TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY

On the Epistemic Functions of Neural Network Models in Cognitive Science

Alice Hein

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung einer

Doktorin der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:	Prof. Dr.-Ing. Martin Buss
Prüfende der Dissertation:	1. Prof. Dr.-Ing. Klaus Diepold
	2. Hon.-Prof. Dr. Sonja Zillner
	3. Prof. Dr.-Ing. Stefan Glasauer

Die Dissertation wurde am 13.06.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 21.10.2024 angenommen.

Acknowledgments

My sincerest thanks to:

My supervisor, Klaus Diepold, for creating the weird and wonderful space that is the LDV and for leading it with such integrity. I will always be tremendously grateful for the years I got to spend here.

Sonja Zillner, for her time, discernment, and trust. And for opening my eyes to a range of futures I would never have considered (or considered possible) before. I could not have asked for a kinder, more constructive mentor.

My lovely colleagues at the LDV, for always being ready to lend a helping hand and for making this place feel like home. To Lukas, for the excellent collaboration. And to the DDI team, Erik and Ronald – working with you was an absolute joy.

Milos and Chris, my supervisors at Konux, whose support and encouragement enabled me to follow my interests and significantly change the course of my life.

Brian and our beautiful little community of Alikes, for all the diversion and comfort over the years – particularly during the pandemic.

Isabel, for helping me learn to let go and find joy in the moment and for many of my most treasured memories.

Linus (whom I will never stop referring to as my "little" brother, no matter how much taller and wiser than me he's become). For freely sharing the topics he loves and being so generous with his time and knowledge. Many of what are now my favorite things in life, I would never have discovered without him.

My parents. For always taking us seriously from the very beginning. For a childhood filled with stories and art and music and explanations. For showing us the world. For being there through every up and down and zig and zag.

Abstract

This thesis investigates the potential of neural networks (NNs) as tools for advancing our understanding of cognition. In recent years, NNs have achieved remarkable performance on a wide range of tasks, prompting renewed interest from cognitive scientists. Many believe that NNs can shed light on the neurological processes of cognition. However, critics argue that NNs lack a theoretical basis, biological plausibility, and the ability to provide explanations. Motivated by this discourse, the thesis addresses two overarching questions: How valid are each of the criticisms against the use of NNs in the cognitive sciences, and what are the broader implications for the epistemic utility of NN models? To answer these questions, the thesis presents six case studies that use NNs to explore different aspects of cognition.

The first study presents a Transformer model for detecting relapses in patients with mental disorders. It illustrates how NNs can have pragmatic utility without being based in theory, biologically plausible, or interpretable. The second study proposes a NN incorporating a cognitively inspired selective attention mechanism and evaluates it on a benchmark for grounded language understanding. It shows how NNs can be explicitly connected to theory by taking inspiration from cognitive science. The third study investigates a hypothesis about the emergence of writing systems using a sender-receiver game involving two NNs. This study demonstrates that NNs can serve as working examples of how certain empirical phenomena may arise.

The fourth study trains a Video Transformer to predict the behavior of animated agents in simplified social scenarios and benchmarks its performance against both other NNs and infant behavior. It illustrates that inspecting what a NN learns can prompt questions about cognitively plausible behavior. The fifth study employs another multimodal Transformer, this time trained on number-related tasks inspired by developmental psychology. It conducts a detailed analysis of the model's behavior and learned representations. This study shows how NNs can serve as exploratory models when comprehensive theories are lacking. The sixth study delves into explicit knowledge and reasoning, presenting an approach to modeling decision-making processes in medical ethics using a fuzzy cognitive map. It exemplifies how models like NNs can help externalize thought processes and build intuitive understanding.

Each case study constitutes an independent scientific contribution. Taken together, the six studies demonstrate that the criticisms against using NNs as models in cognitive science are valid to varying degrees but do not fundamentally undermine their epistemic utility. While NNs often lack a direct connection to theory and biological plausibility, they can be explicitly linked to cognitive science literature through careful design of training environments and architectures. Targeted analyses can often explain the complex dynamics of NNs at different levels, from high-level model comparisons to low-level mechanistic accounts. Even when serving primarily as predictive models, NNs can prompt new questions, provide proof-of-principle demonstrations, and reveal relevant factors for future targeted studies. Ultimately, their flexibility, scalability, and amenability to analysis make NNs a powerful extension of the cognitive science toolkit, enabling the investigation of phenomena that were previously inaccessible to empirical inquiry.

Contents

Acronyms

Artificial Intelligence (AI) the field of developing intelligent machines and computer systems

Baby Intuitions Benchmark (BIB) a dataset to evaluate machines' ability to reason about agents' goals and actions like human infants

Behavioral Cloning (BC) an imitation learning technique where a model is trained to mimic expert demonstrations

Binary Cross-Entropy (BCE) a loss function that measures the dissimilarity between predicted probabilities and true binary labels

Convolutional Neural Network (CNN) a type of deep neural network commonly applied to analyzing visual data

Covariance Matrix Adaptation Evolution Strategy (CMA-ES) an evolutionary algorithm for non-linear optimization problems

Deep Learning (DL) a subfield of machine learning based on artificial neural networks with multiple layers

Echo State Network (ESN) a recurrent neural network with a randomly initialized and fixed sparse reservoir

Epistemically Relevant Elements (ERE) key pieces of information relevant to a task or query

Evolutionary Algorithm (EA) a type of optimization algorithm inspired by natural selection

Explainable Artificial Intelligence (XAI) a field of artificial intelligence focused on making AI models and their outputs more comprehensible

Fuzzy Cognitive Map (FCM) a graph structure representing concepts and their causal influences on each other

Grounded Simplified version of the CommAI Navigation tasks (gSCAN) a benchmark for instruction following in 2D environments

Hierarchically Bayesian Theory of Mind (HBToM) a computational model of intuitive psychology

Histogram of Oriented Gradients (HOG) a feature descriptor used in computer vision

Inductive Logic Programming (ILP) a subfield combining machine learning and logic programming

Large Language Model (LLM) a deep neural network trained on vast data to generate human-like text

Long Short-Term Memory (LSTM) a recurrent neural network that employs gated memory cells to capture dependencies in sequential data

Machine Learning (ML) the study of algorithms and models that learn from data

Mean Absolute Error (MAE) a metric measuring the average magnitude of errors between predictions and ground truth

Mean Squared Error (MSE) a metric measuring the average squared difference between predictions and ground truth

Multi-Layer Perceptron (MLP) a class of feedforward artificial neural network

Natural Language Processing (NLP) a field of artificial intelligence focused on enabling computers to understand and generate human language

Neural Network (NN) a class of machine learning algorithms that use a network of artificial neurons to process data and make predictions or decisions

Recurrent Neural Network (RNN) a type of neural network designed to process sequential data

Reinforcement Learning (RL) a type of machine learning where an agent learns to make decisions through trial-and-error

Root Mean Squared Error (RMSE) a metric measuring the square root of the average squared errors

Video Transformer (VT) a deep learning architecture designed for video processing tasks

Violation of Expectation (VoE) a paradigm used to study infant cognition by measuring responses to unexpected events

1 Introduction

The terminology of Artificial Intelligence (AI), particularly within Deep Learning (DL), is replete with metaphors drawn from the realm of brain sciences. Connectionist models are "neural networks" made up of "neurons" and "synapses." They "learn" through backpropagation. Long Short-Term Memory (LSTM) cells "remember" and "forget," and Transformers "attend" over their inputs. Indeed, many of the early contributions to the field can be traced back to researchers interested in the biological mind: Warren McCulloch, Donald Hebb, Frank Rosenblatt, David Rumelhart, James McClelland, and Geoffrey Hinton were all trained in neuroscience or psychology. Since the early brain-inspired models proposed by these scientists, AI has developed into a discipline in its own right, albeit one more concerned with engineering high-performing systems than with mimicking the human brain. This focus on efficiency and accuracy has yielded tremendously powerful applications. In a 2023 report on the capabilities of GPT-4, OpenAI announced that the model could solve parts of tests like the Bar exam, GRE, or Leetcode [Ope23] – feats that seemed far out of reach only a few years ago.

