the declarative memory module. The activation of chunks holding the same information in one or more of their slots is "spread" between them. For example, the term "birthday" might be associated with a large number of chunks stored in memory. Thus, activation will be divided between the chunks resulting in several possibilities of what kind of chunk could be remembered. In contrast, "birthday of my mother" only refers to one specific chunk. Thus, this chunk would have the highest activation and consequently the highest probability of being remembered. The parameter $W$ controls the amount of spreading activation and corresponds to working memory capacity as has been repeatedly shown (Anderson, Reder, & Lebiere, 1996; Lovett, Daily, & Reder, 2000).

**Algorithm 2:** Retrieval process of interaction neighbors' beliefs for one agent $a$

**Input:** $b_i, b_{l(j)}, \eta_{j,\iota}$ ;
**Output:** $x_j$;
**for** $j \in J$ **do**
  **for** $\iota \in I$ **do**
    $L_\iota := \left(j - t_\iota\right) + T$;
    $B_\iota := \ln\left(\frac{n_\iota}{1-d}\right) - d \cdot \ln\left(L_\iota\right)$;
    $A_\iota := B_\iota + W \cdot \sum_j S_{\iota,1} + \eta_{j,\iota}$;
  **end**
  $\iota^* := \arg\max_\iota A_\iota$;
  **if** $A_{\iota^*} \geq \tau$ **then**
    $x_j := c_{\iota^* 3}$;
  **else**
    break;
  **end**
  **if** $\exists \iota \in I : c_\iota = \left(j, b_i, b_{l(j)}\right)$ **then**
    $n_\iota := n_\iota + 1$;
  **else**
    $n_\iota := 1$;
    $c_\iota := \left(j, b_i, b_{l(j)}\right)$;
    $t_\iota := j$;
  **end**
**end**

Algorithm 2 describes the model dynamics of the declarative memory module for the agent-based belief model. The model parameters are displayed in Table 3.2. For a more detailed description, see Manuscript 1 (Appendix A). The model outlined above could be extended even further by including partial matching (which is controlled by an additional parameter). This would allow agents to "confuse" similar beliefs as well as the order of the agents they last interacted with.

To summarize, equipping each agent with a declarative memory will allow (a) for a more complex simulation of agents' memory, (b) for an investigation of the influence of several other cognitive parameters like memory decay $d$ on belief polarization, (c) for a comparison between modeling WMC in a simple versus a more complex way.

Agents' beliefs are influenced by their openness and cognitive parameters like working memory capacity and confirmation bias. In the next section, we will outline how the cognitive parameter *confidence* can be integrated into the model and discuss simulation results.

**Table 3.2.** Model parameters ACT-R model.

| | | |
|---|---|---|
| **Initial Values** | | |
| $n_{\text{NN}}$ | $\in \mathbb{N}$ | number of neighbors an agent interacted with |
| $n_c$ | $\in \mathbb{N}$ | number of chunks present in the model |
| $I$ | $:= \{1,\ldots,n_c\}$ | index set for chunks |
| $J$ | $:= \{1,\ldots,n_{\text{NN}}\}$ | index set for inputs |
| **Default values of cognitive parameters** | | |
| $\tau$ | $:= -2.5$ | retrieval threshold |
| $d$ | $:= 0.5$ | decay parameter of base level learning |
| $T$ | $:= 0.05$ | time component |
| **Inputs and outputs of ACT-R model** | | |
| $b_{a_{l(j)}}$ | $\in \{0.0, 0.1, .., 1.0\}$ | beliefs of the surrounding neighbors |
| $\eta_{j,\iota}$, | $\iota \in I, j \in J$ | random noise, $\eta \sim U([0,1])$ |
| $x_j$ | $\in \{0.0, 0.1, .., 1.0\}^{N_l}$ | decisions of the model |
| **Declarative memory module** | | |
| $c_{\iota,k}$, | $\iota \in I$, $k \in \{1,2,3\}$ | component $k$ of chunk $\iota$ |
| $n_\iota$, | $\iota \in I$ | number of presentations of chunk $\iota$ |
| $t_\iota$, | $\iota \in I$ | time (i.e. round) of generation of chunk $\iota$ |
| | | 1 = chunk exists, 0 = chunk has not been generated yet |
| $A_\iota$, | $\iota \in I$ | activation of chunk $\iota$ |
| $B_\iota$, | $\iota \in I$ | base level learning of chunk $\iota$ |
| $L_\iota$, | $\iota \in I$ | lifetime of chunk $\iota$ |
| $S_{\iota,k}$, | $\iota \in I$, $k \in \{1,2,3\}$ | strength of association from the components $k$ to chunk $\iota$ |
| $\iota^*$ | | temporary variable for maximum activation index |

## 3.5 Confidence in Climate Change Knowledge

In Manuscript 2, we investigated the accuracy of people's confidence in climate change knowledge in a nationally representative sample ($N = 509$) taken from the German population. Participants were shown nine true or false statements about climate change, and had to state whether or not those statements were correct. Furthermore, participants indicated their confidence in each of their answers. Accuracy of participants' confidence in their own climate change knowledge was assessed by calculating the *relative confidence sensitivity* $M_{ratio}$. $M_{ratio}$ reflects how accurately citizens' insight into the limits of their knowledge is.

We compared citizens' confidence accuracy in climate change knowledge with (a) the accuracy of confidence in science knowledge in a second nationally representative sample ($N = 588$) taken from the German population, and (b) the accuracy of confidence in climate change knowledge in a sample of climate change scientists ($N = 206$). Results showed that citizens' $M_{ratio}$ for their climate change knowledge was only 0.49 (95% CI $[0.33, 0.63]$). Put differently, citizens' confidence sensitivity was only around half of what it could be based on their actual climate change knowledge. Furthermore, relative confidence sensitivity of citizens' climate change knowledge was especially low compared to their science knowledge ($M_{ratio} = .99$, 95% CI $[0.88, 1.16]$). This also hold true for the scientists sample: citizens' relative confidence sensitivity was only about half compared to scientists' relative confidence sensitivity ($M_{ratio} = .95$, 95% CI $[0.85, 1.07]$).

Accurate confidence is important for adequate predictions and decision-making in areas of high uncertainty, e.g., in political (D. D. Johnson, 2009) or medical (Berner & Graber, 2008) areas. Moreover, overconfidence in one's belief (also termed *overprecision*, Moore & Healy, 2008; Moore & Schatz, 2017) makes it less likely to change that belief even when contradicting information is presented (Malmendier, Tate, & Yan, 2010; Ortoleva & Snowberg, 2015; Savion, 2009).

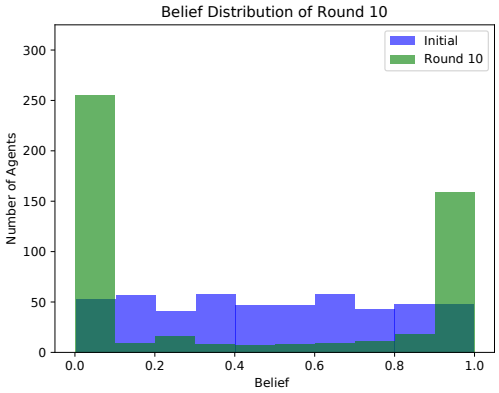
### 3.5.1 Integrating (Over-)Confidence in the Agent-Based Belief Model

One way to integrate confidence into our agent-based belief model is to equip each agent with a confidence parameter $\xi_i \in [0, 1]$. In a first setup, values like belief and position are drawn for each agent from a uniform distribution and confidence in one's knowledge is assumed to change with each new update. The confidence parameter $\xi_i$ determines the probability of whether or not an agent will update its belief based on the belief of its surrounding neighbors. In case of $\xi_i = 0$ the agent will update its belief based on the belief of its surrounding neighbors. In case of $\xi_i = 1$ the agent will not include deviating beliefs into belief updating but rather uses its own belief for updating.

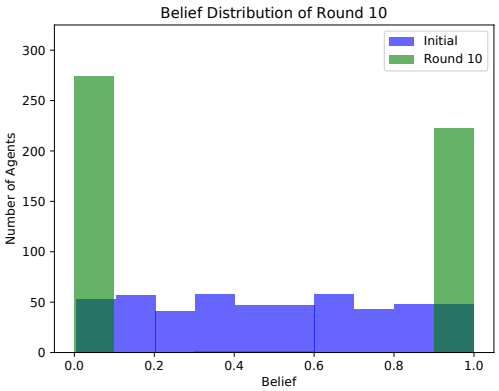
Setting our model into the context of climate change, with $b = 1$ being convinced that climate change is man-made and $b = 0$ believing that climate change has solely natural causes, optimal confidence sensitivity in this specific setting would be if $\xi_i = b_i$. This means agents with $b_i = 1$ would correctly be 100% confident that their belief is correct, while agents with $b_a = 0$ would correctly be 0% confident that their belief is correct and thus more susceptible of changing their beliefs.

Figure 3.8 displays an example of the belief distributions of agents after $R = 10$ rounds in case of (a) the previous model without the confidence parameter, (b) agents' confidences values $\xi_i$ are a v-shaped function of $b_i$: $\xi_i = |b_i - 0.5| + 0.5$. This implies

(a)




(b)




(c)

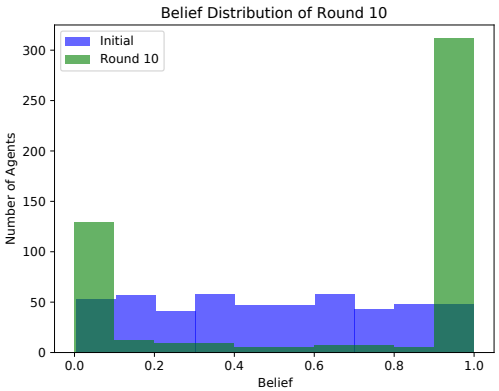


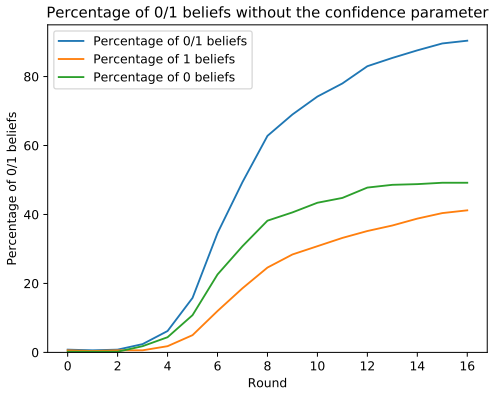

**Figure 3.8.** Belief distribution of $n = 500$ agents, with $n_{NN} = 15$ interaction neighbors, openness $\theta = 1.0$, and confirmation bias $\omega = 0.0$ after $R = 10$ rounds. (a) shows the belief distribution of agents without the confidence parameter, b) shows the belief distribution for agents with $\xi_i$ following a v-shaped function, and (c) shows the belief distribution for agents with optimal confidence sensitivity.

that confidence increases as beliefs are shifting more towards 0 or 1, and (c) optimal confidence sensitivity $\xi_i = b_i$ is assumed.

While confidence values following a v-shape function are even further accelerating belief polarization (48% $b = 1$ and 49% $b = 0$, averaged over 50 simulation runs, total of polarized beliefs: 98%, compared to a total of 65% of polarized beliefs in the model without $\xi$), optimal confidence sensitivity leads to a convergence towards the – in this exemplary setting "correct" belief – that is climate change is man-made (58% $b = 1$ and 13% $b = 0$, averaged over 50 simulation runs).

Furthermore, we investigated how quickly the system converged to a state where more than 90% of agents hold either 0 or 1 as belief. Figure 3.9 displays the number of rounds $R$ the system needs to converge for the three cases: (a) without the confidence parameter the system takes $R = 16$ rounds to converge, (b) for $\xi_i$ following a v-shaped function belief polarization is accelerated, $R = 6$, and (c) in case of optimal confidence sensitivity the number of rounds is $R = 16$. To account for randomness we averaged the number of rounds over 50 simulation runs. Results showed that if agents' confidence values are following a v-shaped function the system converges more than two times faster ($R_{mean} = 6.7$) than in the other two cases ($R_{mean} \approx 17$).

(a)




(b) (c)

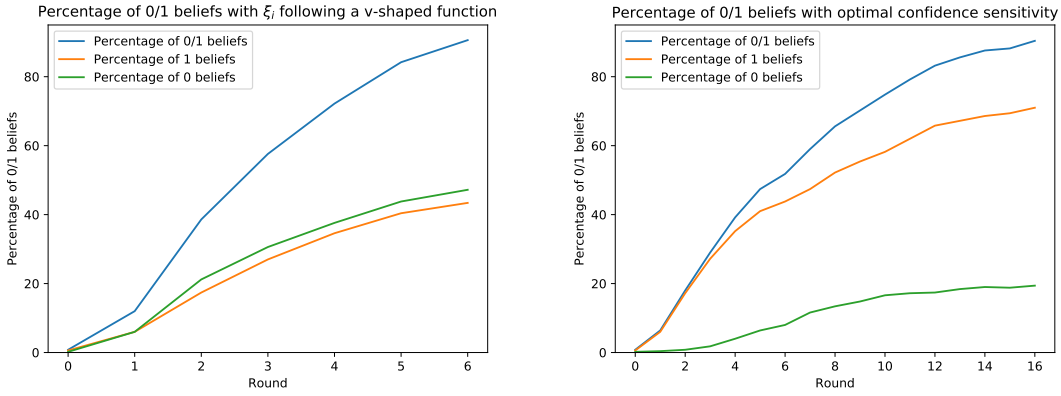



**Figure 3.9.** Percentage of $n = 500$ agents holding either 0 or 1 as belief, with $n_{NN} = 15$ interaction neighbors, openness $\theta = 1.0$, and confirmation bias $\omega = 0.0$. (a) shows $R$ for the model without the confidence parameter, b) shows $R$ in case of $\xi_i$ following a v-shaped function, and (c) shows $R$ in case of optimal confidence sensitivity. The blue line displays the percentage of agents holding 0 or 1 as belief. The orange line displays the percentage of agents holding 1 as belief and the green line the percentage of agents holding 0 as belief.

In Manuscript 2, we showed that citizens' confidence sensitivity regarding their climate change knowledge is only half of what it could be based on their knowledge. The modeling approach outlined above is one way to integrate *confidence sensitivity* into our agent-based belief model and shows how failure of metacognitive awareness could impact belief polarization.

# 4 | Understanding Non-Linear Processes

In Manuscript 3, we investigated people's ability to understand the development of non-linear processes. Using the function-learning (FL) paradigm, it is possible to assess people's understanding of the function-rule underlying those processes by measuring their prediction ability. *Extrapolation studies* consist of a learning and a prediction phase. Participants are shown x-values of a function and have to predict the corresponding y-values. Participants learn through feedback how a process develops over time and then have to predict how this process will progress in the future. People's prediction ability is assessed by calculating the deviation of their predictions from the correct function. However, we argue that while correct predictions clearly do indicate previous rule-learning this does not necessarily hold true for incorrect predictions. We showed in two experiments that about one third of participants who, would be classified as "exemplar-based" learners based on their prediction accuracy in the standard function-learning paradigm, demonstrated correct rule-learning in alternative paradigms.

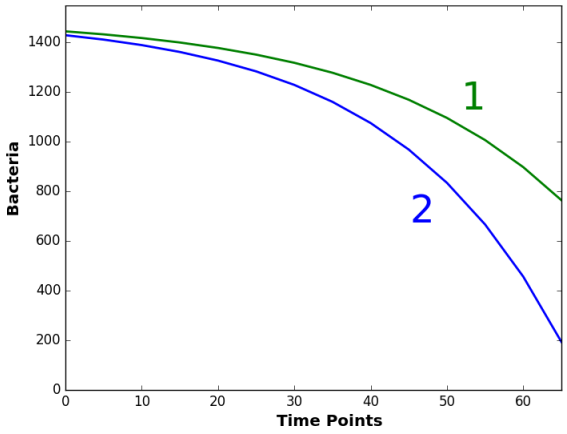



**Figure 4.1.** Functions used in Experiment 2 of Manuscript 3. The figure displays the two functions participants had to predict: (1) "Ain": $y = 1500 - e^{0.045 \cdot (x+50)+2}$ and (2) "Bin": $y = 1500 - e^{0.040 \cdot (x+50)+2}$. The y-axis displays the number of bacteria for a given time point x.

In Experiment 1 we assessed the prediction accuracy of $N = 511$ participants in the standard function-learning paradigm and investigated to what extend those who were classified as "exemplar-based" learners showed correct rule-learning in a *rule-selection task*. In the rule-selection task participants were presented with three graphs (see Figure 4.2) and asked to choose the one which describes the development of the processes best. Prediction accuracy was measured by calculating the *relative root mean squared error* (rRMSE). Based on their prediction accuracy participants fell into different extrapolation style categories. Participants who displayed a linear extrapolation style were classified as "exemplar-based" learners (McDaniel, Cahill, Robbins, & Wiener, 2014).

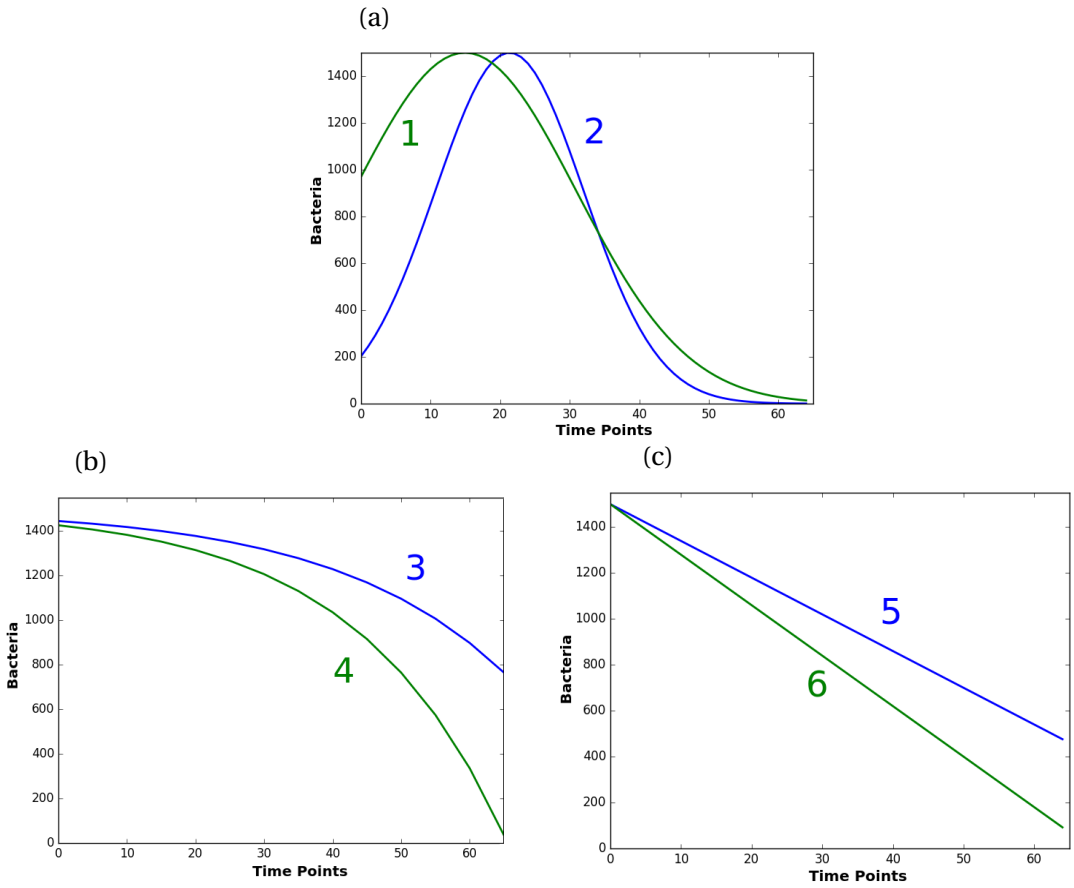



**Figure 4.2.** Rule-selection task. The figure displays three functions with two different slopes each: (a) Gaussian function, b) exponential, and (c) linear. Participants task was to select the function which they thought described the development of the bacteria best.

Results showed that 61% of participants who were classified as ”exemplar-based“ learners based on their prediction accuracy chose the correct function shape and 37% were even able to identify the correct function shape and slope. These results suggest that a substantial proportion of participants who displayed a linear extrapolation style were actually aware of the non-linearity of the processes.

Experiment 1 is limited in two ways: first, the a priori probabilities to guess correctly were much higher in the rule-selection task than in the standard function-learning task. Second, in the standard function-learning task participants were presented the x-y-pairings consecutively while in the rule-selection task participants had access to all function values at the same time. To address these limitations we conducted a second experiment introducing two new conditions in which participants (a) had to draw the function by clicking on a grid (grid condition, see Figure 4.3) and (b) had access to their previously predicted values, that is x-y-pairings were displayed on one screen (summary FL condition). To compare participants' performance in those conditions with the standard function-learning condition maximum and minimum extrapolations were restricted to values between 0 and 1550. Furthermore, the number of clicks on the grid were the same as the number of entries in both of the FL conditions.

In Experiment 2 out of the $N = 918$ participants who completed the experiment the data of $N = 660$ participants were included in the final data set. In the two function-learning conditions prediction accuracy was again measured by calculating the rRMSE.

To asses rule-learning in the grid condition we introduced two approaches: (a) calculating the first derivatives, a method that allows to determine whether participants actually grasped that the processes displayed were *increasingly declining*, and (b) a least squares approach.

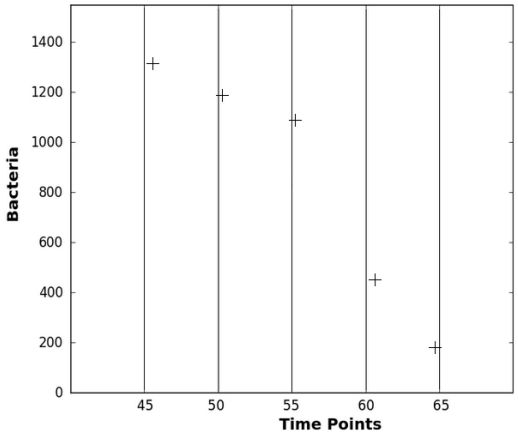



**Figure 4.3.** Grid paradigm. The figure displays the grid task. Participants had to indicate their prediction for each time point by clicking on a grid.

Results showed that about half of participants were classified as rule-based learners in the grid condition compared to less than one third who were classified as rule-based learners in the standard as well as the summary function-learning condition. These results suggest that the proportion of participants who acquired an understanding of the correct function-rule is underestimated in the standard function-learning paradigm. Furthermore, having access to previously entered prediction values did not affect rule-application.

The results of Experiment 1 and 2 suggest that a considerable proportion of participants acquired an understanding of the correct function-rule but failed to apply the learned rule in the prediction phase. Put differently assessing rule-learning by measuring people's prediction ability underestimates the proportion of rule learners.

In both experiments outlined above, we also assessed participants working memory capacity (WMC) and replicated the finding of McDaniel et al. (2014). The authors found that participants' prediction ability is positively associated with WMC. However, as our main focus was on investigating the question whether incorrect predictions do exclude rule-learning, we did not include those findings into Manuscript 3. Therefore, we will shortly present the results of the working memory capacity assessment conducted in the second experiment of our study in Section 4.1. In Section 4.2, we will outline how the findings of Manuscript 3 could be integrated into our ACT-R exemplar-based function-learning model.

## 4.1 Working Memory Capacity in Function-Learning

In the second experiment of Manuscript 3, we assessed participants understanding of the function-rule by comparing the standard function-learning paradigm (standard FL condition) to an alternative paradigm in which participants had to draw the function on a grid (grid condition). We also added a further condition in which participants'

predictions were displayed at the same time instead sequentially (summary FL condition). After a learning phase, participants had to predict the development of two exponential functions (Figure 4.1) with different slopes termed "Ain" and "Bin". Extrapolation accuracy was measured by calculating the relative root mean square error for each participant. To investigate whether participants with low(er) working memory capacity actually grasped the development of the process but were not able to correctly extrapolate because they did not remember the previous numbers they entered in the prediction phase, we assessed working memory capacity using a shortened version of the symmetry span task and compared prediction accuracy in (a) the standard FL ($n = 248$) with (b) the summary FL condition ($n = 247$).

### 4.1.1 WMC Assessment

**Symmetry Span Task.** To assess working memory capacity, we used a shortened version of the symmetry span task (Oswald, McAbee, Redick, & Hambrick, 2015; Unsworth, Brewer, & Spillers, 2009). Participants had to indicate whether the design of a $8 \times 8$ matrix filled with black squares is symmetrical about its vertical axis. After participants' decision, a red square was shown for $0.65s$ within a $4x4$ matrix. This procedure was repeated with either 3, 4, or 5 red squares in total. At the end of each trial, the task was to recall the sequences of red squares by indicating their positions on a grid. In total, participants had to complete 24 trials, consisting of $2 * (3 + 4 + 5)$ sets. Performance, i.e., working memory capacity, was assessed by calculating an abbreviation of the partial storage score. A trial was considered successful if participants were able to recall the positions of the red squares correctly. We did not take into account whether participants also recalled the positions in the correct order.

**Sample tasks.** To make sure participants understood the tasks instruction correctly, we included 4 sample tasks in which 2 red squares needed to be remembered. Participants who were not able to correctly complete the second sample task received the feedback "You did not answer correctly. Please try again." and were shown the same sample task again. If they were still not able to complete the task, they received the correct solution and had to complete a fourth sample task.

**Robustness check: Influence of scrolling.** As some participants in our pretests reported that they had to scroll, since they were not able to view the symmetry span task fully on their screens, we included the question: "Were the boxes with the squares visible in their entirety? Or did you need to scroll?". In the standard FL condition, 58% of the participants reported this issue. Dividing those into two groups (scrolling: yes/no) and conducting a Welch Two Sample t-test, there was no significant difference in performance between participants who did not have to scroll ($M = 2.09$, $SD = 1.84$) and participants who had to scroll ($M = 2.11$, $SD = 1.73$), $t = 0.1$, $p = .92$. In the summary FL condition, 65% of the participants reported this issue. Dividing those into two groups (scrolling: yes/no) and conducting a Welch Two Sample t-test, there was also no significant difference in performance between participants who did not have to scroll ($M = 2.21$, $SD = 1.96$) and participants who had to scroll ($M = 2.28$, $SD = 1.64$), $t = 0.3$, $p = .76$. These results suggest that scrolling did not influence performance in either of the two conditions.

### 4.1.2 Experimental Results

To investigate whether having access to all previous predicted values impacts prediction accuracy and if there is a relationship with working memory capacity, we compared prediction accuracy in (a) the standard function- learning, with (b) the summary function-learning condition by calculating a multivariable linear regression. Results indicated that for both processes Ain and Bin there was a significant effect between prediction accuracy and WMC ($t_{\text{Ain}} = -6.02$, $p_{\text{Ain}} < .001$; $t_{\text{Bin}} = -6.37$, $p_{\text{Bin}} < .001$). However, there was no significant relationship between prediction accuracy and condition ($t_{\text{Ain}} = 0.17$, $p_{\text{Ain}} = .87$; $t_{\text{Bin}} = 0.01$, $p_{\text{Bin}} = .10$). Results suggest that while working memory capacity may be a restricting factor for prediction accuracy (McDaniel et al., 2014), it seems to be less relevant for successful rule-application during prediction.

## 4.2 First Steps Towards Modeling Function-Learning in ACT-R

To investigate the influence of individual differences in cognitive parameters like working memory capacity on prediction performance, we have implemented an exemplar-based function-learning model in ACT-R (adapted version of EXAM, DeLosh, Busemeyer, & McDaniel, 1997). In the current implementation (**Model Setup A**), the model starts by learning (x,y)-pairs (x: time, y: feedback) in the learning phase and guesses possible outcomes randomly. The (x,y)-pairs together with the model predictions are stored in the declarative memory. The model's chunks $c_{\iota,k}$ have three slots $k$, with $c_{\iota,1}$ : x-value, $c_{\iota,2}$: prediction, and $c_{\iota,3}$: feedback (y-value). During the prediction phase, the model starts to actively retrieve the former learned (x,y)-pairs (y-values are either the feedback values or, as in the prediction phase no feedback is given, the extrapolated values) and to extrapolate linearly. Thus, productions implemented in this model are as follows:

- **First step:** Request retrieval of first chunk necessary for prediction.
- **Second step:** If there is such a chunk and the activation of this chunk is above the threshold $\tau$, request retrieval of second chunk necessary for prediction. If there is no such chunk or the activation of the chunk is lower than the threshold $\tau$, chose random value ($y \sim U([0, n_p]), n_p \in \mathbb{N}$) to extrapolate.
- **Third step:** If the second chunk is also retrieved successfully, extrapolate linearly through the retrieved (x,y)-pairs.

Whether or not (x,y)-pairs necessary for prediction are retrieved successfully, depends on the cognitive parameters, working memory capacity $W$, memory decay $d$ and the retrieval threshold $\tau$. However, while *Model Setup A* (see Figure 4.4) allows for an exploration of the influence of individual differences regarding cognitive parameters on linear prediction performance, we propose an extension of the model (**Model Setup B**) based on the results of Manuscript 3.

**Rule-based prediction.** One way to allow for rule-based prediction in our model is to use the utility mechanism implemented in the ACT-R architecture: if several productions in ACT-R match the same goal, the one with the highest utility is selected. A utility value can be set for each production in advance. However, ACT-R also includes

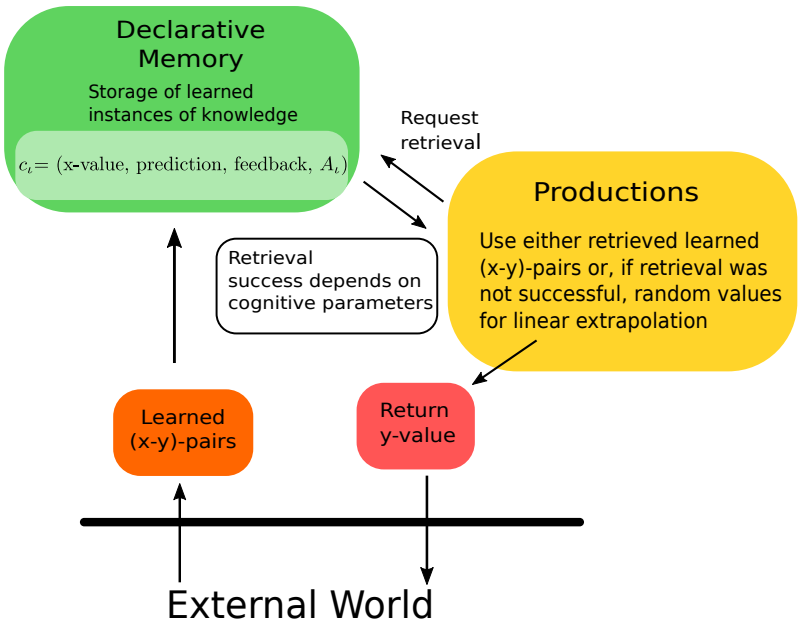



**Figure 4.4.** Description of Model Setup A. The figure displays how the model retrieves the learned (x,y)-pairs from the declarative memory and extrapolates linearly.

a mechanism that allows utility learning, that is the utility of productions increases with each production application [1]:

$$\Theta_v(n) = \Theta_v(n-1) + \alpha[\Lambda_v(n) - \Theta_v(n-1)], \tag{4.1}$$

with $\Theta_v(n)$ being the utility of a production $v$ after its $(n-1)$th application. $\Lambda_v(n)$ is the reward the production receives for its $n$th application and $\alpha$ (with $\alpha \in [0,1]$) is the learning rate. Thus, pre-stored function-rules are associated with an utility that increases with each successful (that is positively rewarded) application. The setup of the model for the learning phase is as follows:

1. The model starts with a number of productions $v \in \mathbb{N}$ containing pre-stored function-rules that have an equal probability to be chosen.

2. During the *learning phase*, the model (at first randomly $v \sim U([0, n_p])$) selects one production $v$ for prediction. With each production, a reward $\Lambda_v$ is associated, based on how close the extrapolated y-values is to the feedback. That is, selecting the production containing the correct function-rule is associated with the highest reward and thus the probability for choosing this production again is increased.

3. After choosing and applying the function-rule, a chunk $c_{\iota,k}$ with the following slot is generated: $c_{\iota,1}$: $v_i$ (encoding of function-rule applied).

In *Model Setup B* (see Figure 4.5), not only the utility of productions are reinforced but also the activation values of chunks containing the encoding of the type of function-rule applied are increasing with the frequency of function-rule application. Thus, it is more likely to retrieve those chunks during the prediction phase. During the learning phase, the model applies different function-rules and learns the correct or at least a

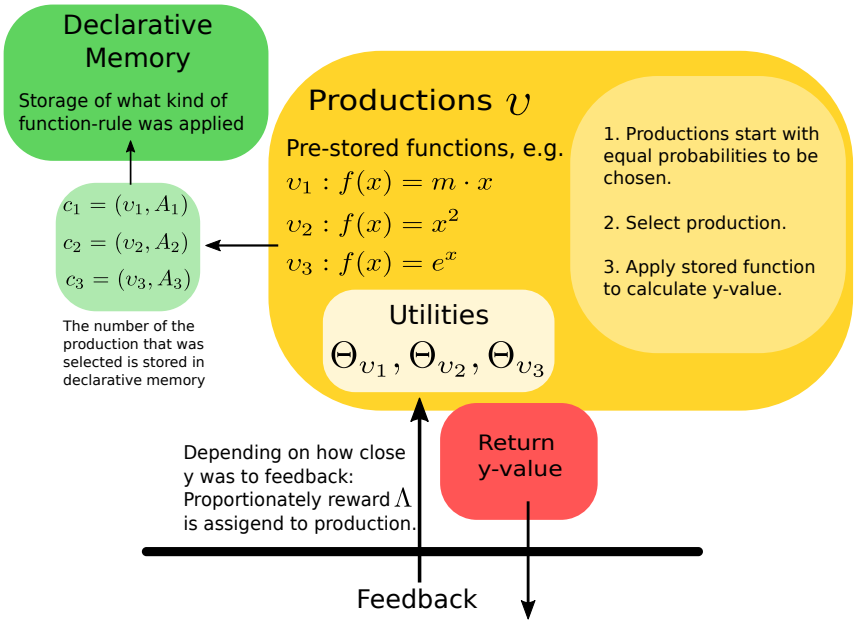



**Figure 4.5.** Description of Model Setup B. The figure displays examples of pre-stored function-rules (linear, quadratic, and exponential) as well as how a reward $\Lambda$ is assigned to productions based on feedback during the learning phase. The chunks $c_i$ and their activation values $A_i$ are stored in the declarative memory module.

function-rule similar to the correct function through feedback. How quickly the model learns a rule depends on the parameter $\alpha$ (learning rate).

**Application-failure.** What kind of production is chosen during the *prediction phase* depends on the $v$-values stored in the chunks. The **Model Setup C** for the prediction phase (see Fig 4.6) is as follows:

1. Request retrieval of $v$-value that is associated with the correct function-rule. If the model successfully learned the correct function-rule during the learning phase, this chunk should have the highest activation value $A_i$ and thus would be retrieved (if $A_i \geq \tau$).

2. Otherwise, a chunk encoding a different function-rule would be retrieved (if $A_i \geq \tau$).

3. In case of $A_i < \tau$, the model would choose a random number for prediction.

As there is no feedback in the prediction phase, the model has no access to the correct y-values and thus might apply a function that returns slightly or even vastly deviating y-values for prediction. In case of the model having learned the correct function-rule, application failure could be modeled by introducing the partial matching parameter $P$. $P$ allows for chunks holding similar $v$-values to be "confused". That is, depending on the values set for $P$, the model either accurately retrieves the chunk with the correct $v$-value and thus the activation of the chunk containing the encoding for the correct function-rule is increased or a chunk with a similar $v$-value is retrieved which, increases the probability to apply a production containing a different function during the prediction phase.

[1]ACT-R 7.0 Tutorial. Unit 6: Selecting Productions on the Basis of Their Utilities and Learning these Utilities

Even if the model has learned the correct or at least a function-rule that is similar to the correct rule, it might be not able to apply this rule during the prediction phase due to the values of $P$.

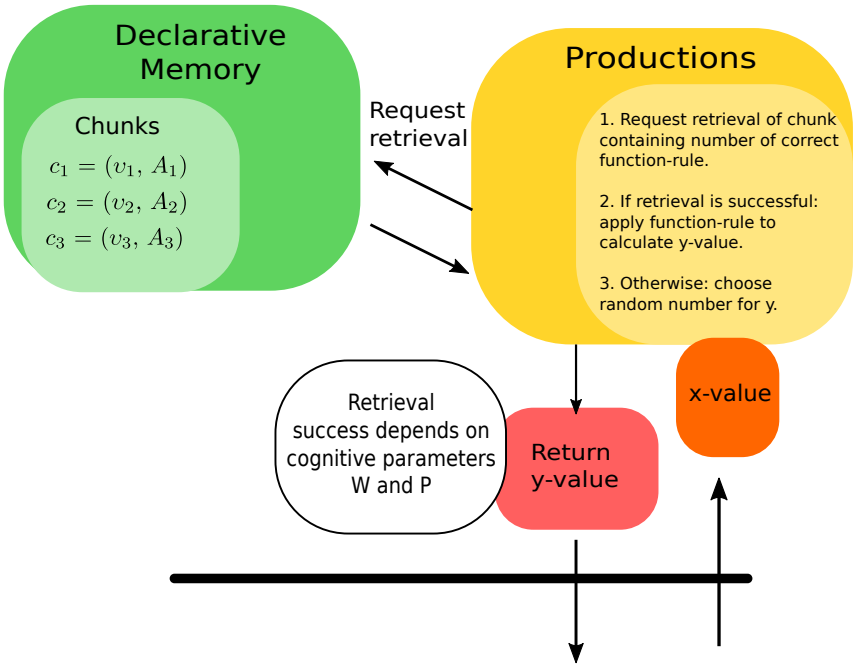



**Figure 4.6.** Description of Model Setup C. The figure displays how the model retrieves the information about the applied function-rules during the prediction phase. Retrieval success depends on the cognitive parameters working memory capacity W and partial matching P.

**Influence of working memory capacity.** In *Model Setup C*, the working memory capacity parameter $W$ influences prediction accuracy such that high values of $W$ increase activation of the chunk holding the information that encodes the correct rule. Working memory capacity thus impacts the model's function-rule choice. Whether the model has access to previous predicted values, does not play any role regarding prediction accuracy.

The model setup outlined above is a first step towards integrating our findings of Manuscript 3 into an ACT-R function-learning model. In concordance with our experimental results presented in Section 4.1 working memory capacity does influence prediction accuracy while correct rule-application does not rely on the ability to remember previous prediction values.

**Future steps: Parameter estimation and sensitivity analysis.** In order to estimate working memory capacity $W$, partial matching $P$ and learning rate $\alpha$ we want to use the interface implemented by Kurt (2017). This will allow us to connect the ACT-R architecture to NLopt, an open-source library (S. G. Johnson, 2010) that provides the necessary derivative-free optimization algorithms. Furthermore, we want to conduct a sensitivity analysis to determine the impact the parameters have on model performance. As the aim of this thesis is to present the experimental results of Manuscript 3 and outline how those can be integrated into an ACT-R model, the rather extensive implementation necessary for both analyses would go beyond the scope of the present work.

# 5 General Discussion

The main goals of this thesis were to present modeling approaches in order to investigate the influence of cognitive parameters on belief polarization in an agent-based belief model (Manuscript 1 and 2) and to outline how the findings from the function-learning experiments presented in Manuscript 3 can be incorporated into an ACT-R function-learning model. Both models allow for the investigation of different cognitive parameters (e.g., working memory capacity) as well as for simulating the impact of individual differences on model performance.

## 5.1 Summary of Core Findings: Belief Polarization

**Influence of Confirmation Bias.** Confirmation bias is an important driver of belief polarization (Lord, Ross, & Lepper, 1979) and is closely related to the formation of echo chambers (Quattrociocchi et al., 2016). Thus, the influence of confirmation bias on belief polarization has been the focus of several agent-based models in recent years (Fryer et al., 2018; Ngampruetikorn & Stephens, 2016; Sobkowicz, 2017). In our model, confirmation bias is integrated by a weighting parameter on the belief dimension. Simulation results showed that high values of confirmation bias accelerated belief polarization given medium to high openness values.

Even though there is substantial literature on confirmation bias and its impact on, i.a., belief polarization and decision-making, research focusing on *individual differences* seems to be rather sparse. Rassin (2008), for example, developed a 10-item self-report to assess individual differences in the tendency to seek out confirming information. Doll, Hutchison, and Frank (2011) examined this tendency from a neuroscientific perspective showing that how susceptible one is to confirmation bias is predicted by dopaminergic genes. As our model allows to simulate individual differences, we want to investigate the influence of individual differences in confirmation bias on belief polarization in the future.

**Influence of Openness.** Being open to new experiences is associated with a range of characteristics, among others, curiosity and tolerance. High levels of openness, for example, lead to more inter-group contact initiation (Jackson & Poulsen, 2005). While tolerance can be defined as acceptance of diversity (Oberdiek, 2001), this does not necessarily include the willingness to try out new practices or integrate deviating beliefs into one's own concepts. In our model, the openness parameter not only controls how much deviation from its own belief an agent "tolerates" in that agents do not "move away" when being set into a heterogeneous environment but goes beyond that as the parameter influences whether agents update their beliefs based on their surrounding neighbors. Simulation results showed that low values of openness led to the emergence of echo chambers while high values of openness decelerated belief polarization. When

introducing approximately Gaussian distributed individual differences of openness, percentages of polarized beliefs were about the same as a population consisting of 40% of agents with openness $\theta = 0.1$ and 60% of agents with openness $\theta = 1.0$. Setting openness for all agents to the maximum value resulted in the lowest percentage of polarized beliefs. Thus, being more tolerant and taking deviating views into account slowed down the process of (perhaps prematurely) shifting one's belief towards 0 or 1.

**Limitations and Next Steps.** Brandt, Chambers, Crawford, Wetherell, and Reyna (2015) however, showed that even though openness was associated with tolerance, this effect was moderated by the conventionality (e.g., majority/minority) of the target group. Experimental results suggested that scoring high on openness did not mean that one is generally open towards others that are different. Rather openness was constrained by the type of target group and the perceived differences associated with that group. Therefore, introducing bounded-openness into our model would be a feasible next step. This would be particularly interesting when comparing the percentages of belief polarization of the Gaussian distributed openness values with those for constrained maximum openness.

**Information Processing and Working Memory Capacity.** Being able to adequately process information is a necessary prerequisite for informed decision-making. However, there is an abundance of constraints and biases regarding how information is processed (Gigerenzer & Goldstein, 1996; Kahneman, Slovic, Slovic, & Tversky, 1982; McClelland & Rumelhart, 1985; Shah & Oppenheimer, 2008; H. A. Simon, 1955, 1990). The constrained capacity of working memory, for example, limits performance in cognitive tasks. Working memory capacity (WMC) correlates with a range of cognitive abilities like attention and intelligence (Conway, Kane, & Engle, 2003; Engle, Kane, & Tuholski, 1999; Kane, Bleckley, Conway, & Engle, 2001) and plays an important role in decision-making (Bechara & Martin, 2004; Fletcher, Marks, & Hine, 2011; Furley & Memmert, 2012; Hinson, Jameson, & Whitney, 2003).

In many agent-based models, information is updated using a Bayesian approach. That is, fully rational agents are assumed (Baker, Saxe, & Tenenbaum, 2011; Zeng & Sycara, 1998). Even though there is evidence that people do follow Bayesian principles in their information processing (Griffiths & Tenenbaum, 2006; Kersten, Mamassian, & Yuille, 2004), there is also substantial literature showing that assuming Bayesian rationality falls short of explaining why human reasoning is often non-rational, based on heuristics, or biased (Albert, 2009; Eberhardt & Danks, 2011; Grether, 1992; Merkle & Weber, 2011). Nonetheless, in line with Acemoglu and Ozdaglar (2011) and Moore and Healy (2008), we consider the Bayesian approach a feasible starting point for our model.

Wilson (2014) showed with an agent-based model in which agents chose between two actions based on informative signals, how introducing a finite memory can account for systematic biases (e.g., confirmation bias). Wilson (2014) concluded that even though people's reasoning is biased, many of those biases are not necessarily in conflict with the assumption of Bayesian rationality given a limited memory.

Horváth, Kovářík, and Mengel (2012) showed in a cooperative agent-based model that limited working memory can be beneficial for the evolution of cooperation. Similar to Horváth et al. (2012), we included a working memory capacity parameter restricting the number of nearest neighbors beliefs an agent "remembers". Simulation results showed that increasing the percentage of agents with the lowest working memory capacity in a population consisting of otherwise agents with maximum WMC accelerated belief polarization. Surprisingly, introducing approximately *Gaussian* distributed

working memory capacities resulted in the lowest polarization values. Put differently, individual limitations in working memory capacity seem to be profitable for the overall population if a slow convergence towards extreme beliefs is the desired outcome.

**Limitations and Next Steps.** These results are preliminary for two reasons: First, even though the choice of maximum WMC is not unreasonable, it is still arbitrary. Second, simulations were only run for high values of openness, as low openness values accelerate belief polarization and thus, an effect of WMC might not be observable. In order to make a more general statement about the impact of WMC limitations on belief polarization, further simulations for different maximum working memory capacities and a wider range of openness values are required.

Like working memory capacity, openness is associated with intelligence and could even be interpreted as cognitive ability (McCrae & Costa Jr, 1997; Moutafi, Furnham, & Crump, 2006). Therefore, running simulations where openness values align with working memory capacity might be another feasible next step.

**The ACT-R Declarative Memory Module.** In Section 3.4, we outlined how the mathematical reformulation of the ACT-R declarative memory module developed in Manuscript 1 can be extended by integrating the spreading activation component of the ACT-R architecture. As has been shown by Anderson et al. (1996) and Lovett et al. (2000), the parameter controlling the amount of activation that is spread throughout the chunks corresponds to working memory capacity. Conceptualizing working memory as a finite resource that is spread among the instances of knowledge to be maintained in memory is supported, for example, by Ma, Husain, and Bays (2014).

Equipping each agent with a cognitive architecture could, on one hand, replace the Bayesian belief updating process. On the other hand, it could allow simulating the influence of not only working memory capacity but also a range of other cognitive parameters, like memory decay, on belief polarization. This would allow us (a) to compare different approaches of modeling information processing and (b) to model agents' memories in a more realistic way.

**Influence of (Over-)Confidence.** Accurate confidence in one's knowledge is important for adequate predictions and decision-making, for example, in political (D. D. Johnson, 2009) or managerial (M. Simon & Houghton, 2003) areas. In Manuscript 2, we showed that citizens had no adequate insight into the accuracy of their climate change knowledge. That is, their relative *confidence sensitivity* was only half of what it could be given their climate change knowledge.

In Section 3.4, we integrated our findings of Manuscript 2 into our agent-based belief model by introducing a confidence parameter. The confidence parameter determines the probability of whether an agent will update its belief based on the beliefs of its surrounding neighbors. Simulation results showed that optimal confidence sensitivity led to a convergence of beliefs toward the correct belief (in our example, climate change is man-made). Introducing a v-shaped confidence function in which confidence increases with the beliefs lead to an acceleration of belief polarization. Put differently, optimal metacognitive awareness resulted in a convergence towards the correct belief in the agent population while high confidence in false beliefs resulted in an acceleration of belief polarization.

**Limitations and Next Steps.** One assumption made in our model set up is that agents' confidences change with each new update. That confidence in one's knowledge changes with new information is supported, for example, by Tsai, Klayman, and Hastie (2008). However, this does not necessarily imply that a change occurs with *each* new

update. Confidence could also change after the second or third update. Thus, one approach would be to vary the number of updates after which confidence is adapted to investigate whether this influences belief polarization.

## 5.2 Summary of Core Findings: Function-Learning

Accurately understanding non-linear behavior of processes is an important skill, not only for tasks in daily life but also when it comes to larger scale processes like climate change (Swim et al., 2011). In our two function-learning experiments (Manuscript 3), participants demonstrated accurate *understanding* of the function-rule of exponential processes in alternative function-learning paradigms despite being classified as "exemplar-based" learners based on their prediction accuracy in the standard function-learning paradigm. One possible explanation for this could be that incorrect predictions in the standard function-learning paradigm reflect alternative processes, such as a failure to correctly apply the learned rule.

In Section 4.1, we presented further experimental results showing that even though working memory capacity is associated with prediction accuracy, not being able to remember previous predictions seems to have no influence on correct rule-application. In Section 3.4, we showed how those results as well as the findings from the experiments presented in Manuscript 3 could be integrated into an ACT-R function-learning model.

**An ACT-R Function-Learning Model.** Our model outlined in Section 3.4, makes use of the ACT-R utility learning mechanism. During the learning phase, application of the correct function is rewarded and thus, the probability to apply this function increases. Application failure during the prediction phase is modeled as failure to retrieve the correct rule.

Three cognitive parameters are essential to whether our proposed ACT-R function-learning model successfully learns and applies a function-rule: the learning rate $\alpha$, the partial matching parameter $P$, and the working memory capacity $W$. The learning rate determines how quickly the model learns the correct rule, or a rule producing similar results, during the learning phase. The partial matching parameter $P$ and the working memory capacity $W$ influence whether the learned rule can be correctly applied in the prediction phase. Both parameters impact the probability to retrieve the correct function-rule from memory. As high values of $W$ and $P$ are having a counteracting effect on the retrieval probability, one could argue that integrating both, $P$ and $W$, into the model might be somewhat redundant. Rutledge-Taylor, Lebiere, Thomson, Stazewski, and Anderson (2012), however, showed that an exemplar-based categorization ACT-R model with only the partial matching component can not account for participants' performance in a categorization task. Combining both partial matching ($P$) and spreading-activation ($W$) seems to describe actual participants' performance more appropriately.

**Limitations and Next Steps.** This model is limited regarding the learning phase, as function-rules are not induced from the learned values but learned values are rather used to reinforce pre-stored rules. Thus, a further extension of the model is necessary to capture this process. Nonetheless, the model captures our main findings of Manuscript 3 as well as the findings presented in Section 4.1, as it allows to simulate (a) application failure and (b) rule-application being independent from the ability to remember previous prediction values.

## 5.3 Future Research

**Different Initial Belief Distributions.** Initial beliefs were uniformly distributed in all simulation runs presented in this thesis. Thus, a next step would be to simulate a variety of different initial belief distributions in the population by using the probability density function (pdf) of the beta distribution. The beta distribution is defined as

$$B(\alpha,\beta) = \int_0^1 z^{\alpha-1}(1-z)^{\beta-1}dz, \tag{5.1}$$

with $\alpha,\beta > 0$. The pdf of the beta distribution is defined as,

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1},\ 0 \le x \le 1, \tag{5.2}$$

with $\alpha,\beta > 0$ (Gupta & Nadarajah, 2004). The parameters $\alpha$ and $\beta$ determine the shape of the distribution (Figure 5.1).

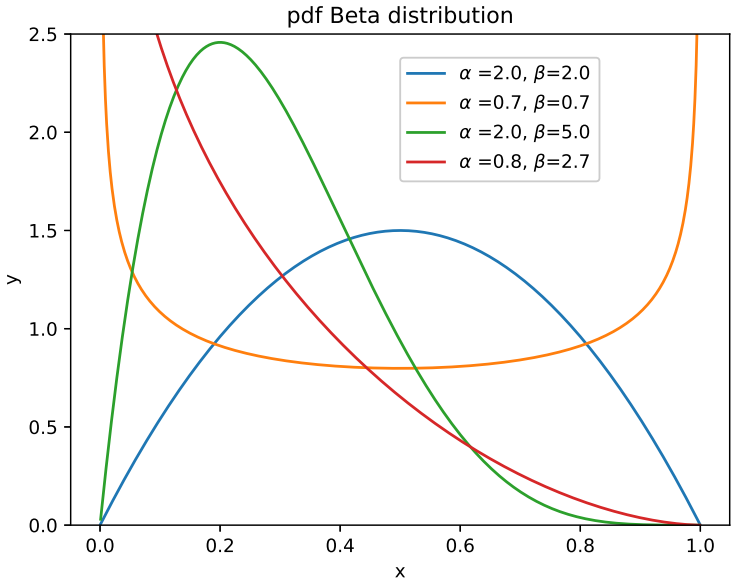



**Figure 5.1.** Examples of the pdf of the beta distribution. The figure displays the pdf for different values of $\alpha$ and $\beta$. For example, $\alpha = 0.7$, $\beta = 0.7$ returns a u-shaped and $\alpha = 2.0$, $\beta = 2.0$ a bell-shaped distribution.

Aside from simulating how beliefs are distributed in a population in a more realistic way, this approach also allows to explore if there is an interaction between belief distributions and cognitive parameters. Thus, we could investigate what kind of parameter combinations are optimal for a convergence of agents' beliefs given a certain initial belief distribution.

**Simulating "Real" Agents: Experimental Approach.** In order to obtain real parameter values for our model, we extended the experiment of Fryer et al. (2018) by including measurements of working memory capacity (Oswald et al., 2015; Unsworth et al., 2009), personality (Rammstedt & John, 2007), and confidence. Fryer et al. (2018) presented participants with research summaries that were either providing evidence for, against, or were neutral regarding a certain topic and assessed their beliefs. Results showed that

participants interpreted the summaries based on their prior beliefs. Using the same experimental setup and including the above mentioned measurements, allows us to (a) obtain real values regarding working memory capacity, openness, and confidence as well as real initial belief distribution, and (b) assess whether there is a relationship between the cognitive parameters and belief updating. Currently, the experiment is prepared to be run for $N = 400$ participants on the MTurk online platform.

**Application of Mathematical Optimization Methods.** In order to assess the influence of the introduced cognitive parameters in both models on model performance, the application of mathematical optimization methods is mandatory.

In agent-based modeling, heuristic optimization approaches (Gilli & Winker, 2003; Oremland & Laubenbacher, 2014; Thiele, Kurth, & Grimm, 2014) are the standard. These heuristic approaches, however, are limited regarding processing time and lack of information about whether an optimum was actually found (for a more detailed discussion see Manuscript 1). Even though there are some approaches to reduce computational complexity (Hinkelmann, Murrugarra, Jarrah, & Laubenbacher, 2011; Kim, Lee, & Levy, 2008), developing more efficient methods is still an important task that needs to be addressed in the future. Regarding our ACT-R function-learning model, we want to use the interface implemented by Kurt (2017) that allows for the application of derivative-free optimization methods.

Applying mathematical optimization methods allows us to (a) make quantitative statements about optimal cognitive parameter values for different scenarios (for example, for different initial belief distributions), (b) explore the influence of the different parameters on model performance, and (c) compare optimal model performance with actual human behavior.

## 5.4 Conclusion

Understanding the cognitive processes underlying belief polarization and function-learning can deliver insight into a range of real-world problems such as climate change. This is because accurate understanding of non-linear processes and being able to adapt one's own belief are necessary prerequisites to make informed decisions, which in turn might influence our future. In this thesis, I provided a framework that allows for modeling belief-polarization and outlined how our experimental results could be integrated into an ACT-R function-learning model. Both models allow for the investigation of different cognitive parameters as well as for simulating the impact of individual differences on model performance. They thus might contribute to a better understanding of both phenomena.

# References


Acemoglu, D., & Ozdaglar, A. (2011). Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, *1*(1), 3–49.

Albert, M. (2009). Why Bayesian rationality is empty, perfect rationality doesn't exist, ecological rationality is too simple, and critical rationality does the job. *Rationality, Markets and Morals*(3), 49–65.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91*(1), 112–149.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford: Oxford University Press.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, *30*(3), 221–256.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, pp. 2469–2474). Austin, Texas: Cognitive Science Society.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542.

Baumgaertner, B. O., Tyson, R. C., & Krone, S. M. (2016). Opinion strength influences the spatial dynamics of opinion formation. *The Journal of Mathematical Sociology*, *40*(4), 207–218.

Bayes, T. (1763). LII. an essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical Transactions of the Royal Society of London*(53), 370–418.

Bechara, A., & Martin, E. M. (2004). Impaired decision making related to working memory deficits in individuals with substance addictions. *Neuropsychology*, *18*(1), 152–162.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology. General Section*, *45*(3), 180–191.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5), S2–S23.

Borst, J. P., & Anderson, J. R. (2017). A step-by-step tutorial on using the cognitive architecture ACT-R in combination with fMRI data. *Journal of Mathematical Psychology*, *76*, 94–103.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences*, *114*(40), 10612–10617.

Brandt, M. J., Chambers, J. R., Crawford, J. T., Wetherell, G., & Reyna, C. (2015). Bounded

openness: The effect of openness to experience on intolerance is moderated by target group conventionality. *Journal of Personality and Social Psychology, 109*(3), 549–568.

Butrus, N., & Witenberg, R. T. (2013). Some personality predictors of tolerance to human diversity: The roles of openness, agreeableness, and empathy. *Australian Psychologist, 48*(4), 290–298.

CBSNews. (2019). Measles outbreak fueled by anti-vaccination movement, infectious disease expert says. *CBS News. Available at http://www.cbsnews.com/news/measles-outbreak-anti-vaccination-movement-dr-anthony-fauci/. Accessed March 14, 2019.*

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7*(12), 547–552.

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., & Way, R., Jacobs, P., Skuce, A. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters, 8*(2), 024024.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90*(2), 272–292.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 968–986.

Dixit, A. K., & Weibull, J. W. (2007). Political polarization. *Proceedings of the National Academy of Sciences, 104*(18), 7351–7356.

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *Journal of Neuroscience, 31*(16), 6188–6198.

Dubé, E., Vivion, M., & MacDonald, N. E. (2015). Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert Review of Vaccines, 14*(1), 99–117.

Duggins, P. (2014). A psychologically-motivated model of opinion change with applications to American politics. *arXiv preprint arXiv:1406.7770.*

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines, 21*(3), 389–410.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 102–134). New York: Cambridge University Press.

Fletcher, J. M., Marks, A. D., & Hine, D. W. (2011). Working memory capacity and cognitive styles in decision-making. *Personality and Individual Differences, 50*(7), 1136–1141.

Fox, M. (2019). Measles outbreaks make 2018 a near-record year for U.S. *NBC News. Available at http://www.nbcnews.com/storyline/measles-outbreak/measles-outbreaks-make-2018-near-record-year-u-s-n961276. Accessed March 14, 2019.*

Fryer, R. G., Harms, P., & Jackson, M. O. (2018). Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association*, 1–32.

Furley, P. A., & Memmert, D. (2012). Working memory capacity as controlled attention in

tactical decision making. *Journal of Sport and Exercise Psychology*, *34*(3), 322–344.

Gabler, N. (2016). The internet and social media are increasingly divisive and undermining of democracy. *Alternet. Available at http://alternet.org/culture/digital-divide-american-politics. Accessed March 14, 2019*.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gilli, M., & Winker, P. (2003). A global optimization heuristic for estimating agent based models. *Computational Statistics & Data Analysis*, *42*(3), 299–312.

Grether, D. M. (1992). Testing Bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, *17*(1), 31–57.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Gupta, A. K., & Nadarajah, S. (2004). *Handbook of beta distribution and its applications*. Boca Raton: CRC Press.

Hajamini, M. (2015). The non-linear effect of population growth and linear effect of age structure on per capita income: A threshold dynamic panel structural model. *Economic Analysis and Policy*, *46*, 43–58.

Hinkelmann, F., Murrugarra, D., Jarrah, A. S., & Laubenbacher, R. (2011). A mathematical framework for agent based models of complex biological networks. *Bulletin of Mathematical Biology*, *73*(7), 1583–1602.

Hinson, J. M., Jameson, T. L., & Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 298–306.

Horváth, G., Kovářík, J., & Mengel, F. (2012). Limited memory can be beneficial for the evolution of cooperation. *Journal of Theoretical Biology*, *300*, 193–205.

IPCC. (2018). *IPCC, 2018: Summary for Policymakers*. Retrieved from https://www.ipcc.ch/sr15/chapter/summary-for-policy-makers/.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology a social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431.

Jackson, J. W., & Poulsen, J. R. (2005). Contact experiences mediate the relationship between five-factor model personality traits and ethnic prejudice. *Journal of Applied Social Psychology*, *35*(4), 667–685.

Johnson, D. D. (2009). *Overconfidence and war*. Cambridge, Massachusetts: Harvard University Press.

Johnson, S. G. (2010). *The NLopt nonlinear-optimization package*. Available at http://ab-initio.mit.edu/nlop.

Jost, J. T. (2006). The end of the end of ideology. *American Psychologist*, *61*(7), 651.