1.1 Neural networks as models of the mind

As Neural Networks (NNs) have grown in complexity and capabilities in recent years, the cognitive sciences have turned their attention to these models with renewed interest [SNS21]. Many neuroscientists and psychologists consider NNs promising models of neural computation [KMK19; Kri15] and have started incorporating them into their research (see Figure 1.1). They believe that "ongoing advances in DL bring us closer to understanding how cognition and perception may be implemented in the brain – the grand challenge at the core of cognitive neuroscience" [SK20, p. 703]. However, not everyone shares this belief, and the issue has sparked a heated debate in the community [CK19]. In the following, I will briefly outline the main arguments on both sides, starting with the proponents of a new convergence of AI and the cognitive sciences. This faction usually stresses two main points:

NNs exhibit activation patterns that correlate with high-level cortical processing in humans. Models in computational neuroscience before the advent of DL were typically shallow and used hand-crafted features [KMK19; YD16]. These models could account to some extent for lower-level perception but were less successful for later stages of neural processing [KM19]. In contrast to these early models, NNs are not designed to fit neural data. Instead, they are usually optimized on large datasets to perform downstream tasks like categorization [YD16]. Yet, several studies have found that NNs outperform other models in accounting for human judgments of perceptual similarity [KBO16; PAG17] and for neural activity while processing images [KK14; Cic+16; HK17; Yam+14; Eic+17; Cad+19; GV15] or language [CGK22; Sch+21]. Given that NNs can handle inputs and tasks of higher complexity than traditional models of perception, they are seen as a way to expand the scope of cognitive computational modeling [MP20]. As Storrs and Kriegeskorte put it, "[...] they are the closest we have yet come to explicit end-to-end models of how perception and cognition might be performed in brains" [SK20, p. 711].

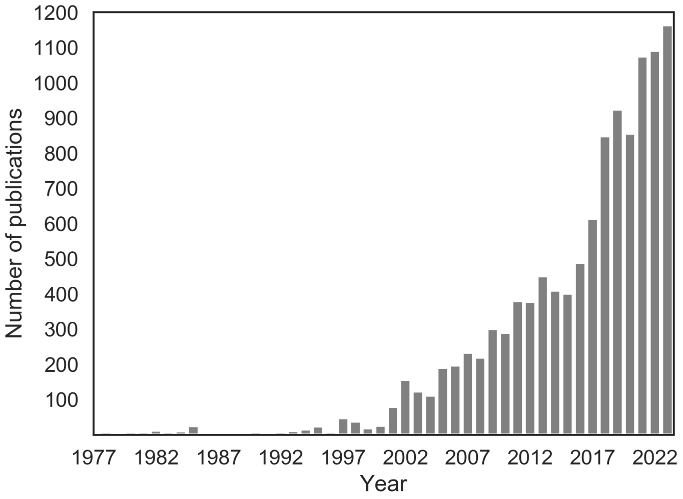


Figure 1.1: Number of papers with AI-related keywords in cognitive science venues between 1975 and 2023. Data retrieved from Scopus by searching for publications in journals and conferences containing "cognitive" or "cognition" in the title. Search results filtered by the keywords "artificial intelligence," "machine learning," "learning systems," and "neural networks."

NNs allow for unprecedented experimental access. The second argument in favor of NNs is that they offer scientists a higher level of experimental control and access to information than is available to them with *in vivo* studies [BMM19]. Researchers can easily record digital neuron activations, network weights, stimuli, learning trajectories, or gradients. They can then analyze the receptive fields of neurons or cross-correlations between activations. They can also perform ablation studies or systematically observe the effect of input perturbations [BMM19]. If one considers NNs idealized stand-ins for biological brains, these models allow for the fast and relatively cheap prototyping of new analysis methods [Kri15; CLS18]. Given the possibilities to "look under the hood" of an AI model, there is also the hope that the investigation of NNs may lead to valuable insights for cognitive neuroscience [MP20].

On the other side of the divide, opponents of NNs as cognitive models put forth three main criticisms, namely [CK19]:

NNs have no basis in theory. In contrast to early cognitive computational models, NNs were not designed to test hypotheses about biological brains [CK19]. Design decisions thus tend to be informed more by heuristics and engineering goals than by pre-existing theories. As a result, it is not possible to map components of NNs to corresponding neural circuits or brain areas [PP20]. Some argue that this theoretical disconnect renders NNs ill-suited for formulating new hypotheses about brain functions and irrelevant to cognitive science. Merely using brain-inspired terminology is insufficient to establish a theoretical link between the disciplines [PP20]. According to these critics, the enthusiasm surrounding DL has led some researchers to disregard existing knowledge of neural computation in favor of tinkering with NNs to elicit more human or animal-like behavior. However, such endeavors are deemed suspiciously akin to engineering-oriented AI research, offering little promise of yielding scientifically valuable insights [SNS21]. Critics contend that rather than pursuing superficial behavioral similarities between NNs and humans, efforts should be directed toward theories of how the mind works [Hom20].

NNs are not biologically plausible. Contrary to what terms like "neurons" or "neural networks" might suggest, NNs lack most of the dynamics of "wetware". Even proponents concede that NNs can, at best, be considered highly abstracted models of the brain in that they consist of simple units that compute linear combinations of their inputs, which are then passed through some kind of nonlinearity [KMK19]. However, most NNs do not produce spike-based representations; they have vastly different constraints regarding power efficiency and memory compared to biological brains, and the standard optimization algorithm in DL, namely backpropagation, is widely considered biologically implausible [BMM19]. On a behavioral level, humans and machines exhibit strikingly different error patterns across many tasks, suggesting divergent underlying representations [SNS21]. In response to the argument presented above that NNs can account for neural activations, critics have argued that the equivalence between brains and AI models has been overstated. They contend that the observed correlations could exist even for activation patterns with vastly different sparsity and dimensionality [SNS21].

NNs do not offer explanations. The third criticism is that trying to use NNs to understand the brain merely substitutes one black box for another [MP20]. AI models may fit neural or behavioral data, but they fall short of arguably one of the most important goals of science: providing explanations [Gel19; Kay18; SK20]. Cognition researchers are interested in the capacities of biological agents at different levels, from perception to learning to social dynamics. Providing a satisfying account of these capacities would mean concisely expressing how they emerge from the interplay of basic elements [Hom20]. Given the complexity and dynamic nature of NNs, they are notoriously difficult to interpret [MP20], and tracing the contribution of components to a specific behavior is a challenge [CK19]. Even if it were possible to garner good explanations from NNs, these might not translate to biological systems. Although AI models have shown behavioral similarities to humans in some regards, an infinite number of systems could produce such behavior using very different processing strategies [KM19]. There are thus two facets to the explainability problem: Understanding the models *themselves* and understanding the target phenomenon, namely biological brains, *through* these models.

1.2 Guiding questions and approach

As outlined above, proponents argue that NNs can account for higher-level cortical processes in humans and offer unprecedented experimental access. On the other side, critics contend that NNs are not informed by theory, are not biologically realistic, and cannot offer explanations. These arguments warrant a closer examination, as they call into question the epistemic utility of such models. In this thesis, I will investigate both the merits of the criticisms against NNs and the potential benefits of employing these models in the study of cognition. To paraphrase a famous quote by statistician George Box: "No model is perfect, but some are useful". Can NNs be useful to the cognitive sciences, and if so, in what way? To sum up, the guiding question for this work as an integrated thesis is the following:

The use of NNs in the study of cognition has been questioned due to these models' lack of theoretical basis, biological implausibility, and inability to provide explanations. How valid are each of these criticisms, and what are the implications for the epistemic utility of NN models?

The study of cognition is, of course, a vast field that encompasses many sub-disciplines. To cover each of them in depth would be beyond the scope of this thesis. Instead, I will approach the guiding question by discussing six models I built to address problem statements

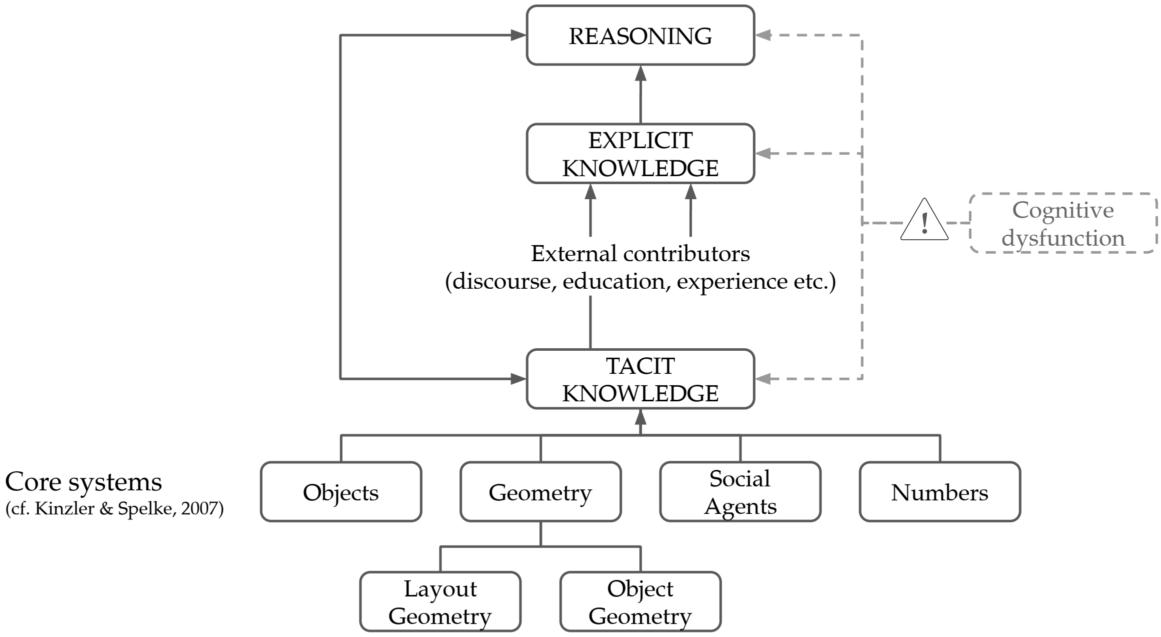


Figure 1.2: Overview of the interrelationships between "core cognition" domains, explicit knowledge and reasoning, and cognitive dysfunction. Adapted from Hast [Has14].

in different, representative areas of cognition research. The choice of problem statements was largely inspired by Kinzler and Spelke's well-known "core cognition" framework. According to their proposal, humans – and, to some extent, animals – have an intuitive understanding of geometry, physical objects, numbers, and the behavior of social agents. These core cognitive domains make up our tacit knowledge. I elaborate more closely on each domain in section 2.2.2.

Core cognition contrasts with the explicit knowledge we gain through socialization and formal education in that it emerges much earlier and is usually not taught or articulated. However, both types of knowledge inform each other and later influence our reasoning. The development and functioning of the cognitive capabilities outlined above may be impacted by factors such as developmental disorders, mental illnesses, or brain lesions. Such conditions constitute an essential part of the study of cognition, as they provide valuable insights into typical and atypical cognitive mechanisms.

In this thesis, I will present a model that relates to each of the aspects of cognition outlined in the previous paragraph (visually summarized in Figure 1.2). For every model, I will then discuss

- How the model relates to the three criticisms against the use of NNs in the study of cognition
- What kind of epistemic value the model has (if any): What type of knowledge can be gained from it?

In the first case study, I present a Transformer model aimed at detecting psychotic and non-psychotic relapses in patients with mental disorders. The second case study focuses on the core systems of physical objects and layout geometry, where I propose a hybrid model incorporating a cognitively inspired selective attention mechanism. I evaluate this model on a benchmark for systematic generalization in grounded language understanding. The

Table 1.1: Relation between overall guiding questions, sub-areas of cognition research (see Figure 1.2), and the modeling case studies presented in the chapters of this thesis.

Meta-questions	Area of cognition		Chapter
• How does the model relate to the criticisms against using NNs for studying cognition? • What kind of epistemic value does the model have, if any?	Cognitive dysfunction		Ch. 3
	Core systems	Objects	Ch. 4
		Geometry	Ch. 5
		Social agents	Ch. 6
		Number	Ch. 7
	Explicit knowledge and reasoning		Ch. 8

third case study explores the cognitive domain of object geometry through a sender-receiver game setup. Here, I use pre-trained vision and speech models to develop artificial writing systems whose geometric regularities I compare against human-made letters. In the fourth case study, I turn to the cognitive domain of social agents. I train a Video Transformer (VT) to predict the behavior of animated characters in simplified social scenarios and benchmark its performance against other NNs and infant behavior. The fifth case study employs another multimodal Transformer, this time trained on number-related tasks inspired by educational and developmental psychology. I then conduct a detailed analysis of the model's behavior and learned representations. Finally, the sixth case study delves into explicit knowledge and reasoning, presenting an approach to modeling decision-making processes in medical ethics using a fuzzy cognitive map.

Each study represents an independent and rather different contribution to the Machine Learning (ML) literature, each of which was or will be published on its own merit. The technical portions of each work occupy the first halves of their respective chapters and can be read independently. In the second half of each chapter, I examine the preceding case study from an epistemic perspective, introducing epistemic concepts as needed to illuminate the relation of the respective model to the overarching guiding questions. Although epistemology is typically associated with the study of scientific processes in natural sciences, I believe that such contextualizations are exceedingly important in empirical fields like ML and cognitive science as well. In both disciplines, the epistemic goals of a study often remain implicit [McC09], and the kind of knowledge we can (and cannot) derive from it is rarely discussed. I, therefore, believe that epistemology can fruitfully inform modeling endeavors in these areas of research, and I use it as a unifying lens throughout this thesis.

The six models will serve as representative illustrations of epistemic functions that NNs can fulfill in the cognitive sciences. Table 1.1 shows how each case study relates to the guiding questions of this thesis and the different sub-areas of cognition research. The more targeted research questions that motivated the independent study of each model are addressed within their corresponding chapters. The framing chapters 1, 2, and 9 focus on the overall contextualization of the individual contributions.

Regarding the use of personal pronouns, for the case studies themselves and the discussions of the first and sixth case studies, which are based on collaborative publications, I will use "we/our/us"; otherwise, I will use "I/me/mine" throughout the thesis.

2 Background

2.1 The epistemology of models

Models join with measurement devices, experiments, theories, and data as one of the key components of scientific practice [MM99]. Many important contributions across disciplines have taken the form of models: Bohr's model of the atom, agent-based models in economics, the DNA double helix model, the Lotka-Volterra model of predator-prey dynamics – the list could go on [Gel16a]. Given the de facto pervasiveness of models, it may come as a surprise that philosophers of science first began to turn their attention to the topic in earnest in the 1980s [Imb17]. Before this, their focus had mainly been on scientific theories [Gel16a]. Although the epistemological study of models began relatively late, their centrality to science is now widely accepted [Gel19].

Despite this agreement on their importance, there is little consensus about what models are and which roles they should play. Part of the obstacle to a comprehensive characterization may be that, as Nelson Goodman puts it, "Few terms are used in popular and scientific discourse more promiscuously than 'model'" [Goo68, p. 171]. Figure 2.1 illustrates this: models come in a bewildering array of guises. They can be drawings, 3D objects, mathematical equations, software – even living organisms [Knu05]. Given that they form such a heterogeneous ensemble, providing a satisfying unitary account that captures the intrinsic nature of all models has proven difficult [Dié15]. Instead, philosophers of science have resorted to defining models according to their function [Gel16c]. Broadly, we can distinguish between two classes of functional characterization: instantiation and representation [Gel16a].