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, Massachusetts: Cambridge University Press.

Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, *130*(2), 169–183.

Kashdan, T. B., Rose, P., & Fincham, F. D. (2004). Curiosity and exploration: Facilitating positive subjective experiences and personal growth opportunities. *Journal of Personality Assessment*, *82*(3), 291–305.

Kata, A. (2012). Anti-vaccine activists, Web 2.0, and the postmodern paradigm–an overview of tactics and tropes used online by the anti-vaccination movement.

*Vaccine, 30*(25), 3778–3789.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271–304.

Kim, P. S., Lee, P. P., & Levy, D. (2008). Modeling imatinib-treated chronic myelogenous leukemia: reducing the complexity of agent-based models. *Bulletin of Mathematical Biology, 70*(3), 728–744.

Kurt, F. (2017). *Analyse der Kapazität des Arbeitsgedächtnisses mittels mathematischer Optimierung* (Master thesis at Ruprecht-Karls Universität Heidelberg Fakultät für Mathematik und Informatik). Heidelberg University.

Litman, J., Hutchins, T., & Russon, R. (2005). Epistemic curiosity, feeling-of-knowing, and exploratory behaviour. *Cognition & Emotion, 19*(4), 559–582.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098–2109.

Lovett, M. C., Daily, L. Z., & Reder, L. M. (2000). A source activation theory of working memory: Cross-task prediction of performance in ACT-R. *Cognitive Systems Research, 1*(2), 99–118.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347–356.

Macal, C., & North, M. (2011). Introductory tutorial: Agent-based modeling and simulation. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, & M. Fu (Eds.), *Proceedings of the 2011 Winter Simulation Conference* (pp. 1456–1469). Piscataway: IEEE Press.

Malmendier, U., Tate, G., & Yan, J. (2010). *Managerial beliefs and corporate financial policies*. Cambridge, Massachusetts: National Bureau of Economic Research.

Marlon, J., Howe, P., Mildenberger, M., & Leiserowitz, A. (2016). Yale Climate Opinion Maps-US 2016. *Yale Program on Climate Change Communication. Available at* http://climatecommunication.yale.edu/visualizations-data/ycom-us-2016/?est=happening&type=value&geo=county. *Accessed March 14, 2019.*.

Marsella, S. C., Pynadath, D. V., & Read, S. J. (2004). Psychsim: Agent-based modeling of social interactions and influence. In M. C. Lovett, C. D. Schunn, C. Lebiere, & M. P (Eds.), *Proceedings of the Sixth International Conference on Cognitive Modeling* (Vol. 36, pp. 243–248). New York: Psychology Press.

Mäs, M., & Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PloS One, 8*(11), e74516.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*(2), 159–188.

McCrae, R. R. (1996). Social consequences of experiential openness. *Psychological Bulletin, 120*(3), 323–337.

McCrae, R. R., & Costa Jr, P. T. (1997). Conceptions and correlates of openness to experience. In *Handbook of personality psychology* (pp. 825–847). Amsterdam: Elsevier.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting

rules. *Journal of Experimental Psychology: General*, *143*(2), 668–693.

McKelvey, R. D., & Page, T. (1990). Public and private information: An experimental study of information pooling. *Econometrica: Journal of the Econometric Society*, 1321–1339.

Merkle, C., & Weber, M. (2011). True overconfidence: The inability of rational information processing to account for apparent overconfidence. *Organizational Behavior and Human Decision Processes*, *116*(2), 262–271.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, *11*(8), e12331.

Moutafi, J., Furnham, A., & Crump, J. (2006). What facets of openness and conscientiousness predict fluid intelligence score? *Learning and Individual Differences*, *16*(1), 31–42.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The Quarterly Journal of Experimental Psychology*, *29*(1), 85–95.

Newell, B. R., McDonald, R. I., Brewer, M., & Hayes, B. K. (2014). The psychology of environmental decisions. *Annual Review of Environment and Resources*, *39*, 443–467.

Ngampruetikorn, V., & Stephens, G. J. (2016). Bias, belief, and consensus: Collective opinion formation on fluctuating networks. *Physical Review E*, *94*(5), 052312.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Oberdiek, H. (2001). *Tolerance: Between forbearance and acceptance*. Lanham: Rowman & Littlefield.

Oremland, M., & Laubenbacher, R. (2014). Optimization of agent-based models: scaling methods and heuristic algorithms. *Journal of Artificial Societies and Social Simulation*, *17*(2), 6.

Ortoleva, P., & Snowberg, E. (2015). Overconfidence in political behavior. *American Economic Review*, *105*(2), 504–35.

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*(4), 1343–1355.

Peterson, C., Seligman, M. E., et al. (2004). *Character strengths and virtues: A handbook and classification* (Vol. 1). Oxford: Oxford University Press.

Pew Research Center. (2014). Political polarization in the american public. *Pew Research Center. Available at https://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/. Accessed March 28, 2019.*.

Pilditch, T. D. (2017). Opinion Cascades and Echo-Chambers in Online Networks: A Proof of Concept Agent-Based Model. In G. G, H. A, T. T, & D. E (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 943–948). Austin, Texas: Cognitive Science Society.

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). Echo chambers on Facebook. *Available at SSRN 2795110*.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of*

*Research in Personality*, *41*(1), 203–212.

R. Anderson, J. (1983, 06). A Spreading Activation Theory of Memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261-295. doi: 10.1016/S0022-5371(83)90201-3

Rassin, E. (2008). Individual differences in the susceptibility to confirmation bias. *Netherlands Journal of Psychology*, *64*(2), 87–93.

Russo, J. E., Schoemaker, P. J., & Russo, E. J. (1989). *Decision traps: Ten barriers to brilliant decision-making and how to overcome them.* New York: Doubleday/Currency.

Rutledge-Taylor, M., Lebiere, C., Thomson, R., Stazewski, J., & Anderson, J. R. (2012). A Comparison of Rule-Based versus Exemplar-Based Categorization Using the ACT-R Architecture. In *Proceedings of the 21st Conference on Behavior Representation in Modeling and Simulation* (pp. 44–50). New York: BRiMS Committee.

Savion, L. (2009). Clinging to discredited beliefs: The larger cognitive story. *Journal of the Scholarship of Teaching and Learning*, *9*(1), 81–92.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, *1*(2), 143–186.

Schneider, S. H. (2004). Abrupt non-linear climate change, irreversibility and surprise. *Global Environmental Change*, *14*(3), 245–258.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*(1), 1–20.

Simon, M., & Houghton, S. M. (2003). The relationship between overconfidence and the introduction of risky products: Evidence from a field study. *Academy of Management Journal*, *46*(2), 139–149.

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, *11*(1), 87–104.

Sobkowicz, P. (2017). Opinion dynamics model based on cognitive biases. *arXiv preprint arXiv:1703.01501*.

Sunstein, C. R. (2018). *# Republic: Divided democracy in the age of social media.* Princeton, New Jersey: Princeton University Press.

Swim, J. K., Stern, P. C., Doherty, T. J., Clayton, S., Reser, J. P., Weber, E. U., ... Howard, G. S. (2011). Psychology's contributions to understanding and addressing global climate change. *American Psychologist*, *66*(4), 241–250.

Thiele, J. C., Kurth, W., & Grimm, V. (2014). Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R. *Journal of Artificial Societies and Social Simulation*, *17*(3), 11.

Tobler, C., Visschers, V. H., & Siegrist, M. (2012). Addressing climate change: Determinants of consumers' willingness to act and to support policy measures. *Journal of Environmental Psychology*, *32*(3), 197–207.

Tsai, C. I., Klayman, J., & Hastie, R. (2008). Effects of amount of information on judgment accuracy and confidence. *Organizational Behavior and Human Decision Processes*, *107*(2), 97–105.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009). There's more to the working memory capacity-fluid intelligence relationship than just secondary memory. *Psychonomic*

*Bulletin & Review*, *16*(5), 931–937.

Webster, J. G. (2005). Beneath the veneer of fragmentation: Television audience polarization in a multichannel world. *Journal of Communication*, *55*(2), 366–382.

Wilson, A. (2014). Bounded memory and biases in information processing. *Econometrica*, *82*(6), 2257–2294.

Zarkov, D. (2017). Populism, polarization and social justice activism. *European Journal of Women's Studies*, *24*(3), 197–201.

Zeng, D., & Sycara, K. (1998). Bayesian learning in negotiation. *International Journal of Human-Computer Studies*, *48*(1), 125–141.

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

**FAKULTÄT FÜR VERHALTENS- UND EMPIRISCHE KULTURWISSENSCHAFTEN**

**Promotionsausschuss der Fakultät für Verhaltens- und Empirische Kulturwissenschaften der Ruprecht-Karls-Universität Heidelberg**
**Doctoral Committee of the Faculty of Behavioural and Cultural Studies of Heidelberg University**

**Erklärung gemäß § 8 (1) c) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften**
**Declaration in accordance to § 8 (1) c) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies**

Ich erkläre, dass ich die vorgelegte Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe.
I declare that I have made the submitted dissertation independently, using only the specified tools and have correctly marked all quotations.

**Erklärung gemäß § 8 (1) d) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften**
**Declaration in accordance to § 8 (1) d) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies**

Ich erkläre, dass ich die vorgelegte Dissertation in dieser oder einer anderen Form nicht anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe.
I declare that I did not use the submitted dissertation in this or any other form as an examination paper until now and that I did not submit it in another faculty.

Vorname Nachname
First name Family name _____

Datum, Unterschrift
Date, Signature _____

# A | Manuscript 1

## Applying Mathematical Optimization Methods to an ACT-R Instance-Based Learning Model

Nadia Said[1,2,*], Michael Engelhart[2], Christian Kirches[2], Stefan Körkel[2], Daniel V. Holt[1],

**1 Institute of Psychology, Heidelberg University, Hauptstr. 47–51, 69117 Heidelberg, Germany**
**2 Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany**

*** nadia.said@psychologie.uni-heidelberg.de**

## Abstract

Computational models of cognition provide an interface to connect to empirically supported theories and models of behavior available in psychology, cognitive science, and neuroscience. In this article, we consider computational models of instance-based learning for dynamic decision making tasks, implemented in the ACT-R cognitive architecture. We propose a framework for obtaining mathematical reformulations of such cognitive models that improves their computational tractability. For the well-known *Sugar Factory* dynamic decision making task, we conduct a simulation study including statistical analysis for central model parameters. We show how mathematical optimization techniques can be applied to efficiently identify optimal parameter values with respect different goals. Beyond these methodological contributions, our analysis reveals the sensitivity of this particular task with respect to initial settings, and yield new insights into how average human performance deviates from potential optimal performance. We conclude by discussing future steps towards applying powerful derivative-based optimization methods to cognitive models represented in ACT-R, an avenue that promises to be an important step towards much improved computational efficiency.

## Introduction

Modern cognitive architectures, such as ACT-R Anderson et al. (2004), allow to construct computational models of behavior that adequately reflect the complexity of human cognition, while still being fully formalized. Cognitive architectures are typically based on empirical behavioral studies and neurophysiological research. Using a model of a cognitive decision process, it becomes possible to answer questions such as “what

can be expected, in the best or worst case, from a decision maker" and "what are the characteristic traits of a typical decision maker".
While cognitive *models* usually focus on particular cognitive phenomena or processes, cognitive *architectures* are concerned with the general structure of the cognitive system across different tasks. Different types of cognitive architectures exist and include symbolic, connectionist, and hybrid architectures, such as Soar (Laird, 2008; Laird, Newell, & Rosenbloom, 1987), Leabra (O'Reilly, Hazy, & Herd, 2012), Nengo (Eliasmith, 2013), and ACT-R (Anderson et al., 2004). The increasing availability and use of formal models of cognition in the behavioral sciences provides an important and growing foundation that calls for application of mathematical tools and methods (Dawson, 2008; Lewandowsky & Farrell, 2010).

**Parameter Identification** The behavior exhibited by a cognitive model typically depends on multiple *model parameters* given the underlying architecture, e.g., the rate of memory decay, or the amount of cognitive noise. Understanding the parameter space of a given cognitive model and efficiently estimating parameter values that best match an expected or measured behavior is a central task in cognitive modeling. This task is made difficult by the large number of function evaluations required, and by the necessary computational complexity of relevant models. Exploring the effects of different parameter values in a cognitive model is important to fully understand its behavior, to identify parameter combinations providing the best fit to human data, and to analyze sensitivity towards parameter variations (Roberts & Pashler, 2000). In practice, for cognitive models this exploration is still often conducted manually, guided by a researcher's intuition, or sometimes just by trial-and-error.

Developing techniques for efficient parameter space exploration and parameter estimation is still a relatively new research area in cognitive modeling, and only few systematic approaches have been described in the literature to date, e.g. (Best et al., 2009; Gluck, Scheutz, Gunzelmann, Harris, & Kershner, 2007; Kase, Ritter, & Schoelles, 2008; Lane & Gobet, 2013; Moore, 2011). Systematic exploration of a given parameter space is often desirable, but quickly runs into difficulties, as processing time increases exponentially with the number of parameters and the resolution of analysis (*curse of dimensionality*). While parallel high-performance computing can improve the speed of parameter space searches to some extent, the combinatorial explosion inherent in this task easily exceeds the capacity even of large computing resources (Gluck et al., 2007).

Another possibility is to try and improve the efficiency of search algorithms. One approach is, for example, to sample the search space selectively by Adaptive Mesh Refinement or Regression Trees (Best et al., 2009; Moore, 2011). Regions of the search space with high-information content, e.g., areas containing discontinuities or nonlinear gradients, are sampled more densely. This strategy allows to preserve most of the information relevant for modeling purposes, while reducing the number of samples required.

Instead of attempting to approximate the full parameter space, it is sometimes sufficient to identify particular points or areas with certain characteristics, e.g., parameter combinations that provide the best model fit to empirical data. To reach this goal, *heuristic optimization methods* such as genetic algorithms have been employed, which use an evolutionary generate-and-select-strategy to find optimal parameter combinations (Kase et al., 2008; Lane & Gobet, 2013). These existing heuristic approaches, however, not only require drastically higher computational resources with increasing

number of dimensions, but also usually do not come with a proof of optimality of the obtained terminal parameter estimate. Using *mathematical optimization methods*, these issues may be avoided by taking information found in (approximations of) first order derivatives of model and objective function into account.

**Contribution** This article proposes an optimization-based approach for evaluating the behavior of a cognitive model represented in the ACT-R modeling language. We propose to rewrite the model in terms of mathematically tractable expressions, and to apply methods from mathematical programming in order to identify parameter values that are optimal with respect to a prescribed criterion. Our approach is generic in the sense that it may be applied to any ACT-R model based on declarative working memory, and may in principle be automated. Extensions to a much wider class of ACT-R models are possible.

To illustrate our approach, we work with a model of the *Sugar Factory* dynamic decision making task, implemented in the ACT-R architecture (Taatgen & Wallach, 2002). We first conduct a simulation study including a statistical analysis for central model parameters. We then show how to address two common optimization problems: Firstly, the identification of parameter values that result in a best model fit to human reference values, and, secondly, the determination of parameter values that maximize the score of a scenario. We propose to apply optimization-based methods that, given an initial guess, construct descent paths to a minimizer instead of searching the entire parameter space, and thereby improve the computational efficiency considerably.

Beyond these methodological contributions, our analysis allows to quantify the sensitivity of the task with respect to initial conditions, and yields new insights into how average human performance deviates from potential optimal performance.

# Mathematical Optimization

Our aim is the application of mathematical optimization methods to models of cognitive processes. We strive to validate models by calibrating them to observed data in order to obtain reliable simulations with predictive capabilites, and to optimize the process behavior. To this end, we formulate mathematical optimization problems and choose appropriate mathematical optimization methods to solve them efficiently.

**Optimization Targets** Our dynamic decision making task setting is round-based, where we denote rounds by $j = 1,\dots,N_r$. For a given parameter vector $\theta$, the model behavior may also depend on a pseudo-random sequence of inputs. Then, evaluations take place over repetitions $i = 1,\dots,n$ with differing realizations of the pseudo-random input sequence. We consider two optimization tasks with respect to the model of the cognitive process:

1. **Parameter estimation.** We determine a parameter vector $\theta \in \mathbb{R}^{n_\theta}$ that gives rise to best model fit to human reference values. For optimizing the model fit, the objective function is the root mean square deviation (RMSD) of the model performance and a human reference value $R_{\text{ref}}$,

$$\min_{\theta} \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R^i(\theta) - R_{\text{ref}}\right)^2}, \tag{A.1}$$

where $R^i(\theta) = \sum_{j=1}^{N_r} R^i_{j+1}(\theta)$. Herein, $R^i_{j+1}(\theta)$ denotes a zero-one indicator that the process was *on target*, i.e. a certain prescribed goal was reached, in repetition $i$ after round $j$ and for model parameters $\theta$.

2. **Process optimization.** We determine a parameter vector $\theta \in \mathbb{R}^{n_\theta}$ with best score. The objective function for the best score is a weighted sum consisting of the performance criterion, here the mean of the rounds on target, and its standard deviation,

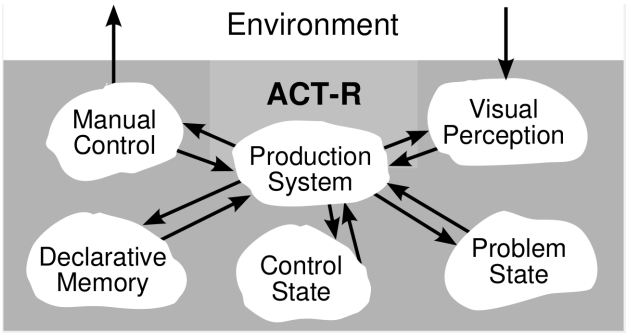
$$\max_{\theta} \quad a \cdot \frac{1}{n} \sum_{i=1}^{n} R^i(\theta) + b \cdot \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left( R^i(\theta) - \frac{1}{n} \sum_{i=1}^{n} R^i(\theta) \right)^2}.$$

Constants $a, b \in \mathbb{R}$ are weighting factors.

## Mathematical Reformulation of ACT-R

The cognitive architecture this article builds on is ACT-R, a computational framework for modeling higher level cognition. ACT-R consists of three main components: modules, buffers, and a pattern matcher (Anderson et al., 2004), which are associated with distinct cortical regions. A central production system coordinates the behavior of these modules, see Fig. A.1. In several functional magnetic resonance imaging (fMRI) studies, Anderson (2007) identified a number of brain regions corresponding to modules in ACT-R, supporting the structure of the architecture.

**Figure A.1.** Connection of modules in ACT-R 5.0 (Anderson, 2005).




One important feature of ACT-R is that it combines the symbolic structure of cognition, i.e., how knowledge is encoded as high-level structures, with a subsymbolic level "[...] abstract characterization of the role of neural computation in making that knowledge available," (Anderson, 2007). As an example, instances of symbolic declarative knowledge (e.g., "The number 2 is followed by the number 3."), called chunks, are stored in the declarative memory. On the subsymbolic level, an activation value is associated with each chunk and determines whether the information is accessible in a certain situation (e.g., when counting). In contrast to purely connectionist models, in which a specific cognitive phenomenon emerges from interconnected networks of simple units (Marcus, 2003), ACT-R operates on different levels of abstraction in order to achieve a representation of how the components of the mind interact.

## Mathematical Description of the Declarative Memory Module

The proposed reformulation of the *Sugar Factory* includes a generic representation of a central part of the ACT-R cognitive architecture, the *declarative memory module*. Our approach can therefore be applied in a straightforward manner to other cognitive tasks that rely on this cognitive module.

A single element of declarative knowledge is called a *chunk*, stored in the *declarative memory* module of the ACT-R architecture. A chunk, see Fig. A.2, is defined by its chunk type and contains a number of *slots* $c_{ik}$ that hold information. Each chunk also has an *activation value* $A_i$ that reflects the usefulness of the stored information for the specific situation at hand (Anderson, 2007).

**Definition 1 (Chunk and Declarative Memory)** *A chunk is a tuple* $(c_{i1}, \ldots, c_{ik}, A_i) \in I_1 \times \ldots \times I_z \times \mathbb{R}$. *The declarative memory is an ordered list* $\mathcal{M}$ *of chunk tuples indexed by consecutive ascending natural numbers.*

The current context and all past experience influence the activation value $A_i$ of a chunk $i$, which is computed from three components: the base-level activation $B_i$, a context component $C_i$, and a noise component $u_{ij}^{\mathrm{n}}$,

$$A_i := B_i + C_i + u_{ij}^{\mathrm{n}}. \tag{A.2}$$

The base-level activation $B_i$ is calculated from the number $n_i$ of presentations, the lifetime $L_i$, and the decay parameter $d$,

$$B_i := \ln\left(n_i/(1-d)\right) - d\ln\left(L_i\right). \tag{A.3}$$

A chunk is said to have been *presented* if a) it first enters the declarative memory, $n_i = 1$, and b) if upon entering declarative memory it is merged with another chunk that already exists. With each presentation of a chunk, the base-level activation $B_i$ increases (Anderson, 2007). The lifetime $L_i$ (i.e., the time since its creation) of a chunk depends on the modeled task. In case of our *Sugar Factory* implementation $L_i$ consists of the round $t_i$ of a chunks creation, the current round $j$, and a time constant $T = 0.05\,\mathrm{s}$,

$$L_i := \left(j - t_i\right) + T. \tag{A.4}$$

When faced with the current situation in turn $j$, a *retrieval request* is be made to retrieve a chunk from declarative memory that best matches the current situation. Then, from the subset of chunks that are a satisfactory match of request $(p_j, p^*)$, comprised of a *situation* $p_j$ and a *desired target* $p^*$, the one with the highest activation value is placed into the *retrieval buffer*.

**Definition 2 (Retrieval of a Chunk)** *Given a request* $(p_j, p^*)$*, the index of the chunk retrieved from declarative memory is*

$$i^* = \underset{i}{\operatorname{argmax}}\left\{A_i(p_j, p^*) \geq \tau\right\}. \tag{A.5}$$

*The retrieval threshold* $\tau$ *defines the minimum activation threshold for a chunk to be retrievable at all. The retrieved chunk is*

$$(c_1^*, \ldots, c_k^*, A^*) = \left(c_{i1^*}, \ldots, c_{ik^*}, A_{i^*}(p_j, p^*)\right). \tag{A.6}$$

To this end, the context component $C_i(p_j, p^*)$ contributes a similarity part that reflects the similarity between the slot values $(p_j, p^*)$ of a retrieval request and the slot values $(c_{i1}, \ldots, c_{ik})$ of any chunks in declarative memory. It is not required that the slots of the chunk have exactly the same values as specified in the retrieval request, but $C_i$ increases if their similarity is high. This mechanism is called *partial matching*,

$$C_i(p_j, p^*) := P \cdot \sum_l M_{i,l}, \tag{A.7}$$

wherein the parameter $P$ reflects the amount of weighting given to the similarities, and the similarity measures $M_{i,l}$ are calculated as

$$M_{i,l}(a, b) := -|a - b| / \max(a, b). \tag{A.8}$$

Maximum similarity between two values is represented by $M_{i,l} := 0$ and maximum dissimilarity by $M_{i,l} := -1$.

Finally, the noise value $u_{ij}^{\mathrm{n}}$ added to the activation consists of two subcomponents: a transient component $u_{ij}^{\mathrm{n}}$, which is computed each time a retrieval request is made, and a permanent component, which is only generated once for each chunk. The transient component is usually sufficient for modeling. To generate the value of the transient noise component a logistic distribution with $\mu = 0$ and noise parameter $s \approx 0.2$ is used (Chung & Byrne, 2008).

**Figure A.2.** Examples of an ACT-R production rule (left) and of ACT-R chunks stored in declarative memory (right).

```
(p start
  =goal>
    isa game-state
    state start
    sugar production =p
  ==>
  =goal>
    state retrieving
    +retrieval>
    isa learned-info
    sugar production =p
    newproduction 9
)
```

```
goal
isa game-state
state start
workers 3
sugar production 7
```

```
learned-info12
isa learned-info
workers 8
production 7
newproduction 9
```

## Mathematical Description of the Production Rules

The single modules of ACT-R interact with each other through a production system. The steps our model runs through are described below. In every round, we

1. compute the activations of the chunks;
2. select the chunk with the highest activation regarding a specific request;
3. if there is such a chunk and the activation of this chunk is above the threshold $\tau$: exhibit specific behavior $b^1$;
4. if there is no such chunk or the activation of the chunk is lower than the threshold $\tau$: exhibit specific behavior $b^2$;

5. update the system state;
6. create a new chunk or merge with an existing chunk that holds relevant information.

Both a particular cognitive process as well as the general production rules for simulating a cognitive process are described in ACT-R by a system of logical relations. In contrast, we aim to formulate a mathematical model of the cognitive process that is a suitable input to mathematical optimization methods.

In our approach, the logical phrases from the ACT-R formalism are modeled by argmax, $|\cdot|$, max, and conditional *if-then* statements. We propose formulations for all three components based on the Heaviside and Delta functions $H(x)$ and $\delta(x)$:

$$H(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}, \quad \delta(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{if } x \neq 0. \end{cases} \tag{A.9}$$

**Formulation of *if-then* statements.** We write *if-then* statements

$$x(s) = \begin{cases} a, & \text{if} \quad s \geq 0, \\ b, & \text{if} \quad s < 0, \end{cases} \qquad y(t) = \begin{cases} c & \text{if} \quad t = 0, \\ d & \text{if} \quad t \neq 0. \end{cases}$$

as $x(s) = H(s) \cdot a + (1 - H(s)) \cdot b$ and $y(t) = \delta(t) \cdot c + (1 - \delta(t)) \cdot d$.

**Formulation of** max **and** $|\cdot|$**.** We substitute max and $|\cdot|$ by

$$\max(x, y) = H(x - y) \cdot x + (1 - H(x - y)) \cdot y,$$
$$\frac{|x - y|}{\max(x, y)} = H(x - y)\frac{x - y}{x} + (1 - H(x - y))\frac{y - x}{y}.$$

**Formulation of** argmax**.** To evaluate the statement

$$i^* = \underset{1 \leq i \leq n}{\operatorname{argmax}}\{A_i\}, \quad x_j(i^*) = \begin{cases} b^1 & \text{if } A_{i^*} \geq \tau, \\ b^2 & \text{if } A_{i^*} < \tau, \end{cases}$$

we first compute $A^* = \max\limits_i \{A_i\}$, and then let

$$x_j(i^*) = \sum_{i=1}^{n} H(A_i - A^*) \cdot (H(A^* - \tau) \cdot b^1 + (1 - (A^* - \tau)) \cdot b^2).$$

In order to obtain a continuously differentiable formulation, we then replace Heaviside and Delta functions by continuous approximations,

$$H(x) := \tfrac{1}{\pi} \arctan\big(hx\big) + \tfrac{1}{2},$$
$$\delta(x) := \exp(-x^2 / a^2),$$

with, e.g., $h = 10.0$, $a = 0.01$.

## The Sugar Factory Problem

In this article, we investigate an ACT-R model of the *Sugar Factory* decision task (Taatgen & Wallach, 2002). The *Sugar Factory* is a turn-based task realized as a computer-based simulation, developed by Berry and Broadbent (1984) in order to answer the question of how participants learn to operate complex systems. Instead of learning and applying explicit generalized rules to a problem, specific situation-responses are stored in memory and retrieved when a similar situation appears, see (Dienes & Fahey, 1995; Taatgen & Wallach, 2002). This cognitive mechanism is known as *instance-based learning* (IBL), cf. (Logan, 1988), and has been shown to play an important role in dynamic decision making situations in Gonzalez and Lebiere (2005). IBL has been implemented successfully in several cognitive models based on the ACT-R architecture as reported in (Gonzalez & Lebiere, 2005; Taatgen & Wallach, 2002).

In the *Sugar Factory* task, participants observe a sugar production rate $p_j$ over turns $j = 1, \ldots, N_r$, and are asked to reach and maintain a specific sugar production $p^*$ by repeatedly changing the number of workers $x_j$ employed at the factory. The initial value is $p_1 = 6$. In every round $j$, the goal is to reach $p_j = p^* = 9$, i.e., to produce 9,000 metric tons of sugar. The following equation describes the behavior of the *Sugar Factory* task.

**Definition 3 (Sugar Factory Simulation Problem)** *The sugar production rate before turn $j = 1$ is $p_1$ and the rate $p_{j+1}$ after turn $j = 1, \ldots, N_r$ is given by*

$$p_{j+1}(x) = \left(2 \cdot x_j - p_j(x) + u_j^{\mathrm{r}}\right)_{[1,12]}, \tag{A.10}$$

*where $x_j \in \{1, \ldots, 12\}$ is a sequence of inputs, $u_{r,j}$ is uniformly distributed random variable from $\{-1, 0, 1\}$, and $(y)_{[a,b]} = \max(a, \min(b, y))$ denotes the clipping of the argument value $y$ to the range $[a, b]$.*

Participants are initially unaware of the relationship (A.10) between workers and sugar production, and are not informed about their results being evaluated in this way.

To measure the performance of a participant on the Sugar Factory, we define the following score.

**Definition 4 (Sugar Factory Score Function)** *The sugar factory score function is*

$$R = \sum_{j=1}^{N_r} R_{j+1} = \sum_{j=1}^{N_r} \chi\left\{|p_{j+1}(x) - p^*| \leq 1\right\}$$

*with $p^* = 9$, i.e., the score counts the number of rounds where the sugar production rate is on target.*

To account for the randomness in $u_j^{\mathrm{r}}$ and to make it possible for participants to be on target 100% of the time, a sugar production of $p_j \in [8, 10]$ is scored as being on target.

## Human Performance in the Sugar Factory

It has repeatedly been found that human participants are able to control the simulated system above chance level but perform far from the rational optimum in this task (Berry & Broadbent, 1984; Dienes & Fahey, 1995). Moreover, even successful participants are

often unable to verbally describe the structure of the system. This is in line with the assumptions of instance-based learning as a cognitive mechanism which do not require the abstraction of formal rules. Surprisingly, even when the structure of the underlying system is made explicit to participants, they are generally not able to improve their performance (Berry & Broadbent, 1984).