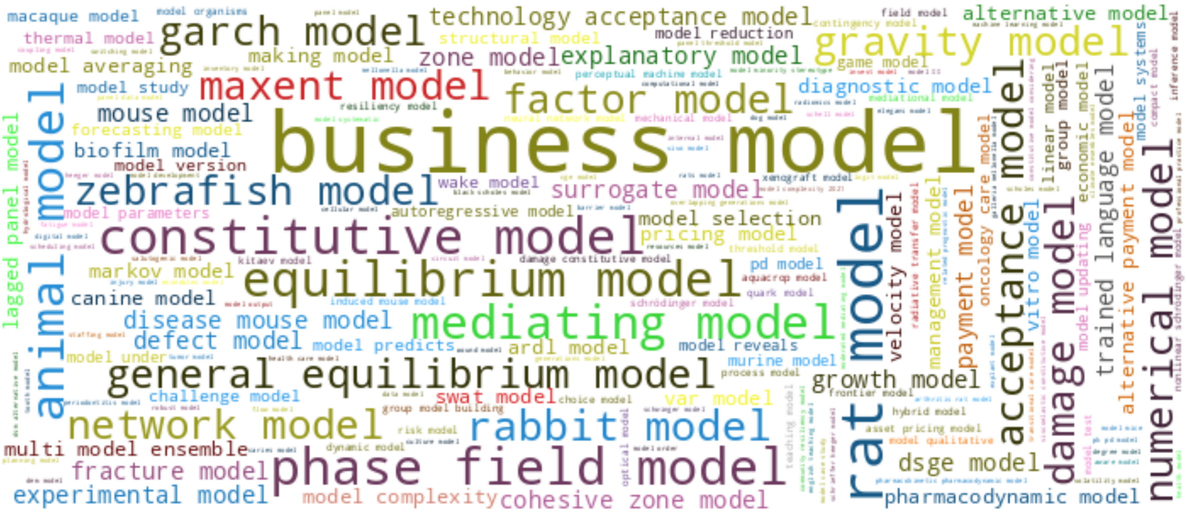


Figure 2.1: Common co-occurrences with the word "model" in the titles of scientific publications, colored by discipline. Data collected from Scopus.

2.1.1 The instantial view

The instantial view regards models as instantiations of a theory [Gel16e]. According to Suppes, a prominent supporter of this conception, "a theory is a linguistic entity consisting of a set of sentences and models are non-linguistic entities in which the theory is satisfied [Sup60, p. 290]." On this account, a model is considered the interpretation of a set of axioms [Gel17; Mor07]. The same set of axioms may allow for different isomorphic interpretations, i.e., models [Gie99]. That is, there will be a one-to-one correspondence between their components [Mor07].

To make this notion more concrete, consider the example of an electric LC circuit and a bouncing spring [Gie99]. Both systems exhibit sinusoidal behavior, which can be described by the same abstract differential equation $a\frac{d^2u}{dt^2} + b\,u = 0$. To calculate the spring's position at any time t, we can substitute its position x for u, its mass m for a, and its constant k for b. To calculate the LC circuit's current at time t, we can substitute its capacitor charge q for u, its inductance L for a, and $\frac{1}{C}$ for b, where C represents capacitance. There are thus direct analogies between the theory, i.e., the differential equation, and its two instantiations, i.e., the spring and the circuit (see Figure 2.2).

As the mechanical and circuit models are sometimes used to teach intuition of electrical circuits, it should be emphasized that the theory being instantiated here is not the lumped electrical circuit but the differential equation. Typically, models constitute simplifications of their target systems, omitting certain details to facilitate understanding or analysis. Interestingly, in this case, the relationship is inverted. The target system – the theory represented by the differential equation – is simpler than its physical instantiations. Specifically, it ignores factors present in real-world systems, such as electromagnetic fields and radiation, high-frequency effects, non-linearities in circuit elements, and resistive losses.

The view of models as instantiations of simple, fundamental theories is attractive in fields like physics, where the pursuit of first-principles explanations is a central goal. However, this approach becomes problematic in domains where such explanations remain elusive. In cognitive science, for instance, academic theories are more modest attempts to bring some initial clarity and organization to our observations and understanding of the mind. The way cognitive science divides cognition into different domains, such as perception, attention, and memory, is somewhat artificial, though, as these categories tend to overlap in reality. This lack of clear separability reflects the absence of overarching theories that are both parsimonious and have full explanatory power. As of yet, cognitive science does not afford the type of elegant, powerful theories that exist in physics and instead has to make do with more piecemeal explanations that do not perfectly capture the integrated nature of cognition. I contend that precisely this state of affairs makes it important to scrutinize the goals and limitations of modeling in cognitive science.

2.1.2 The representational view

The instantial account has come under attack from several fronts. Most recent philosophers of science consider models as tools of representation rather than as embodiments of abstract theories [Knu05]. According to this representational view, models give us knowledge because they capture at least some important aspects of their target systems [Knu21]. In this conception of models, the focus of analysis shifts from the relationship between theory and model to the relationship between model and target [Knu11]. However, scholars differ in their opinion of what this model-target relationship should look like. Within the representational view, we can broadly distinguish between informational and pragmatic views [Gel17].

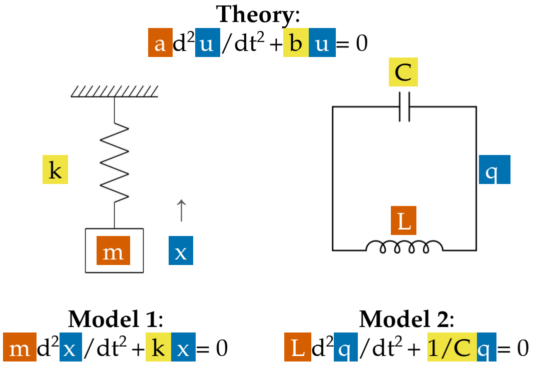


Figure 2.2: Visual representation of the example used to illustrate the instantial view of models in section 2.1.1. One-to-one correspondences between components of the theory (top), the spring model (left), and the LC circuit model (right) marked with the same colors. Adapted from Giere (1999).

The informational view For those who emphasize the informational aspects of models, representation implies an objective relation between the model and its target [Gel16a]. This relation is independent of the model user's beliefs or intentions, i.e., the focus is solely on the model-target dyad [Gel17]. The model's epistemic value stems from the fact that it contains information about the systems it depicts: it is similar to the target in important ways [Gie99]. According to particularly strong versions of the informational view, this similarity should take the form of some kind of morphism [Knu11; Gel17]. There should be a direct mapping between model and target elements that preserves the relations between components [Knu05]. If a model accurately represents its target, it can "stand in" for the other system [Knu11]. Conclusions drawn based on the model will transfer to the real-world target it depicts [Fen67]. Although attractive in theory, this view has been criticized for failing to take into account an essential part of representational practices, namely, the model user.

The pragmatic view Pragmatic views of models shift the focus of analysis from the model-target dyad to a model-target-user triad [Knu05]. Pragmatists maintain that the view of representation as a faithful structural mapping of the elements of a system is too narrow [Knu05; Knu11; RK22]. Instead, models function as representations in virtue of the cognitive activities for which a user employs them [Gel16a]. Essentially, a model is considered a representation if it is used as such [RK22]. Therefore, one needs to take into account the user's cognitive interests, beliefs, and intentions [Gel16a]. A user might, for example, want to derive predictions from their model, gain a deeper understanding of a phenomenon, or communicate an idea. According to pragmatic views, a model's success should be judged by how well it serves its user's epistemic goals – not necessarily by its ability to mirror some target [MM99; GP17].

In the extreme, this view risks resulting in epistemic myopia. A model's success should not be judged solely by its ability to satisfy a user's subjective goals, as users may harbor misconceptions or misjudge whether a model truly serves its purpose. Within scientific disciplines, models typically need to demonstrate some level of empirical grounding or theoretical coherence in addition to serving a specific, pragmatic function. Peer-reviewed publications rarely accept models that exhibit no correspondence with relevant features of a target phenomenon. However, the aspects of reality that a model is expected to capture, and the extent to which it must do so, depend on the specific questions a model is designed to address. Ultimately, the pragmatic perspective cautions against focusing solely on representational faithfulness without considering the context of a model's intended use.

2.1.3 Relation to the guiding questions

NN skeptics in the cognitive science community rarely express their concerns in epistemological terms. Yet, there are striking parallels between their statements and the different schools of thought on models in the philosophy of science.

Consider the first criticism against NNs: They have no basis in theory. This objection mirrors the instantial view that models are, or should be, interpretations of a set of axioms. The second criticism, namely that NNs are biologically implausible, points to an underlying assumption that the value of a model hinges on the similarity to its target – in this case, the brain. An ideal model would constitute a detailed cortical reconstruction à la "Blue Brain Project" [Mar06], allowing for accurate simulations. This perspective aligns with the informational view of models. It also relates to the instantial view in that the brain can be seen as an instantiation of certain physical principles and biological mechanisms (e.g., differential equations governing neuron spiking, synaptic plasticity rules, etc.). From the instantial perspective, for NNs to be considered valid models of the brain, there should be a one-to-one correspondence between the components of NNs and the underlying theory that the brain instantiates.

The third criticism – NNs do not offer explanations – implies that cognitive scientists have a goal, namely understanding biological cognition, which NNs do not meet. This critique reflects a pragmatic view: A model's purpose is to support its user's epistemic aims, and failing to do so invalidates its legitimacy. The pragmatic perspective also encompasses the other two criticisms, as some cognitive scientists may aim for models that accurately represent theoretical frameworks or faithfully mirror the biological brain.

In summary, the recent debate surrounding the use of NNs in the study of cognition mirrors an older discourse about the nature of models. It stands to reason that the epistemological literature on the topic can fruitfully inform the discussion of NNs in cognitive science. After each of the six modeling case studies in this thesis, I will include a section that briefly presents some contributions from the philosophy of science related to the case study in question. These "epistemic background" sections will provide the basis for my evaluation of the epistemic functions of each model.

2.2 The study of cognition

The framing chapters of this thesis address the usefulness of NNs in the study of cognition, aligning with the pragmatic view of models discussed previously. As we have seen, pragmatic accounts maintain that a model's evaluation must consider its user's epistemic aims. In the following section, I will, therefore, attempt to define the scope and goals of cognitive science – in as much as this is possible, given the inherent breadth and interdisciplinary nature of the field. I will also provide an overview of Kinzler and Spelke's ideas on "core cognition", as this framework greatly informed my choice of modeling case studies.

2.2.1 History and definition

Cognitive science has its origins in 1950s North America [Núñ+19]. At this time, psychology was moving away from behaviorism, the previously dominant paradigm [Mil03]. Behaviorism, often associated with researchers like Watson and Skinner, focused on the associations between stimuli and outputs. Whatever the "black box" in the middle did to produce a behavior was considered unobservable and thus of little scientific relevance. However, the emergence of the fields of computer science and neuroscience brought with it a growing interest in information processing mechanisms, both digital and biological [Mil03].

The term "cognitive science" was first used by Christopher Longuet-Higgins in a 1973 commentary on the Lighthouse report. This report, commissioned by the UK parliament, was a scathing review of the state of AI at the time. Its author, Sir James Lighthill, concluded that most AI methods only applied to toy problems and would not scale up to the real world. The report is credited with almost completely dismantling AI research in the UK for about ten years [RN10]. In his rebuttal to this report, Longuet-Higgins argued that the principal value of AI was to allow for formulating and testing models in sciences relevant to human thought and perception: the cognitive sciences [Lon73].

From the very beginning, cognitive science was conceived as an interdisciplinary venture. Longuet-Higgins's commentary suggested four main groupings within cognition research: mathematics, linguistics, psychology, and physiology [Lon73]. This proposal has since been revised, and the most salient conceptualization of the field is now one as the product of synergies between psychology, linguistics, AI, anthropology, philosophy, and neuroscience [Núñ+19]. The Cognitive Science Society, a flagship institution of the field since 1979, also includes education in this list [Soc]. The main disciplines are often depicted as a regular hexagon, which continues to serve as the emblem of cognitive science today (see Figure 2.3) [Núñ+19].

What is it, then, that connects these many disciplines? Because cognitive science spans so many research areas, it is difficult to pinpoint a cohesive core of the field. Here, I will adopt a definition proposed by Mekik and Galang, which I feel best captures the center of the "cognitive science hexagon." Mekik and Galang propose to view cognitive science as "the study of how agents perform tasks" [MG22, p. 2]. A task here refers to anything that has to be done – from daily activities to experimental tasks to computational problems. Mekik and Galang identify this task-oriented focus as a distinctive feature of cognitive science.

Researchers in this field are typically concerned with operationalizing hypotheses as tasks that allow for empirical performance measurements. For example, they may study how people learn and remember new information by testing their recall on word lists or visual patterns, or examine decision-making processes by having participants play economic games or solve logic puzzles. Another common research goal in cognitive science is to characterize the effect of agent and task constraints on internal or external behavior, such as how working memory capacity influences problem-solving ability or how time pressure impacts decision quality.

In Mekik and Galang's interpretation of cognitive science, an agent is understood to be any entity capable of action, covering biological and AI agents alike. Different sub-disciplines may focus on specific instances of agents and tasks at different levels of analysis. For example, neuroscientists may study the neuronal dynamics of human or animal brains during cognitive tasks, while anthropologists may examine the behaviors and decision-making processes of entire cultures over long timescales. Cognitive science forms a meta-discipline that draws on these different areas of research to investigate more abstract and general questions [MG22].

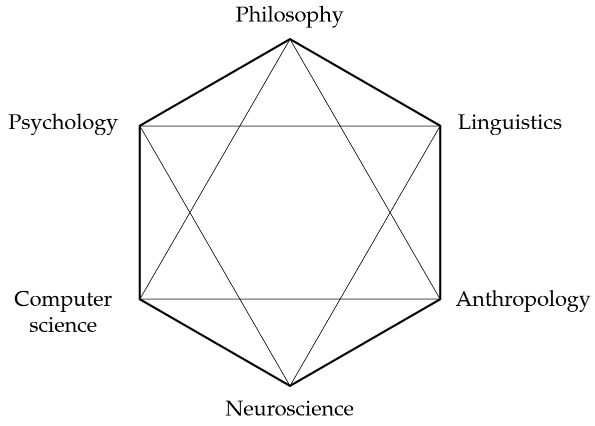


Figure 2.3: “Cognitive science hexagon” representing the six main disciplines of the field.

2.2.2 Core cognition

According to Spelke and her collaborators, humans are endowed with a few specialized systems that form the foundation for higher-level concepts and skills [Spe22]. Spelke’s view on these systems’ exact number and nature appears to have changed over time. In this thesis, I focus on four innate cognitive primitives that have consistently been part of her proposal: 1) intuitions about inanimate objects and their interactions, 2) geometry, 3) agents and goal-directed actions, and 4) numbers. Spelke argues that these systems have been “hardwired” into our brains in the course of cognitive evolution [HM19]. Drawing on empirical investigations from developmental psychology and neuroscience, she shows that core cognitive abilities are present in infants and animals too early to be a product of learning [Spe22]. Each system is characterized by a set of signature limits allowing its identification across tasks, ages, species, and human cultures [SK07]. Spelke, therefore, suggests that these domains form a set of universal building blocks of cognition [HS04].

Objects The first core system that Spelke proposes represents physical objects and the principles that govern their motion [SK07]. Objects move as bounded, cohesive entities on continuous, unobstructed paths. They only influence each other’s motions if they come into contact with each other. Knowing this allows humans and animals to predict object trajectories and to perceive shapes and boundaries, even when objects are partly occluded [SV93]. Evidence of these abilities has been found in newborn infants and baby chicks [LSR96; RV95; Val+06]. Crucially, the object system is constrained: our attentional resources only seem to allow us to represent and track about three or four objects at a time [SK07]. This signature set-size limit has been identified in humans across various ages and cultures [RFJ01; ROL03; Gor04]. Our limitations on object-based attention and visual working memory thus appear to be present almost from birth and to stay relatively constant throughout our lives [Car00].

Geometry Spelke’s second core system represents geometry. It encompasses two distinct subsystems: one for layout geometry and one for object geometry [SL12]. Layout geometry supports spatial navigation. It encodes egocentric distance (proximal-distal) and sense (left-right), which enables inferences about the relative positions of items and places in one’s environment [HM19]. This ability has been found in diverse animals, children, and adults [SL12]. Object geometry supports the recognition of smaller objects and their visual properties. Studies have shown that, throughout the world, people are highly sensitive to shape and geometric regularities such as symmetry [Sab+21; Deh+06]. Spelke's proposal of two core systems for geometry converges with Goodale and Milner's two-streams hypothesis [HM19]. According to this view, object recognition and localization are handled by two separate pathways in the primate brain, namely the ventral stream (or "what?" system) and the dorsal stream (or "where?" system) [GM92].

Social agents The third core system represents agents and their actions. Studies have shown that young infants expect agents to behave differently than passive, inanimate objects. Unlike objects, agents need not be cohesive or move along continuous paths [SPW95]. Agents can cause their own motion and can interact with other objects or agents [SK07]. From a very young age, infants represent agents' actions as intentional, and they expect agents to move towards their goals efficiently, i.e., without unnecessary detours [GC03; Woo99]. Infants use their representations of agents to guide their own actions and to interpret the actions of others [Spe22]. The agent core system may, therefore, play an essential role in social learning [SK07].