Analyzing individual decision behavior, Dienes and Fahey (1995) found that up to 86% of the initial ten choices $x_1, \ldots, x_{10}$ made by participants can be explained by the following rules, which also form the basis for the cognitive model further below:

- Initially, a workforce of $x_1 = 7$, 8, or 9 is chosen;
- If the sugar production is below or above target, $p_j < 8$ or $p_j > 10$, then $x_j = x_{j-1} + u_j^{\text{off}}$, where $u_j^{\text{off}} \in \{-2, \ldots, 2\}$ is added to the current workforce;
- If the sugar production on target, $8 \leq p_j \leq 10$, then $u_j^{\text{on}} \in \{-1, \ldots, 1\}$ is added to the current workforce.

**Algorithm 3:** Mathematical formulation of the ACT-R model of the *Sugar Factory*.

**for** $j = 1, \ldots, N_r$ **do**

**(1)** **for** $i = 1, \ldots, N_c$ **do**

$$L_i := \left(j - t_i\right) + T;$$
$$B_i := \ln\left(n_i/(1-d)\right) - d \cdot \ln\left(L_i\right);$$
$$M_{i1} := -\left|p_j - c_{i2}\right| / \max\left(p_j, c_{i2}\right);$$
$$M_{i2} := -|9 - c_{i3}| / \max\left(9, c_{i3}\right);$$
$$A_i := B_i + P \cdot (M_{i1} + M_{i2}) + u_{ij}^{\mathrm{n}};$$

**end**

**(2)** $i^* := \operatorname{argmax}_i \{A_i\};$

**(3)** $A_{i^*} \geq \tau$?

(i) if $A_{i^*} \geq \tau$ then $x_j := c_{i^*1}$;

(ii) else $x_j := u_{w,j}$;

**(4)** $p_{j+1} := 2 \cdot x_j - p_j + u_j^{\mathrm{r}}$;

(i) if $p_{j+1} > 12$ then $p_{j+1} = 12$;

(ii) if $p_{j+1} < 1$ then $p_{j+1} = 1$;

(iii) $p_{j+1} = 9$?

(a) if $p_{j+1} = 9$ then $u_{w,j+1} := u_{w,j} + u_j^{\mathrm{on}}$;

(b) else $u_{w,j+1} := u_{w,j} + u_j^{\mathrm{off}}$;

**(5)** if $u_{w,j+1} > 12$ then $u_{w,j+1} = 12$;

**(6)** if $u_{w,j+1} < 1$ then $u_{w,j+1} = 1$;

**(7)** $p_{j+1} \in \{8, ..., 10\}$?

(i) if $p_{j+1} \in \{8, ..., 10\}$ then $R_{j+1} := 1$;

(ii) else $R_{j+1} := 0$;

**(8)** $\exists\, i = 1, \ldots, N_c : c_i = \left(x_j, p_j, p_{j+1}\right)$?

(i) if $\exists\, i$ then $n_i := n_i + 1$

(ii) else

$$N_c := N_c + 1;$$
$$c_{N_c} := \left(x_j, p_j, p_{j+1}\right);$$
$$n_{N_c} := 1;$$
$$t_{N_c} := j;$$

**end**

As an example and to demonstrate the mathematical description of the production rules, the limits on the sugar production rates in the *Sugar Factory* are implemented by *if-then* statements. These rules appear as follows in our mathematical description:

$$\text{if } p_{j+1} > 12 \text{ then } p_{j+1} = 12, \quad \text{if } p_{j+1} < 1 \text{ then } p_{j+1} = 1.$$

In our reformulation, these *if-then* statements are smoothened using the Heaviside

function $H$:

$$\begin{aligned}\tilde{p}_{j+1} &= 2 \cdot x_{j+1} - p_j + u_{rj},\\ p_{j+1} &= H(\tilde{p}_{j+1} - 12) \cdot 12 + \big(1 - H(\tilde{p}_{j+1} - 12)\big)\\ &\quad \cdot \Big(H(1 - \tilde{p}_{j+1}) \cdot 1 + \big(1 - H(1 - \tilde{p}_{j+1})\big) \cdot \tilde{p}_{j+1}\Big).\end{aligned}$$

### Nonlinear Recurrence Model

For the *Sugar Factory* problem, let $N_r$ be the number of rounds. Each chunk $i$ has three slots ($c_{i1}$, $c_{i2}$, $c_{i3}$), where $c_{i1}$ holds the information about the new workforce, the value $c_{i2}$ represents the current production and $c_{i3}$ is the new production calculated from $c_{i1}$ and $c_{i2}$. The maximum number $N_c$ of chunks can be calculated from the number of values $c_{ik}$ possible for slot $k \in \{1,2,3\}$ of chunk $i$. Feasible values for new workforce $c_{i1}$, current production $c_{i2}$, and for new production $c_{i3}$ are $\{1,\ldots,12\}$ each. Thus, $N_c = 12 \cdot 12 \cdot 12 = 1{,}728$. We allocate every possible chunk and set its initial activity to a sufficiently small negative value $-M$ to make sure that it is possible to activate it only after information has been stored in the slots of the chunk.

The mathematical model contains different types of variables:

- *states* including the activation $A_i$ of the chunks, the current number of workers $x_j$, and the current sugar production rate $p_j$ in the *Sugar Factory*,
- *parameters* $\theta = (\tau, d, P)$ and $s$ describing the cognitive properties of the individual participant, and
- *pseudo-random vectors*, containing the cognitive noise $u^{\mathrm{n}}$, random decisions by the participants $u_w + u_j^{\mathrm{on}}$ resp. $u_w + u_j^{\mathrm{off}}$ and system inputs $u^{\mathrm{r}}$. They describe the particular settings under which the cognitive task is run. The sequences of random values are generated a priori as reproducible pseudo-random numbers.

All inputs are vectors of length $N_r$, except $u_{ij}^{\mathrm{n}}$, which is of length $N_r \cdot N_c$. The value $R_{j+1}$ is used as an indicator whether the participant has reached the target in round $j$, i.e., whether the new sugar production $p_{j+1}$ equals 8, 9, or 10. The overall score $R^i$ is computed by summation over all $R_{j+1}$,

$$R^i = \sum_{j=1}^{N_r} R_{j+1}^i.$$

with $R_{j+1}^i$ as the indicator *on target* in round $j = 1,\ldots,N_r$ for input $i = 1,\ldots,n$. This modeling approach leads to a systems of nonlinear recurrence relations as shown in Algorithm 2.

## Properties of the Model and Choice of Optimization Methods

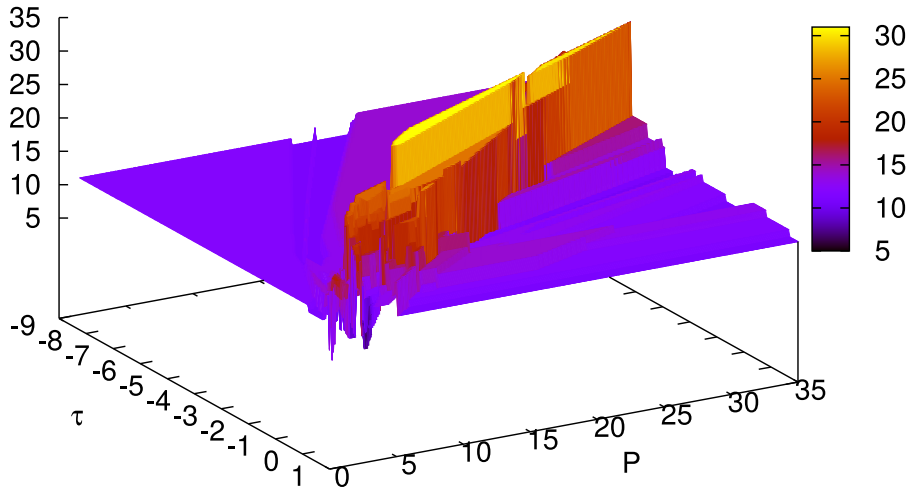
We have implemented the mathematical reformulation of the *Sugar Factory* model in *Python*. Our implementation is modular and object-oriented, with model specific components encapsulated in a problem class that is easy to substitute in order to transfer it to similar tasks. In this section, we report simulation results obtained using this implementation.

## Simulation and Sensitivity Analysis

Computations of the simulations were run on a non-dedicated 48-core machine ($4 \times 12$-core *AMD Opteron 6176*) with 256 GB RAM. Depending on the grid resolution (i.e., step sizes of the parameters variations) the maximum runtime did not exceed one day. We focused on an analysis of the parameters $P$ and $\tau$, which have considerable effect on the achieved score while no strong empirically based recommendation for default values exists. The decay rate was set to its default value $d = 0.5$. The activation noise $u^{\mathrm{n}}_{ij}$ was set to zero as it does not lead to a noticeable change of the mean score. We describe the random components $u^{\mathrm{on}}$, $u^{\mathrm{off}}$ and $u^{\mathrm{r}}$ by pseudo-random input sequences. Different simulation runs are characterized by different pseudo-random input sequences.

Fig. A.3 shows simulation results for one fixed input (i.e., for every parameter combination $(\tau, P)$ the same sequences $u^{\mathrm{on}}$, $u^{\mathrm{off}}$, and $u^{r}$ were used). Parameter ranges are $\tau \in [-8.50, 1.00]$ with step size $\Delta\tau = 0.05$, and $P \in [0.5, 35.0]$ with step size $\Delta P = 0.1$, which results in a total of 66,086 sampling points. There are certain parameter combinations for which the model is on target about 87.5 % of the time ($\tau \in [-3.3, -0.85]$ and $P \in [8.4, 23.7]$) and others where the score drops to less than 25 %. The structure of the plots, especially in the area of the *best learners*, strongly depends on the inputs, compare Fig. A.3 and Fig. A.4. In the latter, the *best learners* are on target no more than 50% of the time, the score drops to 10% near the edge.

**Figure A.3.** Rounds on target for input$_0$ (= $u^{\mathrm{on}}_0$, $u^{\mathrm{off}}_0$, $u^{\mathrm{r}}_0$) over 40 rounds on fine parameter grid with 66,086 grid points.




Hence, we conducted a second simulation in which the input sequences were varied pseudo-randomly. Fig. A.5 (left) shows the mean of 100 different samples for the pseudo-random sequences. Not only does the total number of rounds on target differ compared to the single inputs, but also the area of parameter combinations that yield good results is much broader. To check whether or not the declarative memory has truly "learned" an implicit strategy, only instances in which a chunk was actually retrieved were counted as being on target in Fig. A.5 (center) — again for the same 100 samples for the pseudo-random sequences — and compared to our previous results. Compared to Fig. A.5 (left), there is a drop of the score in the upper right quarter of Fig. A.5 (center). The standard deviation of the mean value for the activated chunks version is shown in Fig. A.5 (right).

**Figure A.4.** Rounds on target for input$_1$ (= $u_1^{\text{on}}$, $u_1^{\text{off}}$, $u_1^{\text{r}}$) over 40 rounds on fine parameter grid with 66,086 grid points.

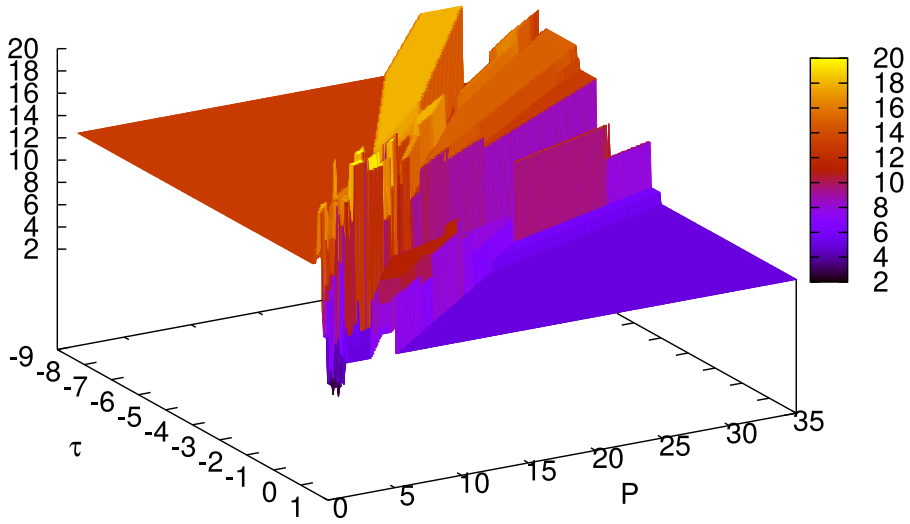
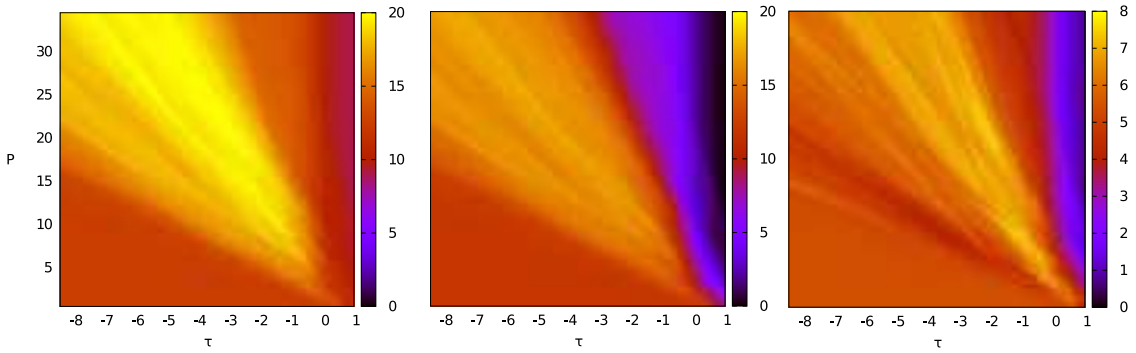



**Figure A.5.** Mean value and standard deviation of rounds on target for 100 inputs over 40 rounds. Initial sugar production rate $p_1 = 6$, medium parameter grid with 8,256 grid points. Left: Mean value of rounds on target. Center: Mean value, activated chunks only. Right: Standard dev., activated chunks only.




In a further simulation, we investigated the sensitivity of the scores with respect to different initial sugar production values. The default value used in experiments is $p_1 = 6$. Results for an initial value of $p_1 = 9$ show, compared to the default value $p_1 = 6$, not only a much broader region of *best solvers* but also a higher overall score. On the other hand, an initial value of $p_1 = 1$ yields a lower overall score as well as a smaller region of *best solvers*.

In general, all simulation results show a similar pattern in response to parameter variations. Fig. A.5 (left) shows a characteristic interaction of parameters $\tau$ and $P$, with the *highest scoring* parameter combinations located in a wedge-shaped area at the center of the plot and lower scores in both lower left and upper right corners. Considering whether model responses were based on memory retrieval as opposed to random behavior, Fig. A.5 (center) reveals that learning occurs in the lower left corner. In contrast, in the upper right corner behavior is almost exclusively driven by random behavior.

## Choice of Optimization Methods

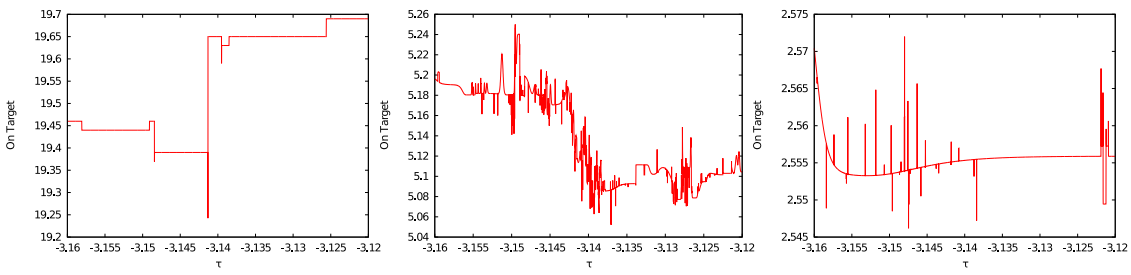
In this section we discuss the results of our simulation study regarding the numerical differentiability of our reformulated model as well as the appropriate optimization methods. In order to apply derivative-based optimization methods, a tractable formulation is necessary. However, the ACT-R architecture contains many logical statements (*if-then-else*) and absolute value expressions that challenge a derivative-based approach. As shown before, such non-differentiabilities can be smoothed using continuous approximations of the Heaviside and Delta functions (A.9). This is similar to the approach of Gurney, Prescott, and Redgrave (2001), in which the authors also used a smooth approximation of the Heaviside function in order to model action selection.

A different approach is described by Chapelle and Wu (2010), who used a softmax activation function and showed that it is possible to perform gradient-based optimization on their smooth approximation. This approach however requires, inter alia, that the index selected by argmax is unique. How to deal with chunks having (almost) the same activation remains an open question.

**Choice of the Smoothing Parameter $h$** We concentrated on the influence of the parameter $h$ of the Heaviside function, as the parameter $a$ of the Delta function turned out to be uncritical. Larger values of $h$ correspond to a sharper transitions at $x = 0$. To identify the value of $h$ for which our model becomes numerically differentiable, we ran simulations for $h \in \{0.1, 1.0, 10, 10^2, \ldots, 10^7\}$ with $P = 20$, and focused on the particular parameter interval $\tau \in [-3.16, -3.12]$ sampled at step size $\Delta\tau = 10^{-5}$. We also separately varied $h$ for smoothing of the similarity calculation, denoted by $h_{\text{sim}}$, smoothing of argmax, denoted by $h_{\text{argmax}}$, and for computation of sugar production and workers, denoted by $h_{\text{env}}$.

Results in Fig. A.6 show that, the *argmax* term proves to be critical for matching the behavior of the *Sugar Factory* model for the ACT-R and the Python implementation. Larger values of $h$ (Fig. A.6, left) are required but yield a numerically non-differentiable function. Decreasing the values of $h_{\text{argmax}}$ leads i.a. to a random choice of chunks that undermines the learning process. Fig. A.6 shows that the score drops from about 19.5 (left) to approximately 5.2 (center). For the similarity calculation and the calculation of sugar production and workers, the choice of $h$ is less critical, but $h_{\text{argmax}} = 100$ is still too large to yield a numerically differentiable model (right).

**Figure A.6.** Mean value of rounds on target for 100 inputs over 40 rounds with $P$ = const. Left: $h_{\text{argmax}} = 10^7$, $h_{\text{sim}} = h_{\text{env}} = 10$. Center: $h_{\text{argmax}} = 10^3$, $h_{\text{sim}} = h_{\text{env}} = 10$. Right: $h_{env} = 10$, $h_{\text{argmax}} = h_{\text{sim}} = 100$. Note the different vertical axis ranges.




We may conclude that, even though smoothing the argmax can be a feasible approach, cf. (Gurney et al., 2001), (Chapelle & Wu, 2010), precise modeling of the argmax is crucial at least for the particular case of the *Sugar Factory* model.

**Heuristic and Mathematical Optimization Approaches** Optimization methods such as genetic algorithms (Mitchell, 1998) or particle swarm (Trelea, 2003) search the global parameter space based on *heuristics*. Such algorithms however rely on the computational time, for example, as termination criterion, as they have little information on whether or not they have actually found an optimum. Two examples for such heuristic optimization methods are ESCH (Beyer & Schwefel, 2002), a modified evolutionary algorithm, and Controlled Random Search (CRS) with local mutation (Kaelo & Ali, 2006). CRS starts with a population of random points, and evolves them heuristically, a method comparable to genetic algorithms.

On the other hand *mathematical optimization solvers* are characterized by the use of derivatives as a source of additional information to make sure that an optimum is reached. Those *mathematical* optimization solvers are e.g. Newton-type algorithms, e.g. (Gill & Murray, 1974), or steepest descent methods, e.g. (Sun & Yuan, 2006), but also include derivative-free methods such as Nelder-Mead (Lagarias, Reeds, Wright, & Wright, n.d.) or BOBYQA (Powell, 2009), which approximate the behavior of gradient based solvers. Nelder-Mead is a downhill simplex method while BOBYQA uses an iteratively constructed quadratic approximation for the objective function. Whereas *heuristic* optimization methods are quickly stretched to their limits with an increasing dimensionality of the parameter space, the number of iterations for mathematical optimization methods, in particular for derivative based ones, ideally is independent of the problem dimensions.

## Numerical Results for the Sugar Factory

We applied a selection of heuristic and mathematical optimization algorithms that promise to cope with the non-differentiability of the nonlinear recurrence model. Our selection comprises Nelder-Mead Simplex (Nelder & Mead, 1965) with explicit support for bound constraints (Box, 1965), BOBYQA, ESCH, and CRS. All optimization algorithms were applied using the Python interface NLopt (Johnson, 2010).

The stopping criterion for BOBYQA and Nelder-Mead was a relative tolerance on the optimization parameters of 0.1. For the heuristic global solvers ESCH and CRS we successively increased the time limit up to about 1000 s. The stopping criterion was then set to the minimum run time for which there was no improvement of the found maxima observed.

**Table A.1.** Maxima found by different solvers for $n = 100$ inputs. Objective was to find the parameter combination best fitting a human reference value using RMSD (parameter estimation).

| **Solver** | $\tau$ | $P$ | **Max.** | **#Eval.** |
|---|---|---|---|---|
| Nelder-Mead | 0.5 | 28.13 | 4.05 | 67 |
| BOBYQA | 0.5 | 27.80 | 4.05 | 54 |
| ESCH | 0.45 | 27.88 | 4.05 | 6,374 |
| CRS | 0.48 | 32.94 | 4.05 | 4,500 |

**Parameter estimation** Table A.1 shows the results for the best fit to human reference performance, with $R_{\text{ref}} = 7.9$ taken from the literature (Dienes & Fahey, 1995). Using multiple start values, all solvers found the same point as a maximum. For ESCH and CRS the results displayed are for a time limit of 5.0 seconds.

**Table A.2.** Maxima found by different solvers for $n = 100$ inputs. Objective was to find the parameter combination with the best score (process optimization).

| **Solver** | $\tau$ | $P$ | **Max.** | **#Eval.** |
|---|---|---|---|---|
| Nelder-Mead | -4.00 | 27.00 | 20.15 | 36 |
| BOBYQA | -4.00 | 27.00 | 20.15 | 43 |
| ESCH | -3.13 | 22.36 | 20.13 | 863 |
| CRS | -4.21 | 28.52 | 20.2 | 860 |

**Process optimization** For the single input displayed in Fig. A.3, all solvers found the global maximal score of 31, however depending on suitable and differing choices of the initial values for parameters $\tau$ and $P$. Table A.2 shows the results for $a = 1$ and $b = 0$ and 100 inputs using multiple start values (see Fig. A.5, left). The local solvers Nelder-Mead and BOBYQA both found the same local maximum ($\tau = -4.00$, $P = 27.00$ with objective = 20.15). Table A.2 shows the maxima found by the heuristic global solvers after 960 seconds (see Fig. A.7). For $a = 1$ and $b = -1$, all solvers found the same point as a maximum ($\tau \approx -6.5$, $P \approx 30$ with objective $\approx 13.87$), except CRS which found a slightly better point ($\tau \approx -8.15$, $P \approx 34.9$ with objective $\approx 14.04$).

**Figure A.7.** Mean of 100 inputs over 40 rounds, $p_0 = 6$, medium grid (8,256 grid points). Points 1–4 show 1: best score found by ESCH, 2: best score found by local solvers, 3: best score found by CRS, 4: best fit to reference human.

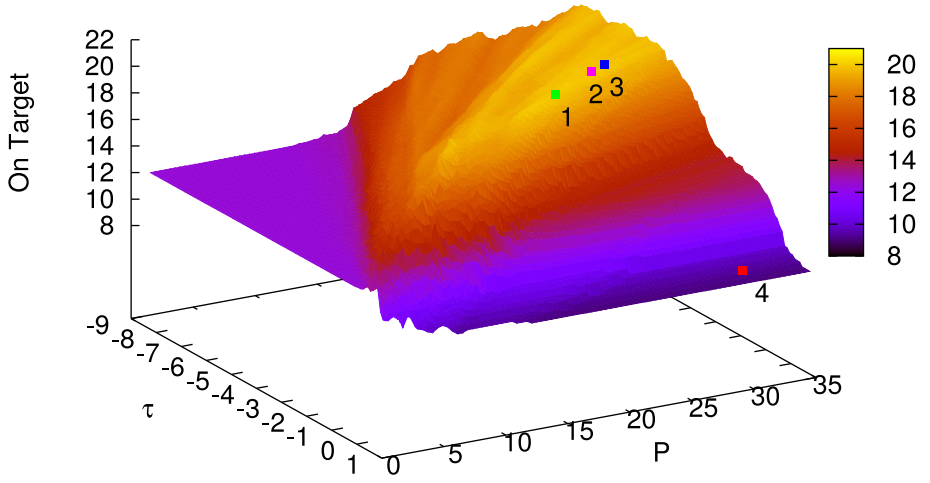


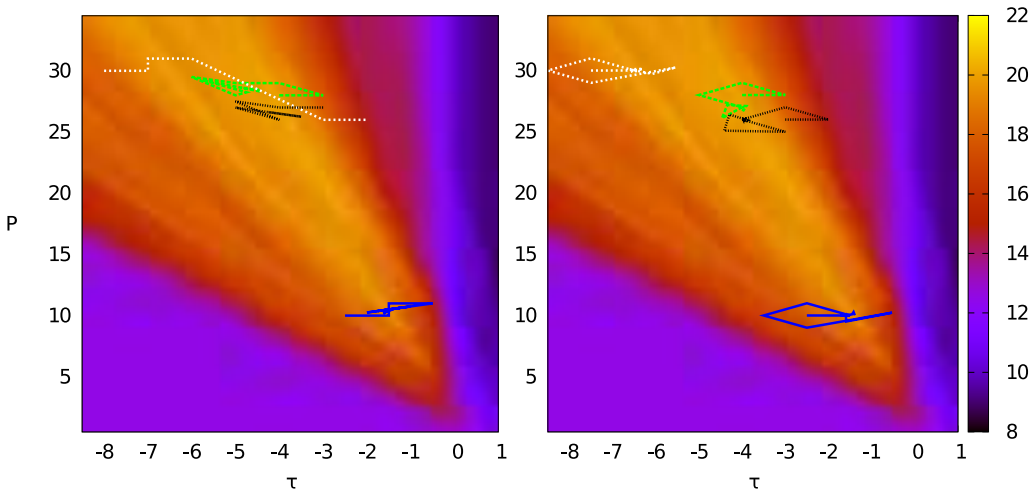

Fig. A.8 shows the optimization trajectories for Nelder-Mead and BOBYQA.

## Cognitive Interpretation

Our results show how high-performance simulation and optimization can provide relevant insights beyond just quickly optimizing the fit to aggregate human data. Interestingly, optimizing for ideal performance shows that this combination is far away from the possible optimal performance, located at $\tau = -4.21$ and $P = 28.52$ (see Figure A.7). This raises two questions, namely why the $\tau$ value of the model with the best fit to human performance diverges from the optimal model, and how a lower $\tau$ value leads to better performance.

A simple answer to the first question is that people do not behave formally optimal in many decision situations (Klein, 2001) in general, and in the *Sugar Factory* in particular

**Figure A.8.** Optimization trajectories for Nelder-Mead (left) and BOBYQA (right) with four different start values.




(Berry & Broadbent, 1984; Dienes & Fahey, 1995). The structure of the human cognitive system seems to be geared towards robust information processing in typical human environments with incomplete or uncertain information (Gigerenzer & Gaissmaier, 2011), rather than formal optimization given strict assumptions. Another possibility is that the model of human performance used in this study is not valid to start with. However, given that existing studies of the *Sugar Factory* task and its derivatives show that implicit learning is a strong explanatory mechanism (Berry & Broadbent, 1984; Dienes & Fahey, 1995), that the implementation of implicit learning based on the ACT-R architecture generally matches human data in other studies (Gonzalez, Lerch, & Lebiere, 2003), and that the specific model used here has been empirically supported (Taatgen & Wallach, 2002), we think there are good grounds to assume that the basic structure of the model is appropriate.

The second question is how a lower $\tau$ value leads to better performance. Apparently, being more open to considering vague memories (i.e., a low retrieval threshold $\tau$) is mostly a good strategy in this task. We think that this is a task-specific effect, as participants are provided only with noise-free and correct information. This means that any memory, however vague, that is is sufficiently similar to the target situation is on average a valuable source of information and likely to increase performance. Similarity is important though, as the estimate for parameter $P$ (mismatch penalty) shows, which lies close to the theoretical optimum. The more conservative memory threshold (low $\tau$ value), shown by most human participants may represent a suitable trade-off across a range of different situations, given that information often is noisy or unreliable and a higher selectivity therefore advisable. This is supported by the fact that the $\tau$ value of 0.5 we found is close to the value of 0 recommended by the architecture as a suitable default for many situations (ACT-R 6.0 Tutorial, 2012).

We also investigated how choosing different initial values for the sugar production has an influence on performance. That an initial value of $p = 9$ yields the best overall performance is not surprising, as this part of the optimal value range of the workforce (7, 8 or 9) and therefore produces many memory instances of trials on target early in the learning phase, which are important for guiding control behavior. This insight is practically relevant for behavioral studies, as the sensitivity to starting conditions has so far not been considered in studies using the *Sugar Factory*.

# Discussion

Cognitive architectures are powerful tools for modeling higher-level cognition. However, methods for efficient exploration of parameter spaces and quick parameter estimation in this type of models are currently still in the process of development. In this article, we have demonstrated first steps towards a differentiable formulation of an instance-based learning model implemented in the ACT-R cognitive architecture. We conducted a simulation study including statistical analysis for central model parameters and showed how mathematical optimization techniques can be applied for efficient parameter identification.

We implemented a mathematical reformulation of the *Sugar Factory*, a simple instance-based learning task, in *Python* and showed that it is possible to derive a tractable formulation. The generic part of this formulation, related to the ACT-R declarative memory module, can in principle be transferred to ACT-R models of other tasks like Backgammon (Sanner, Anderson, Lebiere, & Lovett, 2000), Air-Traffic Control (Lebiere, Anderson, & Bothell, 2001), the beer game (Martin, Gonzalez, & Lebiere, 2004), or even more complex tasks like the Water Purification Plant (Gonzalez et al., 2003).
We conducted simulation studies to determine model properties by varying the parameter $h$ of an approximation of the Heaviside function, which we used for smoothing the non-differentiable parts of our model.