Numbers The fourth core system is that of numbers. What is meant here is not the mathematical concept of numbers but a more fundamental ability to represent and compare magnitudes and quantities. Spelke and her collaborators distinguish between two core systems of numerical representations [FDS04a]. These are present in various animal species, human infants, and adults from indigenous groups who were never exposed to mathematical concepts [Hau+03; XS00; Pic+04a]. The first system allows for precisely representing distinct objects and their continuous properties, such as size or length [FDS04a]. Due to attentional constraints, this system can only keep track of a small number of concrete entities at a time [Lei+17]. The second system can represent much larger numerosities, but only approximately [FDS04a]. It can also compare groups of objects if the difference in number is significant enough [VV82].

Beyond core cognition According to Spelke, core systems occupy an important middle ground in cognitive research: they are distinct from sensory or perceptual systems in that they represent abstract properties and relations. They also differ from the explicit belief systems underlying our decision-making and reasoning because they operate automatically and appear to emerge too early to be a product of learning [Spe22]. However, they may serve as scaffolds for later, more explicit knowledge. E.g., children may build on their core number system to learn mathematics, and their core agent system may be the foundation for more complex moral reasoning. Even after acquiring more advanced concepts, core systems persist throughout our lives and complement, or sometimes compete with, our explicit knowledge [SK07].

Not all researchers concur with Spelke's proposal of core cognition [Car00; Rip17]. There continues to be much debate about the existence and nature of "built-in" knowledge in the brain. My thesis is relatively agnostic to these issues. I take the view that, regardless of questions about origins and exact definitions, the evidence put forward by Spelke points to core systems as important "clusters" of cognitive abilities shared by a range of agents, including animals, children, and adults. They, therefore, provide a useful lens through which I will explore the broader field of cognition in this thesis. With this brief review, I now turn to the first case study.

3 Modeling patient-specific activity patterns with Transformers to detect psychotic and non-psychotic relapses

The sciences do not try to explain, they hardly even try to interpret, they mainly make models.

John von Neumann

3.1 Study

The first case study relates to the research area of cognitive dysfunction, specifically psychotic disorders (see Figure 3.1). The model presented in this study was developed as a submission to the 2nd e-Prevention Grand Challenge hosted at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2024 [Fil+24]. The objective posed in the e-Prevention challenge was to identify psychotic and non-psychotic relapses in patients using biosignals captured by wearable sensors. Our proposed solution is an unsupervised anomaly detection approach based on Transformers. The submission ranked 3rd on detecting non-psychotic relapses (Track 1) and 1st on detecting psychotic relapses (Track 2). Therefore, a version of the following text was included in the 2024 ICASSP proceedings [HGD24]. This case study represents an engineering-driven approach that is typical for the field of ML. It serves mainly as a contrast to the less conventional uses of NNs presented later in this thesis, and its presentation is kept relatively brief.

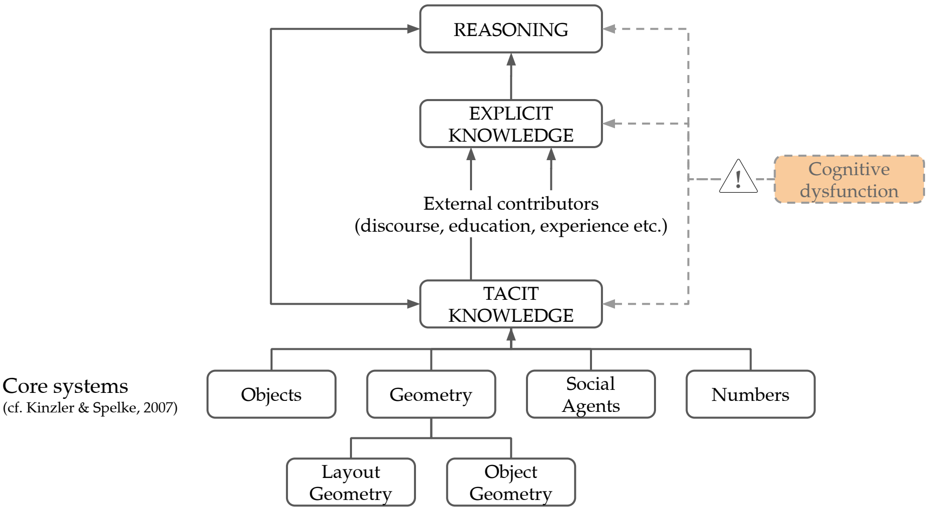


Figure 3.1: Situating the first case study in the broader study of cognition. Relevant parts of the framework marked in orange.

3.1.1 Introduction

Millions of individuals worldwide struggle with psychotic disorders, which significantly impact their quality of life and pose substantial challenges to strained healthcare systems [Wai+17]. The increasingly widespread adoption of wearable sensors, paired with advancements in ML, offers the potential to aid such patients through digital phenotyping [Zla+22]. By continuously and unobtrusively monitoring physiological and behavioral patterns, subtle indicators of impending or ongoing relapses can be detected, thereby facilitating timely interventions and personalized treatment strategies. The e-Prevention project aims to stimulate research in this nascent but promising area [Fil+24]. The first e-Prevention challenge was held in 2023 and included two tasks: identifying smartwatch wearers and detecting psychotic relapses. While submissions scored high on the first task, results on the second task were more mixed. The second e-Prevention challenge, therefore, focuses on relapse detection, this time including both psychotic and non-psychotic relapses.

3.1.2 Dataset

The data for this challenge was sourced from the e-Prevention project [Zla+22]. It contained features extracted from accelerometer, gyroscope, heart rate monitor, sleep, and step data recorded on several contiguous days by 9 patients for Track 1 (non-psychotic relapse detection) and 8 patients for Track 2 (psychotic relapse detection). The primary difference between psychotic and non-psychotic relapses lies in the nature of their symptoms. Psychotic relapses are characterized by a loss of contact with reality, which can manifest as hallucinations or delusions. Non-psychotic relapses instead involve symptoms like a depressed mood, fatigue, or anxiety. Data for both tracks were split into a training set and a validation set by the challenge organizers. The training set contained only non-relapse data, and the validation set included relapse and non-relapse days. A held-out test set only available to the organizers was used for the final evaluation of the challenge. The validation set could not be used for training, i.e., only unsupervised approaches were allowed.

3.1.3 Methods

Pre-Processing We pre-processed the data using a baseline script provided by the organizers[1], which extracted a set of features from the raw sensor data. These included features such as heart rate variability statistics or the norm of the recorded linear acceleration, which had proven helpful in previous work on digital phenotyping [Eft+23]. We added step information as a feature but did not use the provided sleep data as these were quite unreliable and negatively impacted model performance. The dataset contained many missing values due to sensor malfunctions or patients not wearing their watches. For Track 1, we imputed missing values with the features' median value for each 5-minute segment of the day for the individual patients. For Track 2, we omitted segments with missing values.

Models and Training Our implementation built closely on the challenge baseline[1]. This baseline was a Transformer-based model, which the organizers trained to predict a patient's ID based on the data captured by that patient's wearable during the day. Transformers are described in some technical detail in section 7.1.3. The intermediate features from this patient identification model were then used to train an Elliptic Envelope classifier [RD99] to detect anomalies, i.e., relapses [Eft+23]. However, we noticed in our data exploration a significant

[1] https://github.com/filby89/spgc-eprevention-icassp2024

Table 3.1: Experiment settings used for Tracks 1 and 2 of the e-Prevention challenge.

	Track 1	**Track 2**
Handling of Missing Values	Impute median	Discard
Sequence Length	72	24
Encoder Layers	2	2
Encoder Heads	8	8
Output Dimension	64	64
Encoder Dimension	64	64
MLP Ensemble	no	yes
Batch Size	64	16
Dropout Rate	0.1	0.2
Optimizer	Adam	Adam
Epochs	100	50
Early Stopping	yes	yes

variability across patients. We also saw that the daily activity patterns of individual patients tended to change when they relapsed. This observation is consistent with previous findings that changes in routines, especially sleep behavior, can serve as an important feature in detecting relapses [Avr+23; Zho+22]. Therefore, we made two major changes to the baseline.