Simulations showed that in order to obtain exactly the same results like the ACT-R model a large $h$ for the smoothened *argmax* is necessary, contrary to other parts of our model like the similarity calculation and the environmental setting (i.e. calculation of sugar production and workers). This however, leads to a piecewise constant behavior of our *Python* implementation. For smaller $h$ our model becomes numercially differentiable, however at the same time the learning process is replaced by random behavior. Therefore, at this stage, even though the derivatives can be calculated, using gradient-based optimization methods is not feasible.
We then showed how to address two common optimization problems: Firstly, the identification of parameter values that result in a best model fit to human reference values, and, secondly, the determination of parameter values that maximize the score of a scenario. We applied both heuristic and mathematical optimization algorithms that promise to cope with the non-differentiability of our nonlinear recurrence model and showed that *mathematical optimization* solvers like Nelder-Mead Simplex or BOBYQA turned out to be the best choice for the model at hand. Not only do they have the advantage of using approximations of the derivatives to determine if an extremum is found, thus needing a lower number of iterations than the *heuristic optimization* solvers, but they are also, in principle, able to deal with higher dimensional problems. Therefore it should be also possible to apply those to parameter spaces with $n > 2$. Furthermore, we conducted a simulation study for the central model parameters: the retrieval threshold $\tau$ and the *partial matching* parameter $P$ using high-performance computing. Results revealed a sensitivity of the task to initial settings, which has not been considered in the empirical literature so far. These results also indicate that human performance in this specific tasks seems to be hampered in part by a tendency to be overly conservative in considering memory instances.

## Outlook

As the *argmax* turned out to be the crucial part of the transscribed ACT-R model, we pursued a non-differentiable approach in this article and developed a nonlinear recurrence relation that could be optimized with a selection of heuristic or derivative-free solvers. This approach has the advantage of allowing for the computation of a single round of the cognitive process by a mere function evaluation.

We envision in a next step to derive exact reformulations of IBL problems and ACT-R cognitive processes that are amenable to derivative-based optimization methods, as follows: Returning once more to the statement $i^* = \arg\max\{A_i\}$ for data $A_1, \ldots, A_k$, consider the following constrained optimization problem:

$$\begin{cases} \min\limits_{A^*, w} & A^* & & & & \\ \text{s.t.} & A^* & \geq & A_i, & & 1 \leq i \leq k, \\ & w_i \cdot (A_i - A^*) & \geq & 0, & & 1 \leq i \leq k, \\ & w_i & \in & [0,1], & \sum_{i=1}^k w_i = 1. & \end{cases}$$

Herein, $A^*$ is a free variable set to the maximum activation value by virtue of minimization and the first inequality. We seek the argmax, i.e. the index $i^*$ with $A_{i^*} = A^*$. All differences in the second inequality are non-positive, and all with $A_i < A^*$ are negative. This forces the corresponding indicators $w_i$ to zero. Then, the indicator $w_{i^*}$ is forced to one by the equality in the third line. A function $f(i^*)$ depending on $i^*$, the argmax, may then be expressed as

$$f(i^*) = \sum_{i=1}^{k} w_i f(i),$$

which is now bi-linear, differentiable, and independent of the argmax, but yields the same value because $w_i = 0$ for $i \neq i^*$, and $w_{i^*} = 1$.

This formulation represents the computation of one sample of the dynamic decision making process by the solution of a bi-linear optimization problem. The approach is hence significantly more demanding in terms of computational effort. Moreover, optimizing over process samples computed in this way constitutes a bi-level optimization problem. Treatment of such problems is significantly more demanding also in terms of mathematical optimization algorithms, but has the advantage of precisely reproducing the sequence of chunk activations as determined by ACT-R.

Another possibility that might increase the tractability of our model is a different representation of the production rules, as in (Stewart & Eliasmith, 2008). Instead of using a two-step approach like in ACT-R, production rules only have one feature, their utilities.

**Acknowledgments** The authors gratefully acknowledge support by the Excellence Initiative of the German Research Council (DFG) and by DFG Graduate School 220 (Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences).

# References


ACT-R 6.0 Tutorial. (2012). Retrieved from the ACT-R Web site:http://act-r.psy.cmu.edu/software/.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*(3), 313–341.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology*, *36*(2), 209–231.

Best, B. J., Furjanic, C., Gerhart, N., Fincham, J., Gluck, K. A., Gunzelmann, G., & Krusmark, M. A. (2009). Adaptive mesh refinement for efficient exploration of cognitive architectures and cognitive models. *Proc. Ninth Int. Conf. Cognitive Modeling*, 1-6.

Beyer, H.-G., & Schwefel, H.-P. (2002). Evolution strategies–a comprehensive introduction. *Natural Computing*, *1*(1), 3–52.

Box, M. J. (1965). A new method of constrained optimization and a comparison with other methods. *The Computer Journal*, *8*(1), 42–52.

Chapelle, O., & Wu, M. (2010). Gradient descent optimization of smoothed information retrieval metrics. *Information Retrieval*, *13*(3), 216–235.

Chung, P. H., & Byrne, M. D. (2008). Cue effectiveness in mitigating postcompletion errors in a routine procedural task. *International Journal of Human-Computer Studies*, *66*(4), 217–232.

Dawson, M. R. W. (2008). *Minds and machines: Connectionism and psychological modeling*. New York City: John Wiley & Sons.

Dienes, Z., & Fahey, R. (1995). Role of specific instances in controlling a dynamic system. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 848–862.

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford: Oxford University Press.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gill, P. E., & Murray, W. (1974). Newtontype methods for unconstrained and linearly constrained optimization. *Mathematical Programming*, *7*(1), 311–350.

Gluck, K., Scheutz, M., Gunzelmann, G., Harris, J., & Kershner, J. (2007). Combinatorics meets processing power: Large-scale computational resources for BRIMS. In T. Kelley & L. Allender (Eds.), *Proceedings of the sixteenth conference on behavior representation in modeling and simulation* (pp. 73–83). Orlando: Simulation

Interoperability Standards Organization.

Gonzalez, C., & Lebiere, C. (2005). Instance-based cognitive models of decision-making. In D. J. Zizzo & A. Courakis (Eds.), *Transfer of knowledge in economic decision making*. London: Palgrave McMillan.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. II. analysis and simulation of behaviour. *Biological Cybernetics*, *84*(6), 411–423.

Johnson, S. G. (2010). *The NLopt nonlinear-optimization package*. Available at http://ab-initio.mit.edu/nlop.

Kaelo, P., & Ali, M. M. (2006). Some variants of the controlled random search algorithm for global optimization. *JOTA*, *130*(2), 253–264.

Kase, S. E., Ritter, F. E., & Schoelles, M. (2008). From modeler-free individual data fitting to 3-d parametric prediction landscapes: A research expedition. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1398–1403). Austin, Texas: Cognitive Science Society.

Klein, G. (2001). The fiction of optimization. *Bounded rationality: The adaptive toolbox*, 103–121.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (n.d.). Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on Optimization*, *9*(1), 112–147.

Laird, J. E. (2008). Extending the Soar cognitive architecture. *Frontiers in Artificial Intelligence and Applications*, *171*, 224–236.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*(1), 1–64.

Lane, P. C. R., & Gobet, F. (2013). Evolving non-dominated parameter sets. *Journal of Artificial General Intelligence*, *4*(1), 358–367.

Lebiere, C., Anderson, J. R., & Bothell, D. (2001). Multi-tasking and cognitive workload in an ACT-r model of a simplified air traffic control task. *Proceedings of the Tenth Conference on Computer Generated Forces and Behavior Representation*.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. Thousand Oaks: Sage.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527.

Marcus, G. F. (2003). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, Massachusetts: MIT Press.

Martin, M. K., Gonzalez, C., & Lebiere, C. (2004). Learning to make decisions in dynamic environments: ACT-r plays the beer game. *Proceedings of the Sixth International Conference on Cognitive Modeling*.

Mitchell, M. (1998). *An introduction to genetic algorithms*. Cambridge, Massachusetts: MIT Press.

Moore, L. R. J. (2011). Cognitive model exploration and optimization: a new challenge for computational science. *Computational and Mathematical Organization Theory*, *17*(3), 296–313.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313.

O'Reilly, R. C., Hazy, T. E., & Herd, S. A. (2012). *The leabra cognitive architecture: how to play 20 principles with nature and win!* Oxford: Oxford University Press.

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives* (Tech. Rep. No. NA2009/06).

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Sanner, S., Anderson, J. R., Lebiere, C., & Lovett, M. C. (2000). Achieving efficient and cognitively plausible learning in backgammon. *Proc. Seventeenth Int. Conf. Machine Learning*, 823–830.

Stewart, T. C., & Eliasmith, C. (2008). Building production systems with realistic spiking neurons. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual meeting of the Cognitive Science Society* (pp. 1636–1641). Austin, Texas: Cognitive Science Society.

Sun, W., & Yuan, Y.-X. (2006). *Optimization Theory and Methods: Nonlinear Programming*. Berlin/Heidelberg: Springer Science & Business Media.

Taatgen, N. A., & Wallach, D. (2002). Whether skill acquisition is rule or instance based is determined by the structure of the task. *Cognitive Science Quarterly*, *2*(2), 163–204.

Trelea, I. C. (2003). The particle swarm optimization algorithm: convergence analysis and parameter selection. *Information Processing Letters*, *85*(6), 317–325.

# B | Manuscript 2

## The accuracy of German citizens' confidence in their climate change knowledge

Helen Fischer[1,*], Dorothee Amelung[1], Nadia Said[1,2],

**1 Institute of Psychology, Heidelberg University, Hauptstr. 47–51, 69117 Heidelberg, Germany**
**2 Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany**

*** helen.fischer@psychologie.uni-heidelberg.de**

## Abstract

Just like knowledge can be more or less accurate as a function of its correspondence with reality, so can confidence be more or less accurate as a function of its correspondence with the accuracy of knowledge. Accurate confidence has proven important for making appropriate predictions and decisions in areas of high uncertainty such as managerial, medical, or political decision-making. Research in the area of climate change (CC), however, has mostly dealt with the accuracy of citizens' CC knowledge itself, while the accuracy of their confidence is unknown. Here we provide a comprehensive assessment of the accuracy of confidence in CC knowledge in a nationally representative German sample ($N = 509$). Participants verified a total of nine true and false statements about CC, and indicated their confidence in each verification. Two benchmark comparisons are provided: The accuracy of confidence in science knowledge in a second nationally representative German sample ($N = 588$), and the accuracy of confidence in CC knowledge in a sample of CC scientists ($N = 206$). The main indicator of confidence accuracy was relative confidence sensitivity ($M_{\text{ratio}}$) that allows us to quantify the lack of knowing what one knows that cannot be explained by a lack of knowledge. Results showed that citizens' $M_{\text{ratio}}$ for their CC knowledge was only 0.49 (95% CI [0.33, 0.63]). Thus, citizens' confidence sensitivity was only around half of what it could be based on their knowledge, which indicates a failure of metacognitive awareness of their true level of CC knowledge. Moreover, relative confidence sensitivity of citizens' CC knowledge was particularly low compared to their science knowledge ($M_{\text{ratio}} = .99$, 95% CI [0.88, 1.16]), as well as scientists' CC knowledge ($M_{\text{ratio}} = .95$, 95% CI [0.85, 1.07]). We conclude that citizens' confidence in their CC knowledge represents a particularly fuzzy line between correct versus incorrect verifications relative to science knowledge, and relative to scientists'

CC knowledge; moreover, it represents an *unnecessarily* fuzzy line relative to citizens' own CC knowledge.

# Introduction

Accurate confidence in knowledge is fundamental to competent decision-making (Bruine de Bruin, Parker, & Fischhoff, 2007; Parker & Fischhoff, 2005). Just like knowledge can be more or less accurate as a function of its correspondence with reality, so can confidence be more or less accurate as a function of its correspondence with the accuracy of knowledge. For example, rejecting the statement that natural variation in sunbeam is the main driver of climate change (CC) shows accurate knowledge, but being uncertain about this rejection shows inaccurate confidence. Accepting the statement that greenhouse gas emissions are a main driver of CC shows accurate knowledge, and being certain about this acceptance also shows accurate confidence. While high confidence in accurate knowledge enables decision-makers to leverage their knowledge, unwarranted confidence in false knowledge can lead people astray. Specifically, accurate confidence has proven important for making appropriate predictions and decisions in areas of high uncertainty such as managerial (Simon & Houghton, 2003), medical (Berner & Graber, 2008), or political decision-making (Johnson, 2009). Research in the area of climate change, however, has almost exclusively dealt with the accuracy of citizens' CC knowledge itself (e.g., Shi, Visschers, Siegrist, & Arvai, 2016), while the accuracy of their confidence is largely unknown. Here we deliver a comprehensive estimate of the accuracy of citizens' confidence in their verifications of a series of true and false statements about CC. We put the results into context with the help of two benchmark comparisons: The first, citizens' accuracy of confidence in science knowledge, enables us to compare across domains, and the second, CC scientists' accuracy of confidence in CC knowledge, enables us to compare across levels of expertise and exposure to scientifically accurate information. Jointly, the present approach delivers relative answers to questions such as: How well calibrated is citizens' confidence in their CC knowledge? To what extent are citizens aware of the limits of their CC knowledge? How does citizens' accuracy of CC knowledge relate to their accuracy of confidence?

To assess the accuracy of citizens' confidence, we employ methods that have a long tradition in assessing the accuracy of subjective probabilities in areas as diverse as sensory judgments (Baranski & Petrusic, 1994), eyewitness identification (Brewer & Wells, 2006), or forecasts in strategic intelligence (Mandel & Barnes, 2014). Accuracy of confidence is determined in terms of *calibration*, *bias*, and *sensitivity*. *Calibration* measures how well (or poorly) confidence in knowledge aligns with the accuracy of knowledge: A person shows optimal calibration when for statements verified with 100% accuracy, confidence is 100%, for statements verified with 90% accuracy, confidence is 90%, and so on. Confidence is *biased* to the extent that it over- or underestimates knowledge. *Sensitivity*, also termed *resolution of confidence*, captures how well confidence judgments discriminate between what one knows and does not know. A person shows optimal sensitivity when confidence is highest for all statements verified correctly, and lowest for all statements verified incorrectly. Hence, if citizens possess any insight into their CC knowledge, their confidence judgments should discriminate between correct and incorrect verifications.

Citizens' accuracy of confidence is relevant for three reasons. First, confidence in knowledge can affect decision-making above and beyond knowledge (Hadar, Sood, &

Fox, 2013). For example, both biased and less sensitive confidence have been identified as causes of diagnostic error. Physicians who are overconfident about the accuracy of their diagnosis tend to prematurely narrow down the choice of diagnostic hypotheses(Berner & Graber, 2008), and are less likely to seek out additional information that would increase their chance to correct their diagnosis (Meyer, Payne, Meeks, Rao, & Singh, 2013), and less sensitive diagnostic confidence predicts less optimal medical decisions (Jackson & Kleitman, 2014). We deliver an estimate of the accuracy of citizens' confidence in their CC knowledge. If citizens' confidence in their CC knowledge is inaccurate, they cannot leverage accurate CC knowledge, and can be lead astray by false knowledge. Second, citizens are often faced with the challenge to verify the claims of CC statements of unknown veracity. This is because scientifically valid CC information(Hiles & Hinnant, 2014) exists alongside considerable misinformation in the media discourse, including outright disinformation campaigns (Elsasser & Dunlap, 2013), while the amount of "nonsense" to be found on the internet on topics such as CC appears to be on the rise (Williamson, 2016). To what degree of confidence, and to what degree of accuracy of confidence do citizens reject or accept valid information and "nonsense" about CC? On the one hand, confidence in the veracity of one's beliefs may become better-calibrated with feedback (Miller & Geraci, 2011). Therefore, we could expect to find confidence in CC knowledge to be well-calibrated since citizens receive sufficient feedback via media to adjust their confidence. On the other hand, polarized media coverage of CC may encourage selective processing of novel information to confirm prior beliefs. Polarized or complex media coverage may facilitate natural tendencies to selectively reject (unpleasant) information (Sanbonmatsu, Posavac, Kardes, & Mantel, 1998), or to underweigh conflicting information (Ortoleva & Snowberg, 2015; Park, Konana, Gu, Kumar, & Raghunathan, 2010), both of which have been shown to foster miscalibration. Here citizens' accuracy of confidence in their CC knowledge is contrasted with two benchmarks: (i) The accuracy of confidence in science knowledge compared to CC knowledge in a second citizens sample, and (ii) the accuracy of confidence in CC knowledge in a sample of climate scientists. Science knowledge, first, represents a revealing comparison because it demonstrates the level of accuracy of confidence that can be achieved by citizens in a similar domain (factual scientific knowledge) that is unrelated to CC. In areas of expertise that produce valid and timely feedback, experts tend to be comparatively well-calibrated (e.g., meteorologists; Murphy & Winkler, 1984). CC scientists, second, therefore represent an informative comparison group because they demonstrate the level of accuracy of confidence in CC knowledge that can be reached through regular exposure to scientifically valid CC information, and who should therefore show comparatively low signs of metacognitive confusion about true and false statements. Third, to what extent can citizens determine what they don't know about CC, to what degree are citizens aware of the limits of their knowledge? Typical CC knowledge tests cannot answer this questions, even though they do offer an "I don't know" option (Shi et al., 2016; Tobler, Visschers, & Siegrist, 2012). Since the "I don't know"-option excludes any further knowledge assessment, we cannot determine what citizens would have known. Also, the binary classification of an "I don't know"-option does not allow for a rich metacognitive assessment. Here we employ a graded assessment of confidence which is logically independent from the knowledge assessment and enables us to estimate the extent to which citizens are aware of the demarcation line between knowledge and guessing. Previous research has shown that knowledge and confidence tend to be positively correlated (Sundblad, Biel, & Gärling,

2009). One might thus be inclined to infer that citizens are able to estimate what and how much they know about CC. However, correlation is only an insufficient measure of confidence accuracy for three reasons: (a) Correlation is independent of the mean and therefore a given correlation is compatible with both over- and underconfidence; (b) perfect calibration is compatible with both high and low correlation (Juslin, Olsson, & Winman, 1996); and, most importantly, (c) a single correlation coefficient cannot deliver a comprehensive picture of the three facets of confidence accuracy: calibration, bias, and sensitivity.

To measure calibration, we use the *C-Index* which captures how well confidence aligns with accuracy, at the item-level, by determining the squared difference between confidence and proportion correct for each confidence level (e.g., 60% confidence), weighted by the number of judgments at each level: $\frac{1}{N}\sum n(r-c)^2$ , where $n$ is the number of judgments for each level, $r$ is the numerical value of confidence, and c is the proportion of correct answers on that level (Bornstein & Zickafoose, 1999). Optimal calibration is 0.

To measure bias, we use the *O-/UIndex* which assesses, for each confidence level, the difference between accuracy and confidence identical to the C-Index, except that differences are not squared so that positive values denote overconfidence, and negative values denote underconfidence: $\frac{1}{N}\sum n(r-c)$.

To measure sensitivity, we use $M_{ratio}$ which defines sensitivity in a Signal Detection Theory (SDT) approach, and which is considered the gold standard of all measures of metacognition (Overgaard & Sandberg, 2012) because it allows to factor out the influence of the primary task performance (CC knowledge). In yes/no tasks such as typical CC knowledge tests, knowledge sensitivity is confounded with participants' yes-saying bias, their propensity to say "yes" irrespective of content. SDT can be applied to remove bias by measuring knowledge sensitivity as $d'$, the difference between the hit rate ("yes" responses to statements that were in fact true) and the false alarm rate ("yes" responses to statements that were false). A straightforward approach to transfer this to confidence sensitivity would be to compute the difference between the high confidence-correct rate and the high confidence-incorrect rate. However, it was shown that such confidence sensitivity is confounded with both confidence and knowledge response bias, and is furthermore confounded with primary task performance, that is, the CC knowledge test (Galvin, Podd, Drga, & Whitmore, 2003; Maniscalco & Lau, 2012). Maniscalco and Lau (2012) therefore propose to compute confidence sensitivity at the level of knowledge distributions by defining $d'$ as that level of confidence sensitivity that one would *expect* to occur given a persons' knowledge sensitivity. The observed confidence sensitivity is then computed finding the $d'$ that maximizes the likelihood of the observed confidence data. Because d' and confidence sensitivity have the unique advantage of being expressed in the same signal-to-noise ratio units, the two can be directly compared, and relative confidence sensitivity ($M_{\text{ratio}}$) can be determined as the level of confidence sensitivity given a certain level of knowledge sensitivity. $M_{\text{ratio}}$ can thus be used to assess people's ability to discriminate between correct and incorrect verifications in their confidence, while controlling for their ability to discriminate true and false statements in their knowledge (Appendix B).

To estimate$M_{\text{ratio}}$ we applied the hierarchical Bayesian estimation approach of Fleming (Fleming, 2017). In comparison to other estimation routines (e.g. maximum likelihood estimation) the hierarchical Bayesian estimation approach has the advantage of (i) providing probability density functions instead of inevitably noisy point

estimates of $M_{\text{ratio}}$; (ii) allowing to accurately estimate $M_{\text{ratio}}$o when confidence rating data/participant is limited; and (iii) providing group-level fits that are less influenced by highly uncertain individual-level fits. The accuracy of citizens' confidence in CC knowledge was estimated in a national sample of $N = 509$ German citizens. The sample was nationally representative in terms of gender, age, and geographical distribution. Participants verified of a total of nine statements about CC (five true, four false), and indicated their confidence in each verification. Statements were taken from previous research (Sundblad et al., 2009), and consisted of three questions from each of the knowledge domains *causes*, *state*, and *consequences* of CC (Table B.1). Each statement was introduced with "Science says that...". Participants answered, "Yes, science says that", or "No, science does not say that", and indicated their confidence in their answer: "How certain are you that your answer is correct?". Confidence was indicated on a 6-point scale ranging from 50%: "not at all certain, I was guessing" to 100%: "certain, I know the answer". To test for potential difficulties with this probabilistic answer format, we also included a frequentist confidence judgment by asking, upon completion of all knowledge questions: "Out of the 9 questions: How many did you answer correctly?" (B). In the science knowledge benchmark study, a national sample of $N = 588$ German citizens, again nationally representative in terms of gender, age, and geographical distribution, verified nine statements (five true, four false) about physical and biological science. The statements were taken from the National Science Board's Science and Engineering Indicators (National Science Board, 2016). In the CC scientists benchmark study, a total of $N = 207$ CC scientists working in research organizations under the German Climate Consortium answered the same CC knowledge questions as citizens (B).

# Results

**Knowledge accuracy**. Out of nine questions each, citizens verified an average of 4.3 CC statements correctly, compared to 5.4 science statements, and compared to 7.2 CC statements verified correctly by CC scientists [1] We assessed the extent to which CC knowledge scores were affected by guessing by comparing two different kinds of scoring: (i) The typically used 1-0-0 scoring that treats guessed responses as incorrect (irrespective of their actual accuracy; Shi et al. (2016); Tobler et al. (2012)), and (ii) number-right scoring that adds up all correct responses (irrespective of whether they were guessed). Guessed responses were those where participants chose the lowest confidence category: 50%, I was guessing". Knowledge accuracy was more strongly affected by guessing for citizens (1-0-0 scoring to number-right scoring: $M = .61$, 95% CI [.60, .63] to $M = .47$, 95% CI [.45, .48]) compared to scientists ($M = .84$, 95% CI [.82, .86] to $M = .80$, 95% CI [.78, .83]), $F(1,714) = 77.2$, $p < .001$. (See also Supplementary Material Figure B for reliability of the true and false statements as indicators of CC knowledge).

Concerning the accuracy of verifications of true compared to false CC statements, the following pattern of results emerged: The accuracy of CC knowledge was lower for citizens ($M = .47$, 95% CI [.45, .48]) compared to scientists ($M = .80$, 95% CI [.78, .83]), $F(1,714) = 410.2$, $p < .001$, and generally lower for FALSE statements ($M = .56$, 95% CI [.53, .57]) compared to TRUE statements ($M = .72$, 95% CI [.70, .74]), $F(1,714) = 95.4$,

[1] Descriptive results on the accuracy of CC knowledge refer to typical 1-0-0 scoring that treats guessed responses as incorrect, as this provides the most reliable assessment of CC knowledge (Appendix B), but the results hold for number-right scoring as well (Appendix B).

$p < .001$. However, for citizens the difference in accuracy between TRUE ($M = .62$, 95% CI [.60, .64]) and FALSE statements ($M = .31$, 95% CI [.30, .34]) was considerably larger than for scientists ($M = .82$, 95% CI [.78, .86] and $M = .79$, 95% CI [.75, .82], for true and false statements, respectively), $F(1,714) = 86.7$, $p < .001$. For scientists, accuracy did not even differ between true and false statements, $t(1,206) = 1.4$, $p = .16$. That is, not only were citizens less accurate in general, but citizens had substantial difficulty in identifying false statements as false, a phenomenon that was practically non-existent in scientists.

**Accuracy of confidence: Calibration.** In line with previous results, citizens appeared well- calibrated overall in that average confidence correlated positively with total accuracy of CC knowledge, both for the probabilistic confidence assessment, $r(508) = .21$, $p < .001$ (scientists: $r(206) = .19$, $p = .005$), and the frequentist confidence assessment, $r(508) = .18$, $p < .001$ (scientists: $r(206) = .29$, $p < .001$). However, the association between total confidence and total accuracy was lower for citizens' CC compared to science knowledge, both for the probabilistic, $r(588) = .46$, $p < .001$, $z = 4.7$, $p < .001$, and the frequentist confidence assessment $r(588) = .48$, $p < .001$, $z = 5.6$, $p < .001$.

Citizens were less confident ($M = .73$, 95% CI [.71, 74]) of their verifications of CC statements compared to their verifications of science statements ($M = .82$, 95% CI [.81, 83]), $F(1,1095) = 170$, $p < .001$, and also compared to scientists ($M = .87$, 95% CI [86, 88]), $F(1,714) = 267$, $p < .001$. Concerning the calibration of confidence, Figure B.1 shows that for TRUE CC statements, citizens' confidence was rather well-calibrated in that mean confidence judgments were roughly in line with accuracy, and the calibration curve followed a linearly increasing trend. For FALSE CC statements, however, a different pattern emerged in that confidence and accuracy were detached. This contrasts citizens' calibration curves of science knowledge, as well as scientists' calibration curves of CC knowledge in that in both cases, the FALSE statements followed optimal calibration more closely.

When comparing C-Indices for CC statements in citizens and scientists in an ANOVA, results showed that while calibrations were generally better for TRUE ($M = .06$, 95% CI [.052, .061]) compared to FALSE statements ($M = .08$, 95% CI [.078, .09]), $F(1,714) = 48.9$, $p < .001$, and scientists ($M = .05$, 95% CI [.041, 0.53]) were better-calibrated compared to citizens ($M = .09$, 95% CI [.90, .98]) across true and false statements, $F(1,714) = 167.06$, $p < .001$, the calibration disadvantage for FALSE statements was particularly large for citizens, $F(1,714) = 41.04$, $p < .001$.

**Accuracy of confidence: Bias.** For the true CC statements, citizens' bias was rather low in absolute terms ($M = .02, SD = .15$), and in fact not statistically different from scientists' ($M = .01, SD = .10$), 95% CI of the difference [−.02, .02], $t(714) = .55$, $p = .58$. For the false CC statements, in contrast, citizens were mostly overconfident in absolute terms ($M = .11, SD = .14$), and also more biased than scientists ($M = .03, SD = .09$), 95% CI of the difference [.06, .09], $t(714) = 7.4$, $p < .001$. Interestingly, as Figure B.1 shows, citizens were under-confident for three CC judgments, namely true statements evaluated with 60% and 70% confidence (95% CI of accuracy [61, 74] and [73, 82]), and false statements verified with 50% confidence, which were in fact verified with above-chance accuracy (95% CI of accuracy [57, 67]). For the science knowledge statements, a rather typical calibration pattern appeared in that citizens were overconfident across nearly the entire confidence spectrum, except for statements verified with very low confidence.

**Table B.1.** Proportion correct and confidence judgments for the true/false statements.

| Knowledge domain | Statement (True/False) | Percentage correct verifications, citizens (scientists) | Confidence judgment, citizens (scientists) |
|---|---|---|---|
| **State** | The global average temperature in the air has increased approx. 3.1 °C in the past 100 years. (False) | 32.6% (82%) | M=.71, SD=.16 (M=.88, SD=.15) |
| | The 1990s was the warmest decade during the past 100 years. (False) | 54.4% (65%) | M=.67, SD=.17 (M=.85 SD=.14) |
| | The global change in temperature in the past 100 years is the largest during the past 1000 years. (True) | 62.1% (88%) | M=.71, SD=.17 (M=.89 SD=.12) |
| **Causes** | Climate change is mainly caused by a natural variation in sunbeam and volcanic eruption. (False) | 80% (97%) | M=.74, SD=.18 (M=.96 SD=.14) |
| | Carbon dioxide concentration in the atmosphere has increased more than 30% during the past 250 years. (True) | 70.5% (94%) | M=.69, SD=.17 (M=.92 SD=.12) |
| | The increase of greenhouse gases is mainly caused by human activities. (True) | 84.1 %(99%) | M=.78, SD=.16 (M=.98 SD=.05) |
| **Consequences** | The blanket of snow in the Northern hemisphere has decreased approximately 10% since the 1960. (True) | 77% (83%) | M=.72, SD=.17 (M=.72 SD=.16) |
| | An increasing amount of greenhouse gases increases the risk of more UV-radiation and therefore a larger risk of skin cancer. (False) | 24.2% (81%) | M=.75, SD=.18 (M=.83 SD=.16) |
| | In 100 years from now, sea level will rise approximately one meter. (True) | 80.5% (69%) | M=.75, SD=.17 (M=.82 SD=.14) |

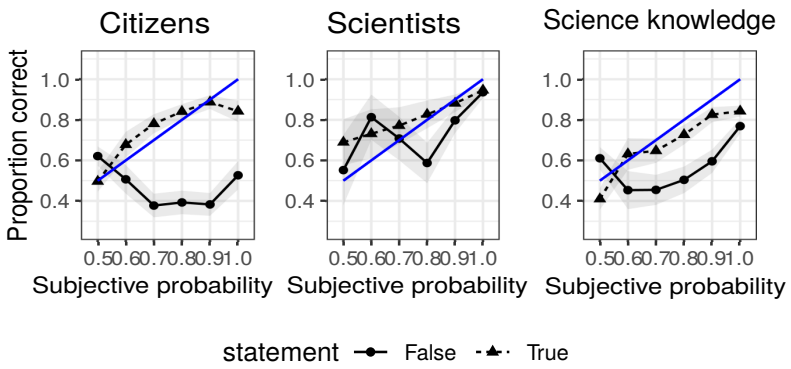



**Figure B.1.** Calibration curves for confidence judgments, separately for separately for climate change (CC) knowledge of citizens and scientists, and for science knowledge of citizens. The figure displays mean proportion of correct verifications on the y-axis against mean confidence for each confidence level on the x-axis, separately for TRUE and FALSE statements. Solid blue line denotes optimal calibration. Shaded grey area: 95% confidence band.