First, we changed the pre-training objective from predicting user IDs to predicting the time of day that a measurement was taken. The idea behind this was that changes in patient routines would reflect in model prediction errors. The other main change we made to the baseline was that we trained an individual model for each patient. Previous work has also found this to better capture the patients' characteristic activity patterns [Cal+23; Pan+22]. Additionally, we made some minor changes in hyperparameters, which we found empirically to improve performance. Table 3.1 provides an overview of the experimental setups used for both tracks.

There was less data available for Track 2, which increased the risk of overfitting. We, therefore, used an ensemble of Multi-Layer Perceptron (MLP) prediction heads, each of which we trained on an 80 % subset of the training data. This served as a form of regularization and improved performance on the Track 2 validation dataset. Preliminary experiments with ensembles on Track 1 yielded no discernible improvement over the single-model approach. Given the added computational overhead of ensembles, we decided to apply them only to Track 2.

Outlier Detection To detect outliers, i.e., relapses, we computed the mean prediction error for each day in the training and validation sets. We then calculated the mean-normalized error for each day in the validation set

$$e_{\text{norm}} = \frac{e_{\text{val}} - \bar{e}_{\text{train}}}{\max(e_{\text{train}}) - \min(e_{\text{train}})}.$$

A 24-hour period was classified as a relapse day when the mean-normalized prediction error was above zero, i.e.,

$$\text{score}(e_{\text{norm}}) = \begin{cases} 0 & \text{if } e_{\text{norm}} \leq 0 \\ 1 & \text{if } e_{\text{norm}} > 0. \end{cases}$$

Table 3.2: Scores on validation and test set for Tracks 1 and 2 of the e-Prevention challenge. Validation scores are averaged over five independent runs. Our approach ranked 3$^{\text{rd}}$ on Track 1 and 1$^{\text{st}}$ on Track 2.

	Track 1			**Track 2**		
Val	PR-AUC	ROC-AUC	AVG	PR-AUC	ROC-AUC	AVG
Random Chance	0.326	0.500	0.413	0.349	0.500	0.424
Baseline	0.472	0.614	0.543	0.452	0.594	0.522
Ours	**0.680**	**0.665**	**0.672**	**0.694**	**0.669**	**0.681**
Test	PR-AUC	ROC-AUC	AVG	PR-AUC	ROC-AUC	AVG
Random Chance	0.500	0.430	0.465	0.500	0.347	0.424
Baseline	0.561	0.485	0.522	0.548	0.412	0.480
Ours	**0.595**	**0.574**	**0.584**	**0.563**	**0.444**	**0.504**

For Track 2, we used an ensemble consisting of $M = 5$ ensemble members to calculate e_{val} based on the squared Euclidean distance of each prediction p_i to the prediction mean $\mu = \frac{1}{M}\sum_i p_i$, i.e.,

$$e_{\text{val}} = \mathbb{E}_{\text{day}}\left[\mathbb{E}_i\left[\|p_i - \mu\|_2^2\right]\right].$$

3.1.4 Results

Table 3.2 shows the scores on the validation and test sets for Tracks 1 and 2. Fig. 3.2 illustrates the rationale behind our approach to use time stamp prediction as a pre-training task. In the example shown, the model's prediction error is lower on validation set days without relapses. Many relapse days show errors especially after midnight and in the early hours of the morning, indicating a change in activity pattern.

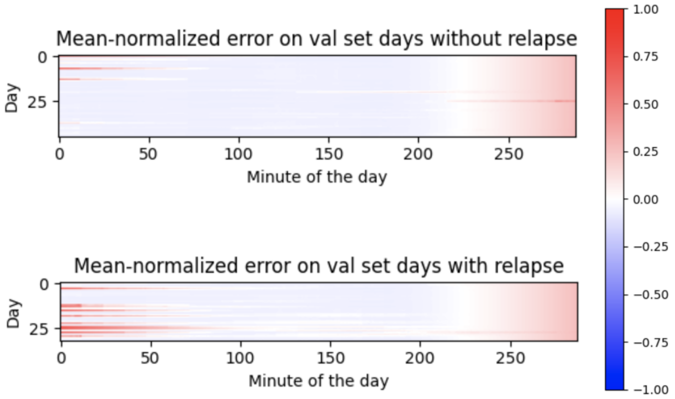


Figure 3.2: Example visualization for the model's prediction error for Patient 5 in Track 1. Shown per 5-minute segment, i.e., there are $288 = 24 \times \frac{60}{5}$ segments per day. Predictions on overlapping segments were averaged.

3.2 Discussion

In summary, we investigated the usefulness of predicting biosignal measurement timestamps as a pre-training task for the unsupervised detection of relapses in psychotic patients. We trained individual Transformer models for each patient to predict the timestamps of biosignal measurements on non-relapse days, implicitly modeling normal daily routines. The models' mean-normalized prediction errors were then used as indicators of atypical behavior and, thus, risk of relapse. We tested this approach in both a single-model (Track 1) and an ensemble set-up (Track 2) and found that it yielded promising results. Our approach ranked 3rd on detecting non-psychotic relapses and 1st on detecting psychotic relapses.

Although this case study relates to psychology through its dataset, it reflects the currently dominant paradigm in ML. The values and goals of the field mainly pertain to performance – i.e., accuracy or computational efficiency. As in the case study, these are the parameters used to justify design decisions: if one approach outperforms another, the reasons are of secondary importance. This attitude is not unique to ML – returning to the epigraph of this chapter, John von Neumann, in his time, already remarked that "The sciences do not try to explain, they hardly even try to interpret, they mainly make models. [...] The justification of such a mathematical construct is solely and precisely that it is expected to work" [Von55, p. 157].

However, this focus on performance is particularly prevalent in ML. What counts is to reach higher accuracies on public benchmark datasets. These serve as easy points of reference and comparison in a highly competitive field, where the number of papers published each month is growing exponentially [Kre+23], and acceptance rates at top-tier conferences have reached lows around 20% [Pin+21]. Challenges, such as the one presented in this case study, are a common mode of engagement. Tuning hyperparameters assiduously is seen as progress for no other reason than that "it works," in the sense that it produces better results. Our own work succeeded with a mixture of insights and parameter engineering.

By its own metrics, the performance-driven paradigm described above has led to great successes in AI. But is such an approach useful when transferred to other fields? This question relates to a broader debate that arose when "Big Data" came to the scene in the late 2000s.

3.2.1 Data-driven inquiry and the "end of theory"

Data-driven inquiry refers to a mode of research that has data accumulation as its starting point. Learning then happens through applying methods to extract meaningful patterns from that data [Zal20]. Although this is certainly not a new approach to science, the technological advancements of the late 2000s made it possible to conduct data-driven research at an unprecedented scale [Des+22]. The proliferation of the internet, social media platforms, sensors, and other digital devices led to a significant increase in the amount of data being generated. This surge in data availability gave birth to the discipline of data science, whose techniques researchers from various fields could operationalize and apply to their respective domains [Des+22]. By some, this was seen as the beginning of a new era of agnostic science with its own norms and "motley" epistemology [Win01; Leo14; Imb17].

Put pointedly, knowledge could now be generated automatically – without the need for hypotheses or a deeper understanding of the data. In a provocative article in 2008, Anderson refers to this shift as "the end of theory" [And08]. He presents a view that "Petabytes allow us to say: 'Correlation is enough.' [...] We can analyse the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot [...] Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. There's no reason to cling to our old ways" [And08]. The view characterized here is an inductivist, Baconian one [Des+22]. It contrasts with the classical conception of the scientific method that prescribes starting with a theory, deducing hypotheses, and then confirming or falsifying them through experiments [Kit14; Des+22].

Many scholars have argued vehemently against this perceived turn to purely data-driven science. They maintain that data by themselves do not speak [Kit14; Des+22; Fri15]. Data collection, algorithm selection, and interpretations must be embedded in wider knowledge and theory to avoid "fishing expeditions," i.e., unsophisticated investigations with little hope of contributing to any fundamental understanding [Ell+16]. The associations found by ML methods do not necessarily reflect causes [Gig20; Vas+21; Des+22]. There is always the risk of spurious correlations and confounders [Lip18; Leo14; Fri15; Zal20; Vas+21]. Critics in the cognitive science community have analogously expressed concern over the field's current focus on documenting empirical regularities and called for more theory development [Gol22].