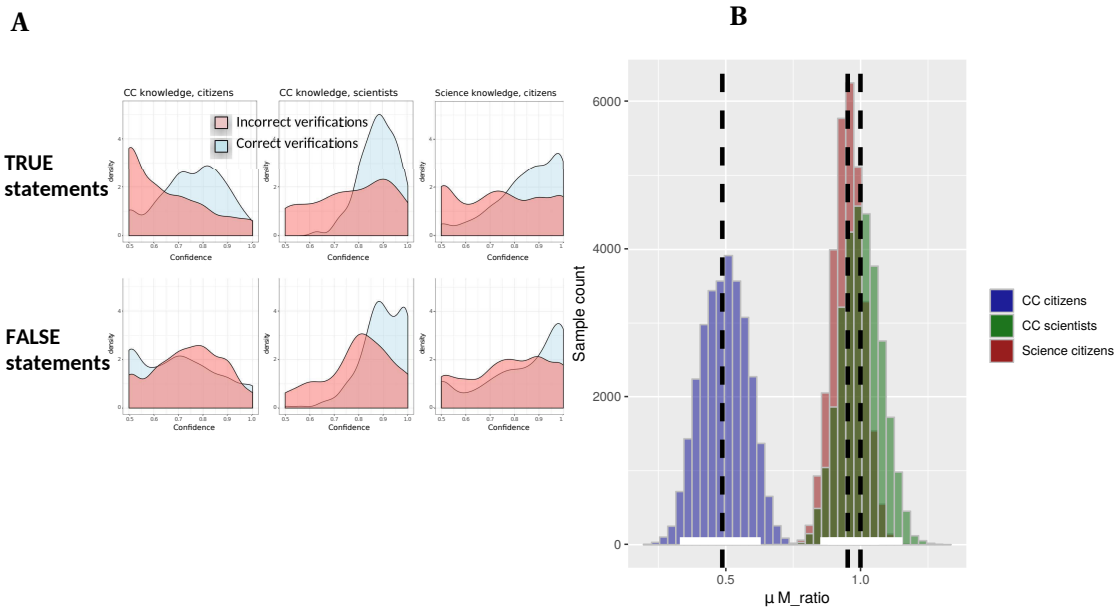



**Figure B.2.** Sensitivity of confidence judgments, separately for climate change (CC) knowledge of citizens and scientists, and for science knowledge of citizens. Panel A displays participants' confidence in correct (blue) and incorrect (red) verifications, separately for TRUE statements and FALSE statements. Panel B displays the aggregated samples of Markov chain Monte Carlo chains (y-axis) for relative confidence sensitivity, $M_{\text{ratio}}$. Vertical lines denote means. White bars: 95% CIs.

**Accuracy of confidence: Sensitivity.** Figure B.2 Panel A displays the distributions for citizens' CC knowledge are relatively flat compared to citizens' science knowledge, and scientists' CC knowledge. The SDT approach produced converging results: Citizens' relative confidence sensitivity of CC knowledge was $M_{\text{ratio}}$ = .49, 95% CI [0.33, 0.63]. Thus, citizens' relative confidence sensitivity was only around 50% of what it could be based on their knowledge sensitivity. For science knowledge, citizens achieved a close-to-optimal $M_{\text{ratio}}$ of $M$ = .99, 95% CI [0.88, 1.16], and scientists achieved an $M_{\text{ratio}}$ of $M$ = .95, 95% CI [0.85, 1.07] for CC knowledge. Taken together, (i) citizens' confidence judgments discriminated considerably less between correct and incorrect verifications than what could be expected based on their CC knowledge. Specifically, their confidence sensitivity was only around half of their knowledge sensitivity; and (ii) estimates of relative confidence sensitivity for citizens' science knowledge, and scientists' CC knowledge were overlapping with optimal sensitivity of 1 (Fig. B.2, Panel B).

# Discussion

The present study provides a comprehensive assessment of the accuracy of citizens' confidence in their CC knowledge. Accuracy of confidence is of particular importance in the area of CC where scientifically correct information exists alongside substantial misinformation in the public discourse and media (Boussalis & Coan, 2016; Hiles & Hinnant, 2014; Lewandowsky, Oberauer, & Gignac, 2013). We employed two different types of indicators of accuracy of confidence: absolute (C-/OU-Index) and relative ($M_{\text{ratio}}$). Absolute indicators determine how much we can trust citizens' confidence judgments as indicators of CC knowledge, which also depends on the accuracy of their CC knowledge itself; relative indicators factor out the influence of knowledge by determining the accuracy of confidence relative to a given level of knowledge. Relative confidence sensitivity can hence be optimal even if knowledge sensitivity is non-optimal, namely when individuals know the limits of their knowledge. As two benchmark comparisons, accuracy of confidence in CC knowledge was estimated for citizens' science knowledge, and for CC scientists.

Concerning absolute confidence accuracy, citizens were remarkably well-calibrated when verifying true statements as indicated by a calibration curve that roughly followed a linearly increasing trend close to optimal calibration. This result suggests that for true statements, citizens could reliably indicate their varying degree of knowledge, and that citizens ' confidence tended to be justified by their accuracy. Such high absolute confidence accuracy shows that German citizens' confidence judgments are informative for accuracy, and can in general be trusted. But for the false statements, citizens appeared to have no insight into their not knowing as indicated by a calibration curve that was practically detached from their accuracy. Even for statements that citizens evaluated with 100% confidence, accuracy was only at guessing rate, compared to approx. 70% for science knowledge, and over 90% for scientists.

Relative confidence accuracy factors out the influence of knowledge accuracy, and thus allows us to quantify the lack of accuracy in confidence in citizens *that cannot be explained by a lack of climate change knowledge*. With a relative confidence sensitivity of $M_{\text{ratio}}$ = .49, citizens' confidence sensitivity was only around 50% of their knowledge sensitivity. Put differently, citizens' confidence sensitivity was around half of what it could be based on the accuracy of their CC knowledge. Citizens' confidence judgments

thus not only reflected a blurry line, but an *unnecessarily* blurry line between correct and incorrect verifications.

Moreover, citizens' relative confidence of their CC knowledge was lower compared to scientists' ($M_{\text{ratio}} = .95$), and lower compared to their science knowledge ($M_{\text{ratio}} = .99$), both of which were close to being cognitively ideal. Cognitively ideal values of $M_{\text{ratio}}$ suggest that participants could use all the information available for the knowledge task when estimating their confidence (Fleming, 2017).

Concerning the question of whether citizens could reliably indicate when they are guessing about the veracity of CC statements, citizens responded with an average accuracy clearly above guessing rate (95% CI [57, 67]) for what they perceived as "guessed" responses to false statements. It is considered a strong indication of unconscious knowledge when people perform above chance levels, but claim to be guessing (Dienes & Fahey, 1995). This result indicates that citizens possess partial knowledge allowing them to recognize false claims as false that they are either not consciously aware of, or of which they consciously underestimate the accuracy. For the true statements, in contrast, citizens were well-calibrated in the lowest confidence category. That is, when citizens answered they were "just guessing", in these cases, typically they really were.

Even though at least for true statements, guessed responses in fact did score at chance levels, the reverse is not true: Just because solution rates were at chance level for some statements does not mean that responses were blindly guessed. Rather, our results show that false statements can be accepted as true with high confidence by citizens. This may reflect what has been termed the "truthiness" of a statement, the feeling or intuition that a statement is true, irrespective of evidence. False "nonsense" statements accepted with high confidence included the ones that confuse CC and UV radiation, or assume an average warming of $3.1°C$ over the past 100 years, even though there is wide-spread discussion on limiting global warming to 2.0, or even 1.5 degrees.

It is furthermore interesting to see for which statements citizens tended to *under*estimate their accuracy. Statements with a positive item-based bias (accuracy-confidence) that was significantly different from zero were "The blanket of snow in the Northern hemisphere has decreased approximately 10% since the 1960s (true)"; "In 100 years from now sea level will rise approximately one meter (true)"; "The increase of greenhouse gases is mainly caused by human activities (true)"; and "Climate change is mainly caused by a natural variation in sunbeam and volcanic eruption (false)". For these statements, citizens on average had a higher level of accurate knowledge than reflected in their confidence. Thus, citizens appeared more doubtful of their CC knowledge than warranted, indicating distrusted knowledge. Distrusted knowledge can have detrimental consequences, since only if one has sufficient confidence in knowledge to actually use it, one truly possesses that knowledge (Burton & Miller, 1999).

Important implications follow from the present results in the light of research demonstrating that accuracy of confidence affects decision-making, above and beyond the accuracy of knowledge (Jackson & Kleitman, 2014). This implies that citizens' blurry confidence judgments do not allow them to fully leverage their accurate knowledge, nor that they can protect them from relying on false knowledge.

To conclude, citizens' confidence accuracy for CC knowledge was lower in relation to scientists' and lower than for science knowledge; but most importantly, it was lower than necessary in relation to the accuracy of their own CC knowledge. Any CC information campaign which only aims at increasing knowledge but does not address confidence in knowledge will fail to acknowledge that citizens not only need accuracy of knowledge,

but also accuracy of confidence. Only if citizens possess high confidence in true statements, and low confidence in false statements, the line between accurate knowledge and false knowledge can, in the future, be less fuzzy.

**Acknowledgments** The authors gratefully acknowledge support by the Excellence Initiative, Institutional Strategy ZUK 5.4 (Scientific Computing in the Social and Behavioral Sciences), and the support of the Heidelberg Center for the Environment, Heidelberg University.

# Supplementary Material

## Supplementary Information (SI) 1: Participants

**Climate Change knowledge: National German sample.** Citizens were recruited via the online polling company YouGov. Data were collected January 2017. The sample (N= 509) was representative for the German population in terms of gender, age, and geographical distribution. Specifically, 262 citizens (53%) were female, mean age was M= 48.5 years (*median* = 51, *range* = 18 – 88 years), and the distribution of number of participants from German federal states was proportional to state size. 72 (14%) held the lowest German school leaving exam (*Hauptschule*), 198 (39%) held middle German school leaving exam (*Realschule*), 103 (20%) the highest German school leaving exam (*Abitur*), 35 (7%) held a Bachelor degree, 67 (13%) a Master's degree, and 5 (1%) a PhD, 20 (4%) other, 9 (1.8%) did not indicate their education. Participants had a range of professions, the largest groups being 48 (9%) commercial clerks, 16 (3%) craftsmen, 12 (2%) engineers, 13 (3%) IT specialists (other than system administrators). The sample‘s distributions of age, political views, and prior CC beliefs are displayed in Figure B.3.

**Climate Change knowledge: Scientists sample.** Scientists were recruited through the German Climate Consortium which distributed the survey invitation among their member organizations. Between June and July 2018, a total of N=449 scientists entered the survey, of which N=207 completed the survey, yielding a completion rate of 46%. Only complete surveys were included in the study. Of the scientists, 138 were male (67%), 69 female (33%), mean age was M= 38.7 years (*median* = 35, *range* = 21 – 78 years), and 181 (87%) were German. Scientists came from a range of different backgrounds such as aerosol physics, atmospheric physics, biology, chemistry, climate research, geoscience, linguistics, mathematics, meteorology, oceanography, paleoceanography, physics, polar science, and sociology; and worked for several organizations, the largest groups being: one of three Max Planck Institutes of Biogeochemistry, Chemistry or Meteorology (79); the Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (27), and the Institute of Meteorology and Climate Research at the Karlsruhe Institute of Technology (23). **Science knowledge: National German sample.** German citizens were recruited through the online polling company YouGov. Data were collected December 2018. For comparability with the CC study, the sample ($N$ = 588) was again nationally representative for the German population in terms of gender, age, and geographical distribution. Specifically, 303 citizens (51%) were female, mean age was $M$ = 48.3 years (*median* = 50, *range* = 18 – 88 years), and the distribution of number of participants from German federal states was proportional to state size. 88 (15%) held the lowest German school leaving exam (*Hauptschule*), 213 (36%) held middle German school leaving exam (*Realschule*), 138 (23%) the highest German school leaving exam (*Abitur*), 45 (8%) held a Bachelor degree, 62 (11%) a Master's degree, and 8 (1%) a PhD, 21 (4%) other, 14 (2%) did not indicate their education.

## SI 2: Method

### Material

**Climate Change knowledge.** Participants judged the veracity of a total of nine true/false statements about CC. Statements were taken from previous research (Sundblad et al., 2009) , with three questions from each of the knowledge domains causes, state,

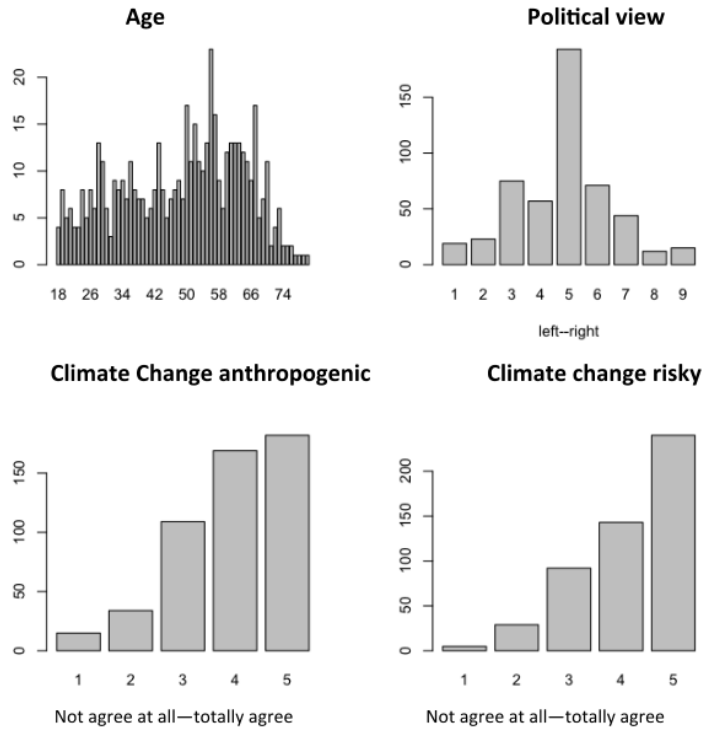



**Figure B.3.** National German sample distributions for age, political view, belief that CC is anthropogenic, and belief that CC is risky.

and consequences of CC. Statements were chosen to reflect a balance of true/false statements across the three knowledge domains. Each statement was introduced with "Science says that..." ("yes, science says that", or "no, science does not say that" [Scientists sample: "yes" or "no"]). Items were carefully selected to ensure item wording does not indicate the veracity of the statement, and such that there was no pattern in the order of true and false statements. A screenshot sample item is given in Figure B.4.

In contrast to some previous CC knowledge scales (Shi et al., 2016; Tobler et al., 2012), an I don't know-option was not included. This was done in order to assess metacognitive confidence in knowledge using a full-range scale (as opposed to a binary classification) that is furthermore logically independent from assessment of knowledge, and therefore does not exclude assessment of (partial) knowledge. Also people have been shown to be differentially adverse to answering "I don't know" versus guessing an answer due to differences in risk aversion, which has been shown to introduce systematic bias in that individuals or groups who tend to skip more questions achieve lower scores, controlling for knowledge (Baldiga, 2013).

**Science Knowledge.** Participants judged the veracity of a total of nine true/false statements about biological and physical science. Statements were taken from the NSF factual knowledge questions published in the National Science Board's Science and Engineering Indicators (National Science Board, 2016). In order to present participants with the same number of true (five) and false (four) statements, and to achieve full structural comparability (statements that can be answered with yes/no) with the CC knowledge items, the question "Does the Earth go around the Sun, or does the Sun go around the Earth" was changed to "The Sun goes around the Earth" (false).

Items were: The center of the earth is very hot (TRUE); the continents on which we live have been moving for millions of years and will continue to move in the future

(TRUE); The Sun goes around the Earth (FALSE); all radioactivity is man-made (FALSE); electrons are smaller than atoms (TRUE); lasers work by focusing sound waves (FALSE); it is the father's gene that decides whether the baby is a boy or a girl (TRUE); antibiotics kill viruses as well as bacteria (FALSE); the universe began with a huge explosion (TRUE).

**Confidence.** After answering each knowledge question, participants were asked: "How certain are you that your answer is correct?" (50%=not at all certain, I was guessing; 100%= certain, I know the answer). Half-range confidence scale (50%−100%) as opposed to full-range (0% − 100%) was used because this was shown to yield better-calibrated confidence judgments (Weber & Brewer, 2003).

**Frequentist assessment of confidence.** We additionally used a frequentist confidence format since this may be easier for participants than the probabilistic scale: "Out of the nine questions please guess: How many did you answer correctly?".

**Prior beliefs.** We assessed prior beliefs about CC by asking "How much do you agree with the following statements?" 1. "CC is mostly caused by humans." 2. "Climate change is risky." (1: not agree at all, 5: totally agree).

**Political view.** Participants were asked to indicate their general political orientation on a nine-point scale (1:left, 9: right).

### Procedure

The studies were conducted online. The procedure was identical for all three studies, except that for the German samples, the survey was conducted in German, whereas for the scientists sample, the survey was conducted in English, and for small other changes given in []. Questions were asked in the following order: Political view, prior beliefs, knowledge and confidence, frequentist assessment of confidence, numerical estimates and demographics (education, profession, [for scientists sample also: Organization], federal state [only citizens sample], gender, age [citizens sample: additional demographics are routinely assessed]).

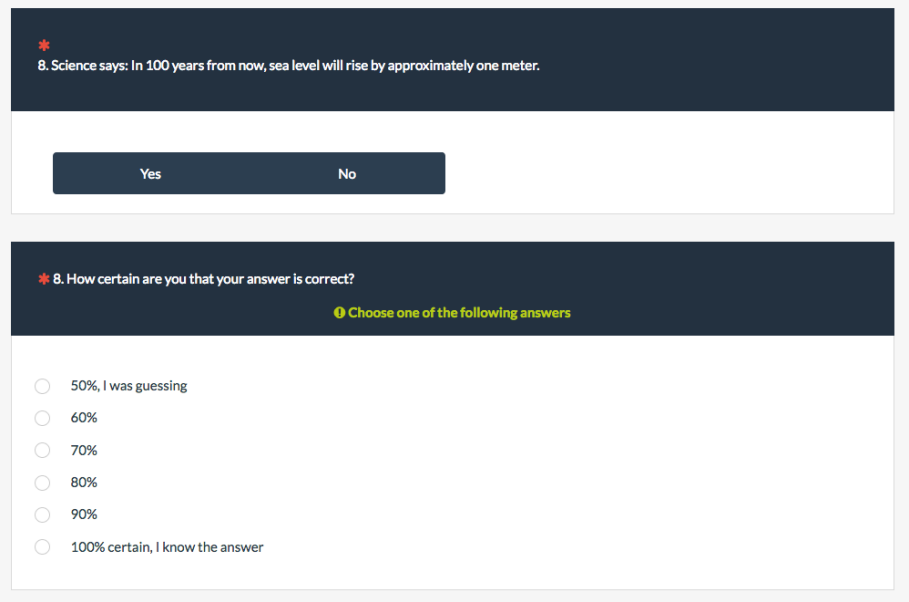



**Figure B.4.** Screenshot of a sample statement together with confidence assessment.

## SI 3: Analyses

**Mokken scale analysis of the CC knowledge scale.**

To test the reliability of the true and false statements as indicators of CC knowledge, we conducted a Mokken scale analysis. Responses were reverse coded for the false statements. Typical 1-0-0 scoring was used such that correct responses were scored with 1, and both incorrect and guessed responses ("50% confidence, I was guessing") were scored with 0 (Shi et al., 2016; Tobler et al., 2012). Counting guessed responses as incorrect was done on line with previous research, and for the theoretical reason that guessed responses should not count as indicators of knowledge [2].

Mokken analysis can be used to construct unidemensional scales that rank participants with respect to the latent trait (in this case, CC knowledge), as well as items with respect to their difficulty. Since Mokken scales thus order not only items, but also respondents, it is subject to stricter assumptions than Cronbach's alpha. Specifically, when relationships between items and a latent trait is given in item characteristic curves (ICCs), the Mokken model puts ICCs to the strict assumptions of double monotonicity, which entails that (1) the ICCs should be monotonically non-decreasing, that is, for any item, the probability of a positive response should never decrease as the latent trait increases; and (2) invariant item ordering, that is, the ordering of the items with respect to their difficulty should be identical for each participant, or: the ICCs should not intersect. The scalability of each item and each scale is measured using Loevinger's coefficient Hi and H, respectively. Cutoff values of $H > 0.3$ are usually assumed for an acceptable scale, and also all items within a scale should be $H_i > 0.3$.

The Mokken scale analysis was conducted using the package *mokken* in R (e.g., Stochl, Jones, & Croudace, 2012). The dimensionality of the scale was assessed. Results showed that the true and false statements constitute two separate subscales, with Loevinger scalability coefficients of $H = .35$ for the scale comprising the four false statements, and $H = .33$ for the scale comprising the five true statements, suggesting that both subscales form unidimensional measures. As Table B.2 shows, the scalability coefficients Hi of all items lie above the threshold of 0.3, suggesting that all respective items form unidimensional subscales. We additionally checked for the assumptions of monotonicity of each item, as well as the assumption of invariant item ordering. There were no significant violations for any item. In sum, the Mokken scale analysis demonstrates that both subscales possess sufficient scalability.

Since 1-0-0 scoring where guessed responses are coded as incorrect proved most appropriate to measure CC knowledge, we use it in all analyses where the accuracy of knowledge is the dependent variable. Results where the accuracy of confidence is the dependent variable are given using number-right scoring where responses are coded solely based on their accuracy. This approach allows us to give both the most reliable estimate of CC knowledge, as well as a full estimate of how well confidence aligns with the actual accuracy of responses.

**Relative confidence sensitivity: Computation of $M_{\text{ratio}}$.** We computed the relative confidence sensitivity $M_{\text{ratio}}$ to measure people's ability to discriminate between correct and incorrect verifications in their confidence, while controlling for their ability to dis-

[2] We also conducted two alternative Mokken analyses: One based on number-right scoring, and one based on three coding categories (2=correct, 1=guessed, 0=incorrect). Both alternative coding schemes resulted in insufficient scales, suggesting that the traditionally used dichotomous coding that treats guessed responses as incorrect is most appropriate to measure CC knowledge.

criminate true and false statements in their knowledge, using the hierarchical Bayes procedure and R code provided under https://github.com/smfleming/HMeta-d (Fleming, 2017). The hierarchical Bayes approach was taken because simulations have shown that hierarchical Bayesian estimation outperforms classical parameter estimation procedures (MLE or SSE point-estimate procedures) when the number of trails/participants is low. One exception is that hierarchical Bayesian $M_{\text{ratio}}$ tends to overestimate $M_{\text{ratios}}$ when d' is high. Thus, true values of $M_{\text{ratios}}$ for scientists and for science knowledge are probably lower by tendency. However, this approach still gives us the most suitable estimate for the most important variable of interest, $M_{\text{ratio}}$ for citizens in CC knowledge.

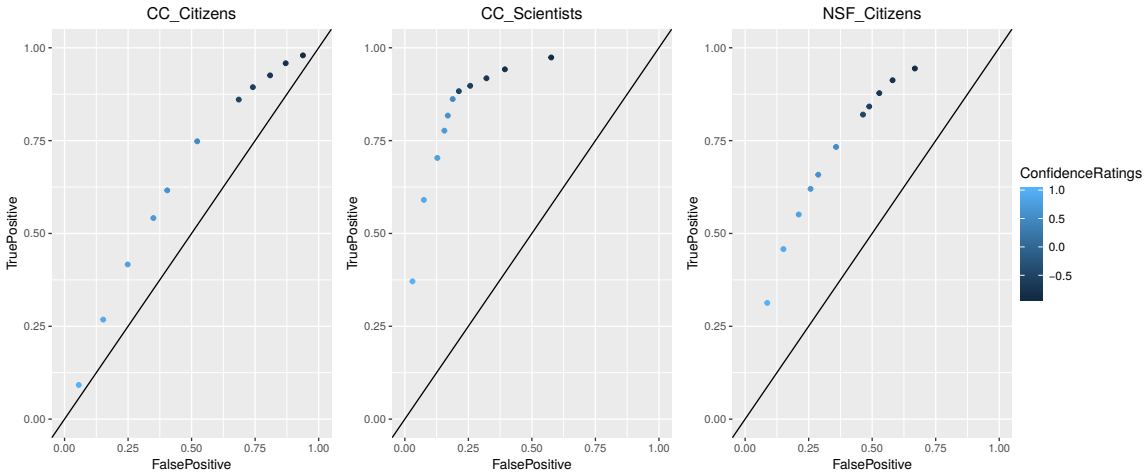
**Equal variance assumption of SDT model.** A prerequisite to estimate $M_{\text{ratio}}$ is equal-variance Gaussian "target" and "lure" distributions. Testing this assumption can be done with help of a z-standardized receiver operating characteristics (zROC) curve, a two- dimensional graph in which the True Positive rate is plotted against the False Positive rate for varying confidence criteria. Equal-variance Gaussian distributions imply a zROC that is linear with a slope of 1. Figure B.5 displays ROC and zROC curves for the three studies, CC knowledge in citizens, science knowledge in citizens, and CC knowledge in scientists. ZROCs are consistent with linearity and approximately unite slope, only scientists' CC knowledge also shows a mild divergence from linearity.

## SI 4: Additional Results

### Accuracy of knowledge for true compared to false statements: number-right scoring of responses.

Concerning the accuracy of verifying true compared to false statements, the following pattern of results emerged: The accuracy of CC knowledge was lower for citizens (M=.61, 95% CI [.60, .63]) compared to scientists ($M$ = .84, 95% CI [.82, .86]), $F(1,714)$ = 387.3, $p < .001$, and generally lower for FALSE statements ($M$ = .65, 95% CI [.62, .67]) compared to TRUE statements ($M$ = .81, 95% CI [.79, .82]), $F(1,714) = 95.4, p < .001$. However, for citizens the difference in accuracy between TRUE ($M$ = .75, 95% CI [.73, .79]) and FALSE statements (M=.48, 95% CI [.45, .50]) was considerably larger than for scientists ($M$ = .86, 95% CI [.84, .88] and $M$ = .81, 95% CI [.78, .85], for true and false statements, respectively), $F(1,714) = 45.0, p < .001$.

**(A) ROC**




**(B) zROC**

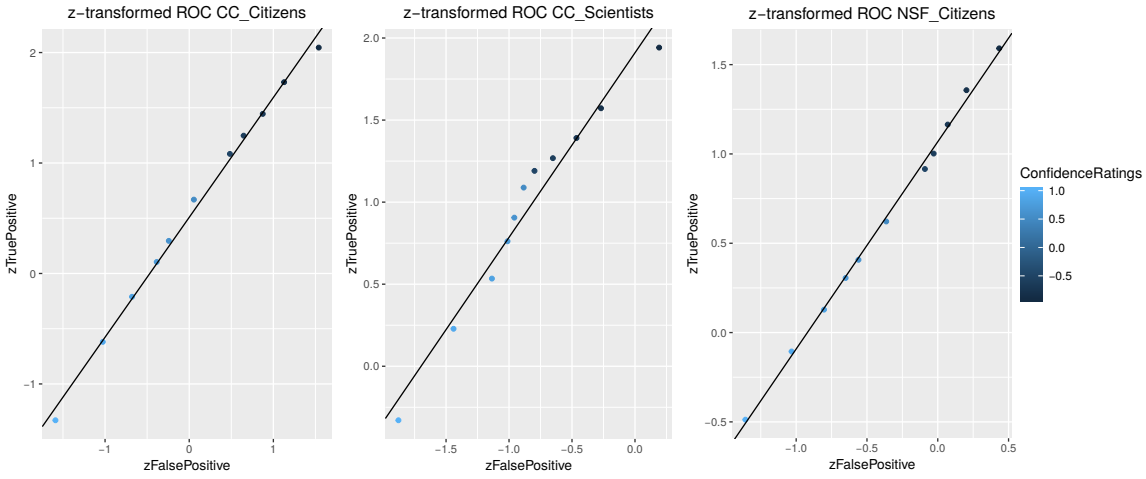



**Figure B.5.** ROC (panel A) and zROC (panel B) curves for the three studies, CC knowledge in citizens, science knowledge in citizens, and CC knowledge in scientists. ROC curves plot the True Positive rate against the False Positive rate for varying confidence criteria; zROCs plot True Positive and False Positive rates in z-space. Equal-variance Gaussian distributions imply zROCS that are linear with slopes of 1.

**Table B.2.** Mokken scale scalability coefficients Hi, within each of the two subscales (true statements, false statements).

| **Veracity of statement** | **Statement** | **Loevinger scalability coefficients $H_i$ (Standard error)** |
|---|---|---|
| FALSE | 1 The global average temperature in the air has increased approx. 3.1 °C the past 100 years. | 0.301 (0.043) |
| FALSE | 6 An increasing amount of greenhouse gases increases the risk of more UV-radiation and therefore a larger risk of skin cancer. | 0.406 (0.046) |
| FALSE | 7 The 1990s was the warmest during the past 100 years | 0.334 (0.044) |
| FALSE | 8 Climate change is mainly caused by a natural variation in sunbeam and volcanic eruption. | 0.400 (0.070) |
| TRUE | 2 The global change in temperature the past 100 years is the largest during the past 1000 years. | 0.314 (0.036) |
| TRUE | 3 Carbon dioxide concentration has increased more than 30% in the atmosphere during the past 250 years. | 0.342 (0.034) |
| TRUE | 4 The increase of greenhouse gases is mainly caused by human activities. | 0.338 (0.045) |
| TRUE | 5 In 100 years from now, sea level rise will be approximately one meter. | 0.319 (0.037) |
| TRUE | 9 The blanket of snow in the Northern hemisphere has decreased approximately 10% since the 1960. | 0.327 (0.034) |

# References


Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science*, *60*(2), 434–448.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–428.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5), S2–S23.

Bornstein, B. H., & Zickafoose, D. J. (1999). " i know i know it, i know i saw it": The stability of the confidence–accuracy relationship across domains. *Journal of Experimental Psychology: Applied*, *5*(1), 76–88.

Boussalis, C., & Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change*, *36*, 89–100.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*(1), 11–30.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, *92*(5), 938–956.

Burton, R. F., & Miller, D. J. (1999). Statistical modelling of multiple-choice and true/false tests: ways of considering, and of reducing, the uncertainties attributable to guessing. *Assessment & Evaluation in Higher Education*, *24*(4), 399–411.

Dienes, Z., & Fahey, R. (1995). Role of specific instances in controlling a dynamic system. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 848–862.

Elsasser, S. W., & Dunlap, R. E. (2013). Leading voices in the denier choir: Conservative columnists' dismissal of global warming and denigration of climate science. *American Behavioral Scientist*, *57*(6), 754–776.

Fleming, S. M. (2017). Hmeta-d: hierarchical bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1), nix007.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876.

Hadar, L., Sood, S., & Fox, C. R. (2013). Subjective knowledge in consumer financial decisions. *Journal of Marketing Research*, *50*(3), 303–316.

Hiles, S. S., & Hinnant, A. (2014). Climate change in the newsroom: Journalists' evolving standards of objectivity when covering global warming. *Science Communication*, *36*(4), 428–453.

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: capturing decision tendencies in a fictitious medical test. *Metacognition and Learning*, *9*(1), 25–49.

Johnson, D. D. (2009). *Overconfidence and war*. Cambridge, Massachusetts: Harvard University Press.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1304–1316.

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). Nasa faked the moon landing—therefore,(climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science*, *24*(5), 622–633.

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, *111*(30), 10984–10989.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430.

Meyer, A. N., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Internal Medicine*, *173*(21), 1952–1958.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: the influence of incentives and feedback on exam predictions. *Metacognition and Learning*, *6*(3), 303–314.

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*(387), 489–500.

National Science Board. (2016). *Science and engineering indicators 2016*. Arlington VA: National Science Foundation.

Ortoleva, P., & Snowberg, E. (2015). Overconfidence in political behavior. *American Economic Review*, *105*(2), 504–35.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1287–1296.

Park, J., Konana, P., Gu, B., Kumar, A., & Raghunathan, R. (2010). Confirmation bias, overconfidence, and investment performance: Evidence from stock message boards. *McCombs Research Paper Series No. IROM-07-10*.

Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual-differences approach. *Journal of Behavioral Decision Making*, *18*(1), 1–27.

Sanbonmatsu, D. M., Posavac, S. S., Kardes, F. R., & Mantel, S. P. (1998). Selective hypothesis testing. *Psychonomic Bulletin & Review*, *5*(2), 197–220.

Shi, J., Visschers, V. H., Siegrist, M., & Arvai, J. (2016). Knowledge as a driver of public perceptions about climate change reassessed. *Nature Climate Change*, *6*(8), 759–762.

Simon, M., & Houghton, S. M. (2003). The relationship between overconfidence and the introduction of risky products: Evidence from a field study. *Academy of Management Journal*, *46*(2), 139–149.

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric irt method in empirical research for applied health researchers. *BMC Medical Research Methodology*, *12*(1), 74.

Sundblad, E.-L., Biel, A., & Gärling, T. (2009). Knowledge and confidence in knowledge

about climate change among experts, journalists, politicians, and laypersons. *Environment and Behavior*, *41*(2), 281–302.

Tobler, C., Visschers, V. H., & Siegrist, M. (2012). Consumers' knowledge about climate change. *Climatic Change*, *114*(2), 189–209.

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology*, *88*(3), 490–499.

Williamson, P. (2016). Take the time and effort to correct misinformation. *Nature News*, *540*(7632), 171.

# C | Manuscript 3

## Extrapolation performance underestimates rule learning: Evidence from the function-learning paradigm

Nadia Said[1,2 *], Helen Fischer[1]

**1 Institute of Psychology, Heidelberg University, Hauptstr. 47–51, 69117 Heidelberg, Germany**
**2 Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany**

*** nadia.said@psychologie.uni-heidelberg.de**

## Abstract

Understanding the development of non-linear processes such as economic or population growth is an important prerequisite for informed decisions in those areas. In the function-learning paradigm, people's understanding of the function rule that underlies the to-be predicted process is typically measured by means of extrapolation accuracy. Here we argue, however, that even though accurate extrapolation necessitates rule-learning, the reverse does not necessarily hold: Inaccurate extrapolation does not exclude rule-learning. Experiment 1 shows that more than one third of participants who would be classified as "exemplar-based learners" based on their extrapolation accuracy were able to identify the correct function shape and slope in a rule-selection paradigm, demonstrating accurate understanding of the function rule. Experiment 2 shows that higher proportions of rule learning than rule-application in the function learning paradigm is not due to (i) higher a priori probabilities to guess the correct rule in the rule-selection paradigm; nor is it due to (ii) a lack of simultaneous access to all function values in the function-learning paradigm. We conclude that rule application is not tantamount to rule-learning, and that assessing rule-learning via extrapolation accuracy underestimates the proportion of rule learners in function-learning experiments.

**Keywords** function-learning, rule-based vs exemplar-based learners, non-linear processes, understanding

# Introduction

Non-linear processes abound in human life, ranging from small-scale examples such as fuel consumption to large-scale, global processes such as the developments of economies, populations, or greenhouse gas emissions. A long-standing question in the cognitive literature is whether humans acquire an understanding of the underlying function rule when making predictions about the development of such processes. This question is often investigated in the function-learning paradigm, where participants learn about the beginning of a process with input-output pairs sampled from the underlying function, and predict the future development of that process. Typically, extrapolation accuracy, the distance between participants' predictions and the actual function values, is used to infer whether participants acquired an understanding of the function rule: It is argued that when predictions are sufficiently close to the correct function, participants must have learned the correct function rule; when predictions deviate sufficiently from the correct function, for example by showing flat extrapolations of highly non-linear processes, participants did not learn the correct function rule.

Here we argue, however, that even though sufficiently correct extrapolations necessitate previous rule learning, the reverse does not necessarily hold: Incorrect extrapolations do not exclude rule-learning. Rather, incorrect extrapolations can mirror alternative processes, such as implementation failure. Based on this theoretical argument, we investigate in how far accuracy of extrapolations coincides with rule-learning of three different exponential declining [1] processes in two function-learning experiments.

In function-learning experiments, participants learn to predict continuous output (y-values) from continuous input (x-values) variables. To do so, participants are presented with an input value (for example, a time point; Fischer & Holt, 2016), and then predict the corresponding outcome value. During training, participants receive feedback on their predictions; during test (interpolation or extrapolation), no feedback is given.

Research has shown that there are two fundamentally different types of learning style that participants may employ in function-learning experiments: Rule-based and exemplar-based learning (McDaniel, Cahill, Robbins, & Wiener, 2014). In exemplar-based models, participants try to memorize the given exemplars, whereas in rule-based models, participants learn the function rule underlying the to-be predicted process. Among the class of exemplar-based models, at least three different accounts exist on what participants do with the stored exemplars during extrapolation. Simple exemplar-based models, first, hold hat participants extrapolate using exemplars that are identical (or at least highly similar) to learned exemplars, thereby for example producing flat extrapolations that correspond to the stored exemplars (DeLosh, Busemeyer, & McDaniel, 1997). The Extrapolation-Association Model (EXAM; DeLosh et al., 1997), second, holds that participants retrieve the two best-matching exemplars, and extrapolate linearly through these exemplars. And the Population of Linear Experts model (POLE; Kalish, Lewandowsky, & Kruschke, 2004), third, holds that participants store mappings between x-values and matching linear functions that they retrieve for extrapolation. Rule-based models, in contrast, hold that participants use the training information provided to abstract a rule describing the ensemble of x-y pairings (Mcdaniel & Busemeyer, 2005).

While function-learning studies differ in many aspects, such as the functions used

[1] Please note that the term "exponential declining" conventionally refers to $e^{-x}$. However, we will use this term throughout the paper to refer to $-e^{x}$ which is the negative of the exponential function.

(for example V-shaped, McDaniel et al., 2014 , or periodic, Bott & Heit, 2004), the input format (entering a number, MacKinnon & Wearing, 1991, or clicking on a bar, McDaniel et al., 2014), the number of learning trials (for example 200, McDaniel et al., 2014, or 10 Fischer & Holt, 2016 ), and the experimental design (such as one learning, followed by one extrapolation phase, Lewandowsky, Kalish, & Ngang, 2002, as opposed to several interspersed extrapolation phases, Bott & Heit, 2004), most function-learning studies have in common that extrapolation accuracy is used as a proxy for learning style. Specifically, not only is high extrapolation accuracy interpreted as signaling rule-learning, but also low extrapolation accuracy is interpreted as signaling exemplar-based or simple exemplar-based learning.

In one of the classic function-learning studies (DeLosh et al., 1997), absolute deviations of participants' extrapolations from the correct quadratic function were used to infer learning-type, and the authors concluded that flat extrapolations to a quadratic function were reflective of simple exemplar-based learning. In another experiment using quadratic functions (Lewandowsky et al., 2002), about 20% of participants were classified as being unable to learn the underlying rule based on the low fit of their extrapolations with the correct function. In a study with periodic functions participants were able to extrapolate (surprisingly) accurately compared to the results in other studies. The authors suggested that this difference in results may be due to participants in other experiments being unable to learn the function rule (Bott & Heit, 2004). And in a more recent study explaining individual differences in learning style, participants who showed relatively flat extrapolations to a V-shaped function were categorized as exemplar-, as opposed to rule-based learners.

The reasoning behind these studies is summarized in a theoretical argument of Kwantes and Neal (2006) who argue: "To show that you have really learned the concept, you need to demonstrate two things: You need to perform reasonably well on new items that fall within the bounds set by the training examples (so-called interpolation items), and you need to perform reasonably well on new items that fall outside the bounds set by the training examples (so-called extrapolation items)". The authors thus argue that extrapolation accuracy separates participants who learned a function rule (or "concept") from those who did not learn a function rule.

In sum, function-learning studies reviewed here share the (often implicit) assumption that provided that, participants did acquire an understanding of the correct rule, they also apply it when extrapolating. If this assumption holds, inaccurate extrapolations can indeed be interpreted as signaling the absence of rule-learning. If this assumption does not hold, however, inaccurate extrapolations are also compatible with accurate rule-learning. In other words: while accurate extrapolations are an implication of rule-learning, inaccurate extrapolations indicate learning styles other than rule-learning if, and only if, the assumption of rule application holds.

Here we put the assumption of rule-application given rule-learning to an experimental test. The reasoning behind is that participants neither need to apply a learned rule per se, nor do they need to apply it correctly. For example, participants may fail to accurately implement a learned rule, potentially because deriving extrapolation points from the abstracted rule requires substantial cognitive resources such as working memory capacity (Fischer & Holt, 2016). Also adjusting each consecutive extrapolation to previous extrapolations may be error-prone. Participants may even deliberately use comparatively simple linear extrapolations despite better knowledge.

We will use the term *(a) function rule* to refer to the general trend (declining), shape

(exponential), and slope of a presented process. Depending on whether participants acquire an understanding of only one, two, or all three of these aspects, increasingly stricter conditions of rule-learning are met. We use the term *(b) extrapolation style* to refer to participants' extrapolations as either (1) simple exemplar-based, that is, linear extrapolation parallel to the x-axis (DeLosh et al., 1997), (2) exemplar-based, that is, linear extrapolation through the two best-matching learning points (DeLosh et al., 1997) or 3. rule-based, that is, extrapolation according to a function rule.

We report the results of two large function-learning experiments demonstrating that a substantial proportion of participants who would be classified as "exemplar-based learners" based on their extrapolation accuracy actually acquired an understanding the correct function rule. These results shed doubt on the assumption of rule-application given rule-learning. Furthermore, these results also deliver a comprehensive estimate of the extent to which rule-learning is underestimated by means of extrapolation accuracy.

# Experiment 1

Experiment 1 investigated the extent to which participants who would be classified as "exemplar-based" based on their extrapolation accuracy in a classical function-learning paradigm had acquired an understanding of the function rule. To do so, participants completed two tasks: A standard function-learning task to assess extrapolation accuracy, and a rule-selection task to assess whether participants could identify the correct function shape and slope. In the function-learning task, participants extrapolated the development of three exponential declining processes. The task consisted of one learning phase, and one extrapolation phase per process (Fischer & Holt, 2016). Participants received the instructions to extrapolate the development of different types of bacteria cultures, "Ain", "Bin", and "Cin". After the learning phase, participants completed the rule-selection task. Participants identified the function rule of the process they had just learned by selecting one of a total of six pictures displaying different function shapes and slopes.

## Method

**Participants.** A total of 520 participants completed the experiment. Participants were recruited over MTurk, and received 1.05$. Data from 9 participants were removed because they already participated in the pretest, thus the data from n=511 participants were included in the final data set. Participants were instructed not to use pen, paper or any other help during the study. The sample size was determined by a power analysis based on a small effect size of $r = .18$ (Fischer & Holt, 2016), $p = .05$ and $\beta = 0.8$, resulting in a sample size of $n = 240$ per condition.

**Materials. (a) Processes.** We used three variations of exponential functions based on the equation:

$$y = 1500 - e^{a \cdot (x+50)+2} \tag{C.1}$$

with $a = [0.045, 0.040, 0.046]$. In the following we refer to the function with $a_1 = 0.045$ as process Ain, with $a_2 = 0.040$ as Bin, and with $a_3 = 0.046$ as Cin.

Exponential declining functions were chosen because they represent the most difficult function to extrapolate (Busemeyer, Byun, Delosh, & McDaniel, 1997), and hence because of their strong deviation from the "cognitive default" of positive linearity, a

particularly strong case for rule-learning can be made (DeLosh et al., 1997; Kalish et al., 2004; Kwantes & Neal, 2006).

**(b) Rule-selection task**. Participants were presented with three graphs, displaying three different function shapes of two different slopes each: two linearly decreasing functions, two exponential declining functions, and two Gaussian functions. Participants were asked to indicate, "Which graph describes the shape of the development of the bacteria best?". Participants entered the number of the graph into a text box (Figure C.1).

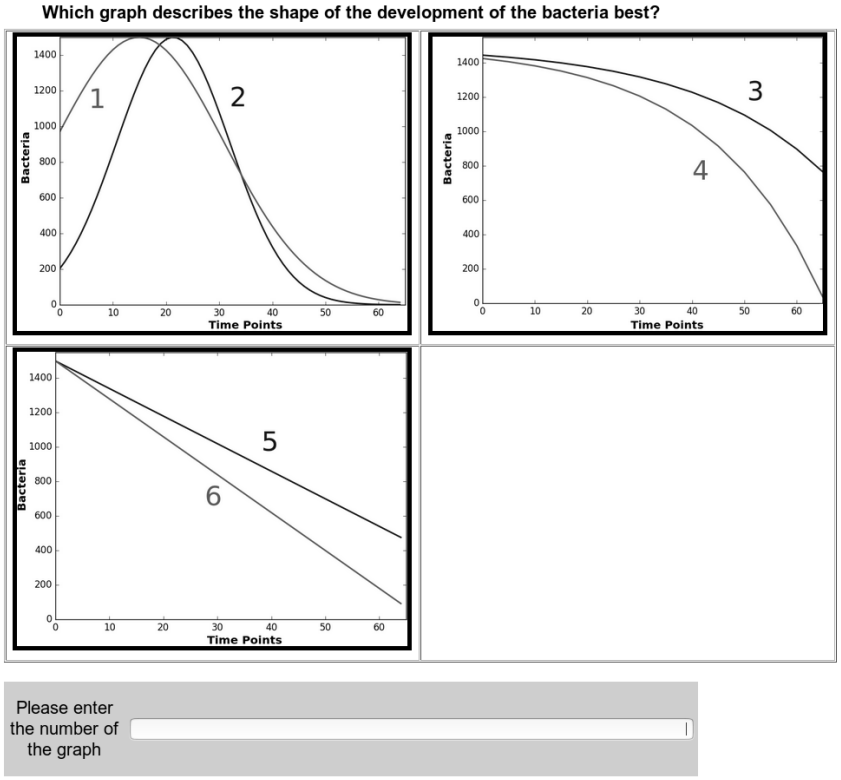



**Figure C.1.** Example of rule-selection task for process Bin. The figure displays the functions given to participants with the instructions to choose the process they have just learned (Correct: Function no. 3.)

Linear functions were chosen as they represent the most basic and frequently found extrapolation style (Busemeyer et al., 1997; Carroll, 1963) that is furthermore employed in exemplar-based, as well as simple exemplar-based extrapolations; Gaussian functions were chosen to assess whether participants believed the process to be non-monotonical; and the exponential declining functions were chosen to assess whether participants could correctly identify the correct function shape, and potentially also slope. The slopes displayed were $0.045_{\text{correct}}$ and 0.040 for Ain, $0.040_{\text{correct}}$ and 0.046 for Bin, and $0.046_{\text{correct}}$ and 0.043 for Cin. Slopes for all functions were chosen in a way such that y-values of functions remained between 0 and 1500.

**Procedure.** Each participant extrapolated 3 processes. Each process consisted of 13 trials, 8 learning and 5 extrapolation trials. At the beginning of each process, participants were given the starting point of that process, that is, the number of bacteria at time point 0 (1430 for Ain, 1445 for Bin, and 1426 for Cin). During each trial, participants were shown the current time point and predicted the number of bacteria for that time

point by entering their extrapolation as a number into a text box ("I guess the number of bacteria is ..."). During the learning phase, participants received feedback in terms of the correct number of bacteria for each time point, immediately after entering their extrapolation ("You guessed: ... . Actual number: ...").

To control for the effect of seeing different function shapes in the rule-selection task on extrapolation accuracy, participants were randomly allocated to complete the rule-selection task immediately before, or immediately after the extrapolation phase.

## Results

**Dependent variables.** To measure extrapolation accuracy in the function-learning task, the relative root mean square error (rRMSE) was used :

$$\text{rRMSE}_i = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left( \frac{(y_i - z_i)}{z_i} \right)^2}, \tag{C.2}$$

with $y_i$ : extrapolation and $z_i$ : correct function value.

To measure understanding of the function rule, the rule selection task distinguished between identifying the correct function *shape*, that is, choosing either of the two exponential functions; and additionally identifying the correct function *slope*, that is, choosing the exponential function AND the correct slope.

**Outliers.** We excluded individual extrapolations more than five standard deviations above or below the mean of each time point (0.51% of the total number of extrapolations). If more than two out of the five extrapolation trials were excluded, the process was treated as missing for this participant (Fischer & Holt, 2016). In total, for 4 participants processes were excluded, resulting in 509 participants for process Ain, and 510 participants for processes Bin and Cin.

**Order of rule-selection and extrapolation.** In order to assess whether showing participants pictures of the correct function impacted extrapolation accuracy, we compared extrapolation accuracy in the group completing the rule-selection task before ($M = 3.52$, $SD = 2.07$) versus after ($M = 3.55$, $SD = 2.02$) the extrapolation phase. Accuracy was marginally but not significantly higher in the group performing the rule-selection task before the extrapolation phase, $F(3, 503) = 2.36$, $p = .07$, Pillais' Trace = 0.014. Thus, in the following, results for both groups are presented together.

**Extrapolation accuracy by function slope.** We assessed whether extrapolation accuracy varied by function slope. As the assumptions of homogeneity of variances ($F(2, 1526) = 324.12$, $p < .001$) as well as of normality ($W = 0.69$, $p < .001$) were not met, Kruskal-Wallis rank sum tests were conducted. Results showed that extrapolation accuracy differed between the three processes as a function of slope $\chi^2(2) = 1204.3$, $p < .001$ in that prediction accuracy was higher for Bin ($M_{Bin} = 0.30$, $SD_{Bin} = 1.00$) compared to Ain ($M_{Ain} = 1.50$, $SD_{Ain} = 1.12$) $z = 17.82$, each $p < .001$. Interestingly, the drop in accuracy was particularly steep from Bin to the steepest function Cin ($M_{Cin} = 8.86$, $SD_{Cin} = 5.4$) $z = -34.70$, $p < .001$. These results are in line with previous findings that participants have a tendency towards linear extrapolation, and hence extrapolation accuracy decreases as function slope increases.

**Proportion of participants per extrapolation style.** Participants' extrapolation styles were categorized based on their extrapolation accuracy (McDaniel et al., 2014). Specifically, we determined the deviation (rRMSE) of each participant's extrapolation

accuracy including a 95% confidence interval from these three cases: (1) **rRMSE$_{\mathbf{Exp}}$**: The deviation from the correct function, (2) **rRMSE$_{\mathbf{LinSlope0}}$**: the deviation from a linear extrapolation with slope 0 through the last learning point, (3) **rRMSE$_{\mathbf{Lin}}$**: the deviation from a linear extrapolation through the last two learning points. Confidence intervals were calculated as follows:

$$\mathrm{CI}_{\pm} = \bar{y}_i \pm 2.776 \cdot \frac{\sigma}{\sqrt{n}} \tag{C.3}$$

with $\bar{y}_i$ : rRMSE for each time point i.

Extrapolations were categorized into the different groups based on whether their entire

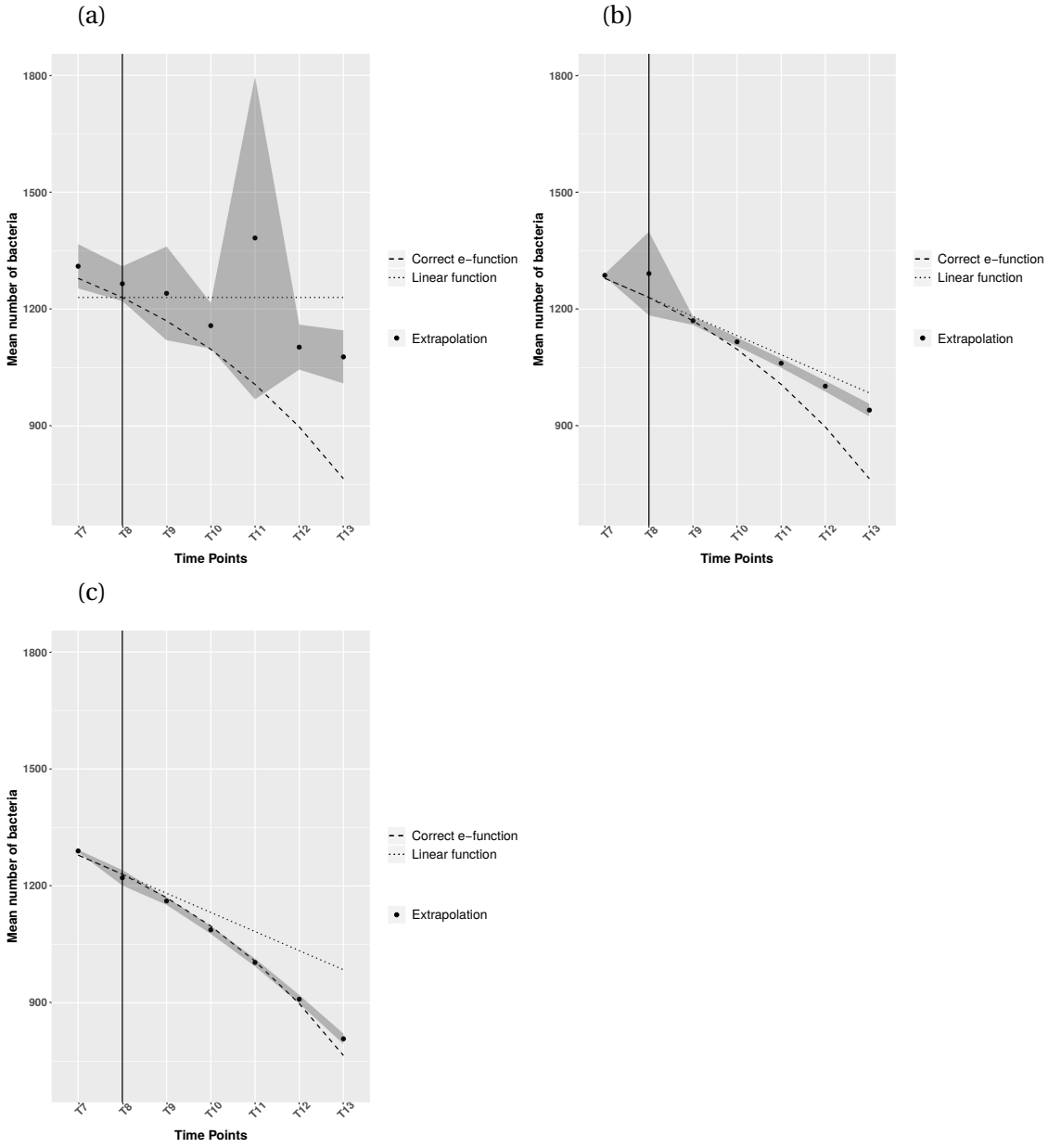



**Figure C.2.** Example of extrapolations for process Bin. The figure displays the number of bacteria predicted by participants who were classified as (a) simple-exemplar based , (b) exemplar-based and (c) rule-based learners based on their extrapolation accuracy. Dark-grey area: 95% confidence band.

$\mathrm{CI}_{\pm}$ was (a) above **rRMSE$_{\mathbf{LinSlope0}}$**, indicating extrapolation parallel to the x-axis (simple

exemplar-based extrapolation), (b) below $\mathbf{rRMSE_{LinSlope0}}$ but above $\mathbf{rRMSE_{Lin}}$, indicating linear extrapolation through the last two learning time points (exemplar-based extrapolation); or (c) below $\mathbf{rRMSE_{Lin}}$, indicating the most accurate extrapolation (rule-based extrapolation). Figure C.2 displays an exemplatory categorization for process Bin.

**Relationship between rule-learning and rule-based extrapolation.** In order to investigate the relationship between rule-learning and extrapolation style, we determined the association between rule selection and extrapolation style in logistic regressions, separately for each of the three processes, Ain, Bin, and Cin. We distinguished between (i) choosing the correct function shape, and (ii) choosing the correct function shape AND slope. Choosing the correct function shape, (i), was significantly related to extrapolation style for each process, $\chi^2_{\text{Ain}}(1) = 11.13$, $\chi^2_{\text{Bin}}(1) = 14.86$, and $\chi^2_{\text{Cin}}(1) = 13.73$, each $p < .001$, suggesting that, unsurprisingly, learning of the correct function shape was related to rule-based extrapolations. Choosing the correct function shape AND slope was associated with extrapolation style for processes Ain and Cin, $\chi^2_{\text{Ain}}(1) = 36.93$, $\chi^2_{\text{Cin}}(1) = 16.83$, $p < .001$, but not for process Bin, $\chi^2_{\text{Bin}}(1) = 1.11$, $p = .29$, suggesting that learning of the correct function slope was related to rule-based extrapolations, except for extrapolation of the process with the lowest slope.

To test the extent to which exemplar-based, or simple exemplar-based extrapolation styles exclude rule-learning, Table C.1 displays the proportion of participants who could correctly identify the correct function shape, separately for each extrapolation style. In the group displaying simple exemplar-based extrapolation, 46% of participants were able to identify the correct function shape, while only 31% of participants estimated the function was actually linear. A similar pattern held for the group displaying exemplar-based extrapolation, where 61% of participants chose the correct function shape, and only 24% estimated the function to be linear. For the group displaying rule-based extrapolation, 65% chose the correct function shape. In sum, the relative majority of participants displaying simple exemplar-based extrapolations, and even the absolute majority of participants displaying exemplar-based extrapolations could identify the correct function shape as exponential declining.

In total, the proportion of participants who had acquired an understanding of the correct function rule in the learning phase (as indicated by the rule-selection task) but did not apply this in the extrapolation phase (as indicated by classifications of their extrapolation style based on extrapolation accuracy) was 47% for process Ain, 45% for Bin, and 36% for Cin (Table C.2).

As the stricter criterion of rule-learning, (ii) we determined the number of participants choosing not only the correct function shape but also function slope, per extrapolation style. As Table C.3 shows, the proportion of participants choosing the correct function slope increased with extrapolation style, from simple exemplar-based, to exemplar-based, to rule-based. Across all three processes, 24% of participants displaying simple-exemplar-based extrapolations, and 37% of participants displaying exemplar-based extrapolations, were able to identify the correct function slope. These results suggest that even among those participants who had acquired a deep understanding of the function rule in that they could identify the correct shape AND slope, a considerable proportion of participants did not apply this understanding when extrapolating, but rather used exemplar-based, or even simple exemplar-based extrapolation styles.

The last column of Table C.3 displays the proportion of participants who could identify the correct slope, out of those who could identify the correct shape. Results

**Table C.1.** Proportion of participants selecting one of the three function shapes in the rule-selection task, per extrapolation style.

| Style | Process | Gaussian | Linear | Exponential | $\chi^2$-test |
|---|---|---|---|---|---|
| Simple exemplar-based | Ain | 59 (21%) | 73 (27%) | 144 (52%) | $\chi^2(2, N = 276) = 45.15^{***}$ |
| | Bin | 59 (28%) | 45 (22%) | 104 (50%) | $\chi^2(2, N = 208) = 27.41^{***}$ |
| | Cin | 49 (18%) | 127 (47%) | 97 (35%) | $\chi^2(2, N = 273) = 34.02^{***}$ |
| | **Total** | **22%** | **32%** | **46%** | |
| Exemplar-based | Ain | 19 (14%) | 26 (19%) | 93 (67%) | $\chi^2(2, N = 138) = 72.57^{***}$ |
| | Bin | 36 (18%) | 41 (20%) | 125 (62%) | $\chi^2(2, N = 202) = 74.27^{***}$ |
| | Cin | 23 (14%) | 56 (33%) | 88 (53%) | $\chi^2(2, N = 167) = 37.95^{***}$ |
| | **Total** | **15%** | **24%** | **61%** | |
| Rule-based | Ain | 16 (17%) | 14 (15%) | 65 (68%) | $\chi^2(2, N = 95) = 52.70^{***}$ |
| | Bin | 16 (16%) | 12 (12%) | 72 (72%) | $\chi^2(2, N = 100) = 67.52^{***}$ |
| | Cin | 15 (22%) | 17 (24%) | 38 (54%) | $\chi^2(2, N = 70) = 13.91^{***}$ |
| | **Total** | **18%** | **17%** | **65%** | $^{***}$ $p \leq .001$ |

**Table C.2.** Proportion of participants identifying the correct exponential function shape, and applying exemplar-based and simple exemplar-based extrapolations, per process (Ain, Bin, and Cin).

| Style | Process | Exponential |
|---|---|---|
| Simple-& Exemplar-based | Ain | 237 (47%) |
| | Bin | 229 (45%) |
| | Cin | 185 (36%) |

show that while for processes Bin and Cin, around half of participants who could identify the correct shape also identified the correct slope, results were different for the steepest process Cin in that the vast majority of participants who identified the correct shape also identified the correct slope.

**Table C.3.** Proportion of participants selecting the correct function shape AND slope, per extrapolation style.

| Style | Process | Correct shape AND slope | **Total** | $\frac{\text{Correct shape AND slope}}{\text{Correct shape}} \cdot 100$ |
|---|---|---|---|---|
| Simple exemplar-based | Ain | 44 (16%) | **24**% | 31% |
| | Bin | 65 (31%) | | 63% |
| | Cin | 70 (26%) | | 72% |
| Exemplar-based | Ain | 51 (37%) | **37**% | 55% |
| | Bin | 76 (38%) | | 61% |
| | Cin | 62 (37%) | | 70% |
| Rule-based | Ain | 36 (45%) | **44**% | 55% |
| | Bin | 50 (36%) | | 69% |
| | Cin | 35 (50%) | | 92% |

Interestingly, for all three extrapolation styles, the proportion of participants who could correctly identify the correct function shape dropped 15% for the steepest function Cin compared to the proportion of participants who could correctly identify the correct function shape for processes Ain and Bin. This result contrasts results on extrapolation accuracy for Cin ($M_{Cin} = 8.86$) which dropped by 83% compared to Ain ($M_{Ain} = 1.50$), and even 97% compared to Bin ($M_{Bin} = 0.30$).

**Prevalence of rule-learning based on rule-selection task vs. extrapolation accuracy.** Table C.4 compares the proportion of participants who would be classified as rule-learners, based on accuracy in the rule-selection task as opposed to extrapolation accuracy. Results show that while the minority (< 20%, $\chi^2(2, N = 1529) = 238$, $p < .001$) of participants would be classified as rule-learners based on extrapolation accuracy, the relative majority of participants could identify the correct function shape (> 50%, $\chi^2(2, N = 1529) = 308$, $p < .001$), and even slope (> 25%, $\chi^2(3, N = 1529) = 53.33$, $p < .001$).

**Table C.4.** Proportion of participants classified as rule-based, exemplar-based or simple exemplar-based learners in the function-learning and rule-selection paradigm, per process (Ain, Bin and Cin).

| | **Traditional FL paradigm** | | | **Rule-selection paradigm** | | | |
|---|---|---|---|---|---|---|---|
| | Simple exemplar-based | Exemplar-based | Rule-based | Gaussian | Linear | Exponential (shape) | Exponential (slope) |
| Ain | 276 (54%) | 138 (27%) | 95 (19%) | 94 (19%) | 113 (22%) | 302 (59%) | 138 (27%) |
| Bin | 208 (41%) | 202 (40%) | 100 (19%) | 111 (22%) | 98 (19%) | 301 (59%) | 177 (35%) |
| Cin | 273 (53%) | 167 (33%) | 70 (14%) | 87 (17%) | 200 (39%) | 223 (44%) | 167 (33%) |
| **Total** | 50% | 33% | 17% | 19% | 27% | 54% | 32% |

### Summary 1

Experiment 1 showed that a substantial proportion of participants who had acquired an understanding of the correct function rule in the learning phase of a function-learning experiment (as indicated by the rule-selection task) did not apply their understanding in the extrapolation phase (as indicated by classifications of their extrapolation style based on extrapolation accuracy). Specifically, out of those participants who would be classified as exemplar-based learners based on their extrapolation accuracy, 61% were able to accurately identify the correct function shape, and 37% were able to identify the correct function shape AND slope. Moreover, only 32% of participants showing simple exemplar-based, and 24% of participants showing exemplar-based extrapolations believed the functions to be actually linear. These results suggest that (i) extrapolation accuracy underestimates rule-learning in the classical function-learning paradigm, and that (ii) up to certain extent, participants are aware of the non-linearity of the process, even if their extrapolations suggest otherwise.

However, there are two limitations to Experiment 1. First, even though selection of the exponential shape was clearly above guessing rate in all three extrapolation styles, a priori probabilities to guess correctly were considerably higher in the rule-selection task compared to extrapolation in the standard function-learning task. To address this limitation, Experiment 2 required participants to indicate their understanding of the function shape not by selecting a picture, but by drawing their understanding of the function shape into a grid. And second, it remains unclear why participants fail to apply their rule understanding in the classical function-learning paradigm. One plausible explanation could be that participants make implementation errors in the classical function-learning paradigm where x-y pairings are given only consecutively, whereas in the rule-selection task participants have simultaneous access to all function values. To address this second limitation, Experiment 2 introduced another control condition, where all extrapolations were displayed on the same page for a given process, so that current as well as all previous extrapolations were visible to participants.

## Experiment 2

Experiment 1 showed that a substantial proportion of exemplar-based or even simple exemplar-based extrapolators had acquired an accurate understanding of the function rule, indicating that the number of rule-learning was underestimated previously by using extrapolation accuracy as a proxy for rule-learning. Experiment 2 provided equal a priori probabilities between a standard function-learning condition and an alternative paradigm in which participants indicated their understanding of the process by drawing the function into a grid (grid condition). Furthermore, we added a third condition in which participants' extrapolations were displayed on one screen instead of in consecutive order (summary function-learning condition).

### Method

**Participants.** A total of 918 MTurk participants completed the experiment. Sample size was determined by computing a power analysis based on $f = .1$, $p = .05$ and $\beta = .8$, resulting in a sample size of $n = 323$ per condition. Data from 176 participants were removed because inspection of MTurk IDs revealed participants had taken part in either

Experiment 1, or a pretest. Participants were instructed not to use pen, paper or a calculator during the study. We included a statement at the end of the study in which participants had to confirm that they did not do so ("I confirm that I did NOT use a calculator, pen or paper"). A total of 15 participants were excluded because they reported having used aids. Additionally, in order for the grid condition to be fully comparable with the function-learning conditions, we checked for the position (one click per time point) and order (clicks starting from time point T1 followed by T2 and so on) of extrapolations. Out of the 232 participants in the grid condition, 67 were excluded because they violated this requirement. A total of N=660 participants were included in the final data set.

**Materials.** Two of the processes from Experiment 1 were used, the development of the bacteria cultures Ain and Bin.

**Procedure.** Participants were randomly assigned to three groups: (a) A *standard function-learning* condition identical to Experiment 1 that served as a baseline. Participants entered their extrapolations as numbers, and extrapolations were displayed sequentially; (b) A *grid* condition where participants drew the function shape by clicking the respective positions on a grid; and (c) A *summary function-learning* condition where participants entered their extrapolations in numbers and all time points were displayed as value-pairs on one screen, so that participants were able to see all previous function values (Figure C.3). To ensure comparability between the three conditions, the maximum and minimum extrapolations were restricted to values $0 - 1550$, in steps of 1; and the clicks on the grid were restricted to the same number as the entries in both function-learning conditions.

## Results

**Dependent variables.** Calculation of **extrapolation accuracy (rRMSE)** in the two function-learning conditions was identical to Experiment 1. We applied the same procedure regarding outliers as in Experiment 1 (0.11% of the total number of extrapolations).

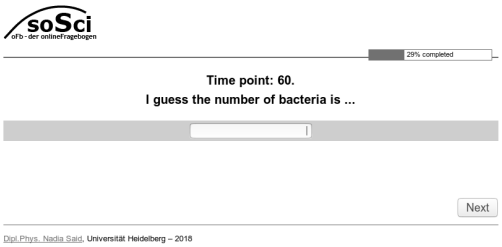
To assess rule-learning in the grid condition, we used two types of approaches:

**(1) Calculating the first derivatives.** Calculating the first derivatives allows us to determine whether participants extrapolated (a) linearly through the last two learning points (exemplar-based extrapolation), or (b) according to the function rule (rule-based extrapolation)(McDaniel et al., 2014). This is because in case of (a), derivatives must be constant, whereas in case of (b), derivatives must be strictly monotonic decreasing as $\frac{d}{dx}(-e^x) = -e^x$ with the negative sign reflecting the trend of the process. This allows us to evaluate whether participants abstracted a rule about the exponentiality of the process in that it is *increasingly declining*. To do so, we used three different approaches varying in strictness of what counts as strictly decreasing derivatives: (1) slopes of the lines through the first and the second, as well as the first and the last extrapolation point were strictly decreasing; (2) slopes through the first and two other extrapolation points have to be strictly decreasing; and (3) slopes though the first and all other extrapolation points have to be strictly decreasing. Please note that other non-linear functions could also meet the condition of strictly decreasing derivatives.

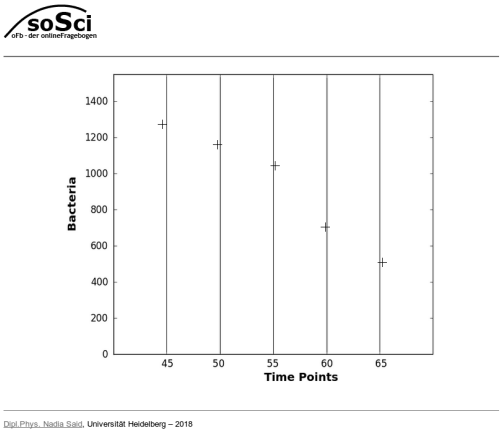
**(2) Least squares approach.** Using a least squares approach, we classified the functions drawn in the grid condition as rule-based, exemplar-based, or non-distinguishable based on the deviation (RMSE) of clicks on the grid from three models: If

$$\text{RMSE}_{\text{linear}} > \text{RMSE}_{\text{exponential}} + \text{RMSE}_{\text{linear}} * 25\%, \tag{C.4}$$

(a)




(b)




(c)

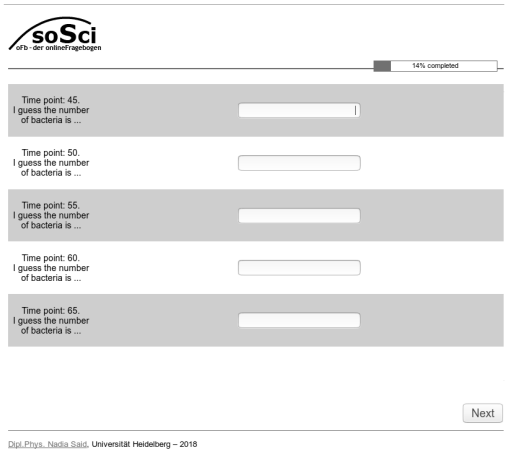



**Figure C.3.** Screenshots of the three conditions. The figure displays the three conditions: (a) standard function learning, (b) grid, and (c) summary function-learning. Valid numbers participants could enter: 0 – 1550.

participants were classified as rule-based learners; if

$$\text{RMSE}_{\text{linear}} < \text{RMSE}_{\text{exponential}} - \text{RMSE}_{\text{linear}} * 25\%, \tag{C.5}$$

participants were classified as exemplar-based learners; for all other cases, participants were classified as non-distinguishable. To ensure comparability between the models, only the slope parameter $a$ was allowed to vary, and all other parameters were fixed.

**Extrapolation accuracy in standard versus summary function-learning.** To assess whether comparatively low proportions of (accurate) rule-application in the classical function-learning paradigm were due to a lack of simultaneous access to all previous function values, we compared extrapolation accuracy in (a) the standard function-learning, with (b) the summary function-learning condition. Results showed that extrapolation accuracy was not higher in the summary ($M_{\text{Ain}_1} = 1.54$, $SD_{\text{Ain}_1} = 0.80$, $M_{\text{Bin}_1} = 0.27$, $SD_{\text{Bin}_1} = 0.19$), compared to the standard function-learning condition ($M_{\text{Ain}_2} = 1.53$, $SD_{\text{Ain}_2} = 0.84$, $M_{\text{Bin}_2} = 0.26$, $SD_{\text{Bin}_2} = 0.19$), $F(2,492) = 0.027$, $p = .97$, Pillais' Trace = 0.0001. This result suggests that having access to all function values did not increase rule application per se, nor did it increase the accuracy of rule-application.

**Proportion of rule-based extrapolation in the two function-learning conditions.** For conditions (a) standard function-learning and (b) summary function-learning, participants' extrapolation styles were classified based on extrapolation accuracy. Table C.5 shows that for both function-learning conditions, between 19% (Ain) and 32% (Bin) were classified as rule-based extrapolators.

**Table C.5.** Proportion of participants classified as showing each of the three extrapolation styles based on extrapolation accuracy, per processes (Ain and Bin).

| | Simple exemplar-based | | Exemplar-based | | Rule-based | |
|---|---|---|---|---|---|---|
| | Standard FL | Summary FL | Standard FL | Summary FL | Standard FL | Summary FL |
| Ain | 148 (60%) | 139 (56%) | 58 (23%) | 58 (24%) | 42 (17%) | 50 (20%) |
| **Total Ain** | 58% | | 23% | | 19% | |
| Bin | 79 (32%) | 72 (29%) | 87 (35%) | 99 (40%) | 82 (33%) | 76 (31%) |
| **Total Bin** | 30% | | 38% | | 32% | |

**Proportion of rule-learning in the grid condition.** *(1) Calculating the first derivatives.* We calculated the first derivatives ($d_{1,2}$, $d_{1,3}$, $d_{1,4}$, $d_{1,5}$) by calculating the slopes between the first extrapolation point and the following 4 points. Participants were classified as having understood the function rule as exponential declining if (a) they captured the trend of the process ($d_{1,j} < 0, \forall\, j \in \{2,3,4,5\}$) and (b) if slopes of the lines through the first and the second, as well as the first and the last extrapolation point were strictly monotonic decreasing ($d_{1,5} < d_{1,2}$). Following that classification, 48% of participants had abstracted the function rule for process Ain, and 56% for process Bin (Figure C.4). We employed the same method for the next stricter criterion (three out of the four slopes), resulting in 28% of participants having abstracted the function rule for process Ain, and 39% of participants having abstracted the function rule for process Bin. Only when using the strictest possible criterion where the values *for all four slopes* have to decrease strictly monotonically, results were comparable to those based on extrapolation accuracy in that 20% of the participants were classified as having abstracted the function rule for process Ain, and 29% for process Bin (Table C.6).

**Table C.6.** Proportion of participants classified as "rule-based learners" in the derivatives approach.

| | $d_{1,5} < d_{1,2}$ | $d_{1,5} < d_{1,3} < d_{1,2}$ or $d_{1,5} < d_{1,4} < d_{1,2}$ | $d_{1,5} < d_{1,4} < d_{1,3} < d_{1,2}$ |
|---|---|---|---|
| Ain | 80 (48%) | 92 (28%) | 33 (20%) |
| Bin | 93 (56%) | 128 (39%) | 48 (29%) |
| and $d_{1,j} < 0, \forall\, j \in \{2,3,4,5\}$ for all four classification criterions. | | | |
| $\chi^2_{\text{Ain}}$ | $\chi^2(1) = 55.87^{***}$ | $\chi^2(1) = 9.33^{**}$ | $\chi^2(1) = 0.08, p = .77$ |
| $\chi^2_{\text{Bin}}$ | $\chi^2(1) = 30.35^{***}$ | $\chi^2(1) = 3.83, p = .05$ | $\chi^2(1) = 0.34, p = .56$ |

Comparison of proportions FL (standard & summary) and derivatives approach, $^{**}$ $p \leq .01$, $^{***}$ $p \leq .001$.

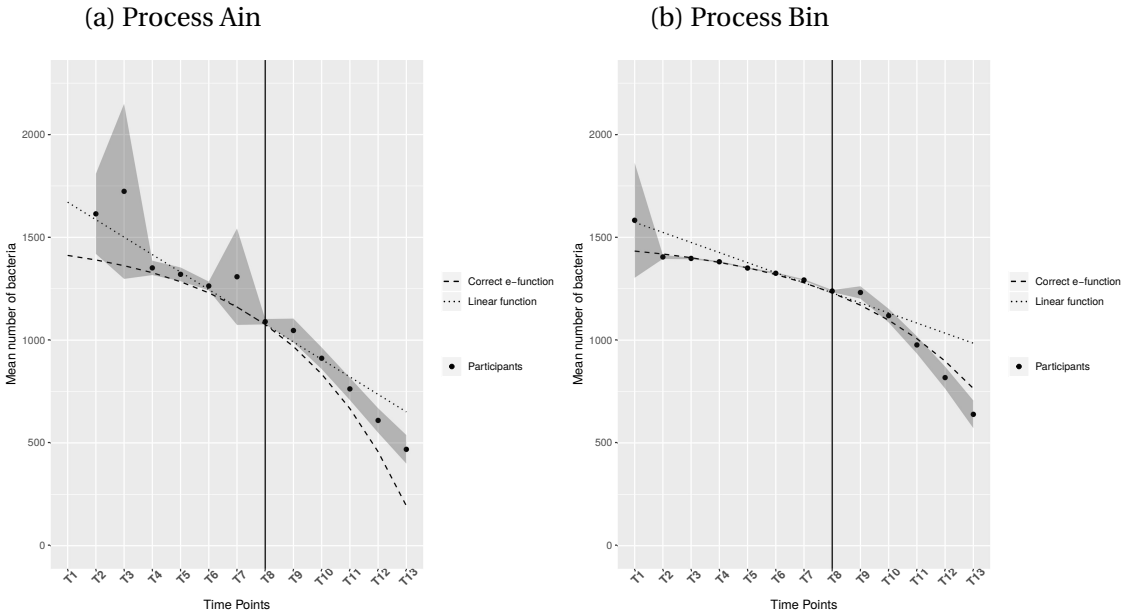



**Figure C.4.** Mean number of bacteria for Ain and Bin for first derivatives approach. The figure displays the mean number of bacteria estimated by participants in the grid condition who were classified as having abstracted a rule about the underlying process for (a) Ain and (b) Bin by calculating the first derivatives.

*(2) Least squares approach.* Using the least squares approach outlined above, 52% of participants were classified as having understood the correct function shape for process Ain, and 59% for process Bin (Figure C.5). These results are broadly in line with results using the derivatives approach.

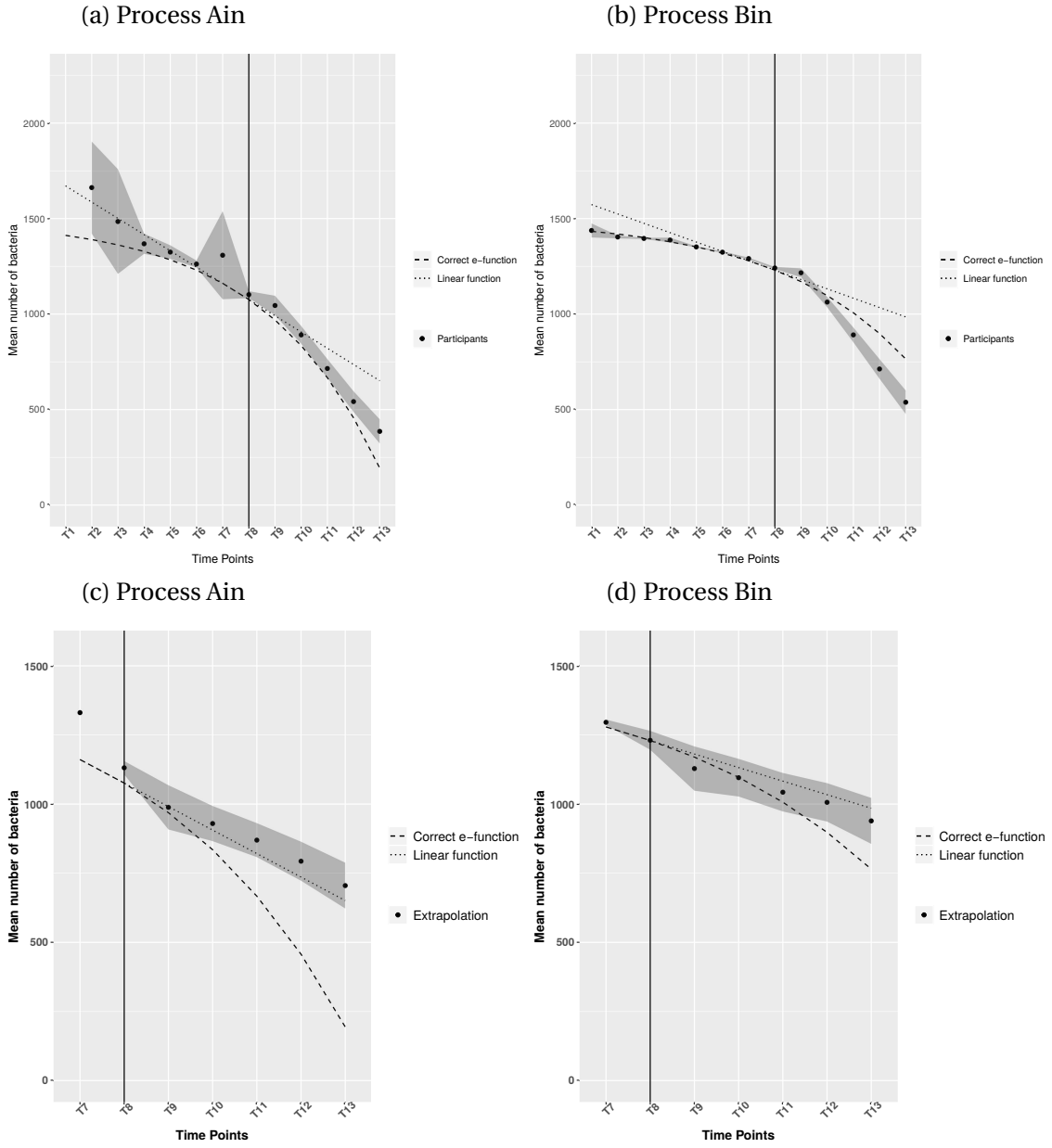



**Figure C.5.** Mean number of bacteria for Ain and Bin for least squares approach. The figure displays the mean number of bacteria estimated by participants in the grid condition who were classified as having abstracted a rule about the underlying process for (a) Ain and (b) Bin applying a least squares approach. Pictures (c) and (d) display participants who were classified as extrapolating linearly.

**Proportion of rule-based learners in all three conditions.** Table C.7 shows the proportion of participants classified as having acquired an understanding of the function rule for both processes, per condition. Across the two function-learning conditions, standard and summary, 19% of participants were classified as rule-based learners for process Ain and 32% for process Bin. In contrast, in the grid condition 50% of partici-

pants were classified as rule-based learners for Ain and 57% for Bin.

**Table C.7.** Proportion of participants classified as "rule-based learners" in all three conditions.

| | Standard FL | Summary FL | (a) Total: Standard & Summary | Derivatives | Least-Squares | (b) Mean: Derivatives & Least-Squares | Comparison of proportions of (a) & (b) |
|---|---|---|---|---|---|---|---|
| Ain | 42 (17%) | 50 (20%) | 92 (**19**%) | 80 (48%) | 86 (52%) | 166 (**50**%) | $\chi^2(1) = 91.20^{***}$ |
| Bin | 82 (33%) | 76 (31%) | 158 (**32**%) | 93 (56%) | 97 (59%) | 190 (**57**%) | $\chi^2(1) = 74.28^{***}$ |

Comparison of proportions (a) FL and (b) grid condition, $^{***}$ $p \leq .001$.

### Summary 2

Experiment 2 investigated whether unequal a priori probabilities to guess correctly could explain increased rule-learning compared to rule application in the standard function-learning paradigm. To do so, we introduced a condition where participants drew their understanding of the progress of the function into a grid where the numbers of clicks as well as the range of possible values were restricted to the same values as extrapolations in the standard function-learning task. The function shapes were evaluated using two different types of approaches, calculating the first derivatives, and a least squares approach of varying strictness. Both methods produced broadly similar results in that 50% of participants were classified as rule-based learners for process Ain and 57% for processes Bin, compared to 19% of participants showing rule-based extrapolation in both function-learning conditions for process Ain, and 32% for process Bin.

To furthermore investigate whether (accurate) rule-application is reduced in the standard function-learning paradigm because participants lack access to all previous function values, we compared extrapolation accuracy in the standard function-learning, with a summary function-learning condition. Results showed that there was no difference in extrapolation accuracy between the two conditions, suggesting that lacking access to all function values did not affect rule-application.

## General Discussion

In the function-learning paradigm, people's understanding of the function rule that underlies the to-be extrapolated process is typically measured by means of extrapolation accuracy (Bott & Heit, 2004; DeLosh et al., 1997; Kwantes & Neal, 2006; Lewandowsky et al., 2002; McDaniel et al., 2014). Here we argue, however, that even though accurate extrapolations necessitate rule-learning, the reverse does not necessarily hold: Inaccurate extrapolations do not exclude rule-learning. Using inaccurate extrapolations to infer learning styles therefore hinges upon the assumption of rule-application given rule-learning. In two function-learning experiments with exponential declining functions, we put this assumption to an experiment test. Results showed that the proportion of participants who demonstrated an understanding of the correct function rule was almost twice as high as the proportion of participants who would be classified as rule-learners based on extrapolation accuracy in the standard function-learning experiment. We therefore conclude that (i) using extrapolation accuracy as a proxy for rule-learning

severely underestimates people's actual ability to abstract the correct function rule; and that (ii) for a substantial proportion of participants, the assumption of rule-application given rule-learning does not hold.

A majority of participants who would be classified as "exemplar-based learners" (61%) or "simple exemplar-based learners" (46%) based on their extrapolation accuracy, were able to identify the correct function shape in the rule-selection paradigm, and 37% of the "exemplar-based learners" were able to identify the correct function shape AND slope. The grid paradigm that ensured equal a priori probabilities between drawing one's understanding of the rule and extrapolations in the classical function-learning paradigm produced broadly similar results: Around half of participants showed an accurate understanding of the function rule, both when analyzing their understanding of the exponentiality of the process via the first derivatives, and via a least squares approach. These results therefore suggest that a substantial proportion of participants did not apply their rule-understanding when extrapolating. In other words, extrapolation accuracy was considerably lower than what would be expected based on participants' understanding of the function rule.

In the grid paradigm, we employed different criteria varying in strictness of what counts as understanding of the function rule as exponential declining. Specifically, in the derivatives approach we varied the number of slopes that had to decrease strictly monotonically. It is important to note that even though strictly decreasing slopes are a characteristic feature of exponential declining functions, other non-linear functions could also meet this condition.

When the first and last slopes, as well as three out of four slopes were strictly monotonically decreasing, the proportion of rule-learners was higher than the proportion of rule-based extrapolators. Only when using the strictest criterion for rule-understanding, that all four slopes be strictly monotonically decreasing, the proportion of participants who were classified as having understood the rule was broadly in line with the proportion of participants who were classified as rule-based learners based on their extrapolation accuracy. This suggests that a considerable proportion of participants who had acquired an understanding of a characteristic feature of the function rule, namely that later extrapolation points should be steeper than earlier extrapolation points, could not implement this understanding when extrapolating.

Contrary to our expectation, extrapolation accuracy was not affected by implementation errors caused by a critical feature of classical function-learning experiments: the successive (as opposed to instantaneous) presentation of function values. Extrapolation accuracy was not higher in an alternative presentation format (summary function learning condition) that provided participants instantaneous access to current, as well as previous function values. This result suggests that while cognitive resources (working memory capacity) may be a limiting factor for rule-induction (McDaniel et al., 2014), they seem to be less relevant for rule-application during extrapolation.

Interestingly, while performance generally dropped as a function of slope, extrapolation accuracy was more strongly affected by function slope (>80% drop for Cin compared to Ain and Bin) compared to learning of the correct function shape (approx. 15% drop for Cin compared to Ain and Bin). Furthermore, learning of the correct function shape AND slope was not influenced by function slope, suggesting that function slope impairs rule learning to a lesser extent than extrapolation accuracy. This result suggests that the well-established tendency toward linear extrapolations (Busemeyer et al., 1997) more strongly reflects a difficulty to extrapolate non-linearly than a more basic difficulty to

recognize non-linear processes as non-linear.

The absolute minority of participants were classified as rule-based learners based on extrapolation accuracy. This pattern holds for both experiments, and for all functions slopes, except for one notable exception: In Experiment 2, about one-third of participants were classified as using rule-based extrapolation. That is, for the function with the less steepest slope, the highest proportion of rule-based extrapolators was reached. In line with this finding, previous research has repeatedly found that people expect linearly increasing function types (Brehmer, 1974; Busemeyer et al., 1997). The generally high proportions of participants who could identify the correct shape and slope of exponentially decreasing functions is therefore particularly telling in the present experiment using exponentially declining functions since these are among the function types with the strongest deviation from participants' expectation of positive linearity.

For participants displaying exemplar-based or simple exemplar-based extrapolations, the relative majorities (61% and 46%, respectively) could identify the correct function shape in the rule-selection task, while considerably smaller proportions of participants (24% and 32%, respectively) believed the trained functions to be actually linear. That is, approximately half of participants extrapolating linearly were well-aware that extrapolations should not in fact be linear. For participants displaying rule-based extrapolations, the pattern was reversed in that the majority believed their extrapolation style to be accurate. These result suggest that participants, up to a certain extent, are aware of what accurate extrapolations should look like, and that around half of exemplar-based and simple exemplar-based extrapolators employed linear extrapolations *despite* their understanding of non-linearity.

It is a common assumption of many function-learning studies that given that participants acquired an understanding of the function rule, they also apply that rule during extrapolation. The present results suggest, however, that a considerable proportion of participants who had acquired an understanding of the accurate function shape, and even slope displayed exemplar-based or even simple exemplar-based extrapolation in the classical function-learning paradigm. We conclude that rule-learning is not tantamount to rule application and that the proportion of rule-based learners in the current function-learning literature likely represents an underestimation.

**Acknowledgments** The authors gratefully acknowledge support by the Excellence Initiative, Institutional Strategy ZUK 5.4 (Scientific Computing in the Social and Behavioral Sciences), Heidelberg University.

# References


Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38–50.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. *Studies in Cognition. Knowledge, Concepts and Categories*, 408–437.

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*(2), i–144.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968.

Fischer, H., & Holt, D. V. (2016). When high working memory capacity is and is not beneficial for predicting nonlinear processes. *Memory & Cognition*, *45*(3), 404-412.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072–1099.

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1019–1030.

Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General*, *131*(2), 163–193.

MacKinnon, A. J., & Wearing, A. J. (1991). Feedback and the forecasting of exponential change. *Acta Psychologica*, *76*(2), 177–191.

Mcdaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*(1), 24–42.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*(2), 668–693.