3.2.2 AI as an epistemic enhancer

However, not everyone sees the rise of data science pessimistically. Some consider it merely a logical next step in our centuries-old history of using scientific tools – or epistemic enhancers, as Paul Humphrey calls them [Hum04]. These are technologies that increase our capacities to acquire knowledge by augmentation, conversion, or extrapolation.

Augmentation means that we gain access to features of the world we are not naturally equipped to perceive. An example here would be computerized axial tomography (CAT scans) [Hum04]. *Conversion* occurs when a phenomenon accessible to one sensory modality is made available in another [Hum04]. E.g., accelerometer data can be visualized on a screen. Computational modeling often involves this kind of information conversion [Alv23]. *Extrapolation* refers to an expansion of our existing abilities [Hum04]. E.g., a microscope extends our visual capabilities. Many ML methods provide this type of epistemic enhancement [Alv23]. While humans have the ability to extract patterns from data, algorithms allow us to do so faster and on a much larger scale [Fri15]. Analyzing the 136 gigabytes of raw sensor data in this case study manually would have posed quite a challenge. From a more optimistic perspective of epistemic enhancement, ML can thus be considered a worthwhile endeavor to exceed our human limitations when confronted with large quantities of information [Des+22].

It is worth noting that not all uses of AI should automatically be considered epistemic enhancement. Much ML work is focused on improving the accuracy of practical applications. In these cases, the trained model may simply be capturing correlations in the data rather than identifying causal mechanisms. For example, a model designed to predict which products a customer is likely to buy may perform well on a test set but then fail to generalize if the underlying drivers of customer behavior shift. This could be due to changes in the customer's income, family situation, or product preferences. The model designer may respond by retraining the model, temporarily narrowing the performance gap, but this does not address the fundamental issue of missing the true causal factors. Such approaches are closer to empirical data mining than to data science. Thus, while ML techniques can be powerful tools for expanding our knowledge, their value as epistemic enhancers depends on how they are applied.

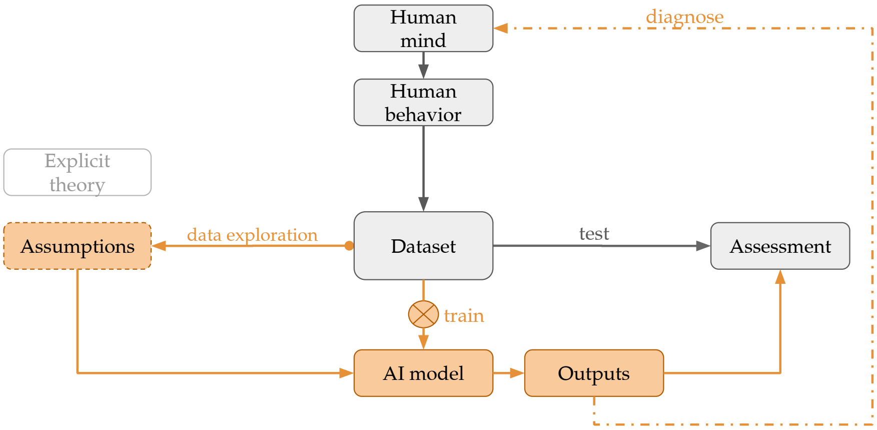


Figure 3.3: Overview of relations between human cognition, theory, assumptions, data, model, and outputs in the first case study. Relevant components performed or generated by us shown in orange. Components provided by third parties shown in gray. Circle at the beginning of an arrow indicates the starting point of investigation.

3.2.3 Relation to the guiding questions

Relating the case study to the criticisms from section 1, it is certainly fair to say that model design was informed more by heuristics than by scientific theory, it is not a biologically plausible model of the brain, and the only kind of explanation it offers is teleological. Teleology means to explain by reference to some purpose, goal, or function. The modeling process is quite typical for NNs (see Figure 3.3). Its starting point is a dataset, often collected and provided by a third party – here, the challenge organizers. This dataset is explored, which leads to certain assumptions. In this case, we concluded that it might be helpful to pre-train the model on predicting the time of day rather than patient IDs and that patient-individual models might work better than one joint model. These assumptions inform the model design and training. The model's outputs are then evaluated – usually by the model designers, but in this challenge, by the organizers. In the future, digital phenotyping models could be used to assess the state of the systems that produced the training data, namely, patients with mental illnesses.

Given that the modeling process began with the data rather than with some scientific hypothesis, the case study could be seen as an example of the much-maligned "end of theory" approach. The criticism against data-driven inquiry that correlations do not automatically impart understanding certainly holds here. Do patients relapse because their routines change? Do the patients' routines change because of a relapse? Is it a reinforcing spiral? I.e., does a lack of sleep or physical activity lead to a worse mental state, which then begets less sleep, less activity, and a further deteriorating psychological condition? The study offers no answers to these questions. We cannot conclude anything beyond the fact that changes in daily routines could be helpful indicators of relapses. However, such purely predictive or diagnostic models might be enough for some use cases. Applications like neural prostheses, brain-computer interfaces, monitoring patient well-being, or developing and optimizing experimental designs can have an inherent value, even if that value is mainly pragmatic [CK19].

Furthermore, a closer look reveals that the kind of modeling practice exemplified in this case study is not as inductivist as it appears at first glance. Although we begin by exploring the data, we do so to identify possible explanations of the data's internal structure. These explanations may relate to the previous scientific literature. For example, we hypothesized that changes in patient routines indicate relapses and found support for this in others' work. This inference from data structure to explanatory hypothesis is an example of abductive reasoning [Des+22].

Together with induction and deduction, abduction is one of the three most commonly recognized modes of inference. The different modes' characteristics can be summarized as follows: "deduction proves that something *must* be, inductive reasoning shows that something *is*, while abduction merely suggests that something *may* be [Mil10, p. 194]". Abductive inference is commonplace in many fields, including social, political, economic, and cognitive sciences [Des+22; Kit14]. Besides the informal hypothesizing involved in NN design, the model could also become embedded into broader scientific practice by inspiring further psychiatric experiments that allow for causal manipulation.

To sum up, we can conclude from the first case study that a NN need not be derived from theory, biologically plausible, or offer much in the way of explanation to be useful. Even models that do not fulfill these criteria can serve as epistemic enhancers. They can convert information between sensory modalities or expand the domain of our human analytic abilities. These models can provide value by virtue of their outputs, e.g., optimized experiment designs or alerts that a patient is relapsing. We can also say that NN modeling is usually an abductive rather than merely an inductive endeavor. That is, it may be data-driven and approximate but not wholly disconnected from theory.

Recall, however, that the epistemic goal of cognitive science is to systematically understand how agents perform tasks. The automatic fitting of a model to a dataset carries little systematic understanding beyond the actual solution of the problem itself [NPS14]. This kind of pragmatic "oracle" model may thus not be satisfactory in all cases [SA19]. The following case study, therefore, tries to go into more explanatory depth.

4 Modeling the emergence of compositional generalization on the grounded SCAN dataset with selective attention

> *"That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"*
>
> *"About six inches to the mile."*
>
> *"Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"*
>
> *"Have you used it much?" I enquired.*
>
> *"It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."*
>
> From "Sylvie and Bruno Concluded" by Lewis Carroll

4.1 Study

The second case study relates to the core systems of objects and layout geometry (see Figure 4.1). Layout geometry supports spatial navigation by encoding sense and egocentric distance [HM19]. The object system allows us to represent and track physical items. However, due to our limited attentional resources, it can only handle a few objects simultaneously [SK07]. We take inspiration from these core cognitive systems to equip an AI model with two inductive biases in the form of a selective attention bottleneck and egocentric location encoding.

The model is trained and tested on the Grounded Simplified version of the CommAI Navigation tasks (gSCAN), a benchmark dataset that requires an agent to navigate a 2D grid world and to interact with objects. Our goal is to determine whether including selective attention and egocentric location encoding can improve accuracy and sample efficiency on this dataset and, if so, to investigate which factors contribute to performance. The proposed model is a hybrid architecture that combines layers trained with gradient descent and components optimized with an evolutionary strategy. Through ablation studies, neuron pruning, and error analyses, we show that both of the human-inspired inductive biases contribute to performance. Particularly, the selective attention mechanism drastically improves the model's sample efficiency.

A version of this case study was published in the Proceedings of the BlackboxNLP Workshop at the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP) [HD22a].

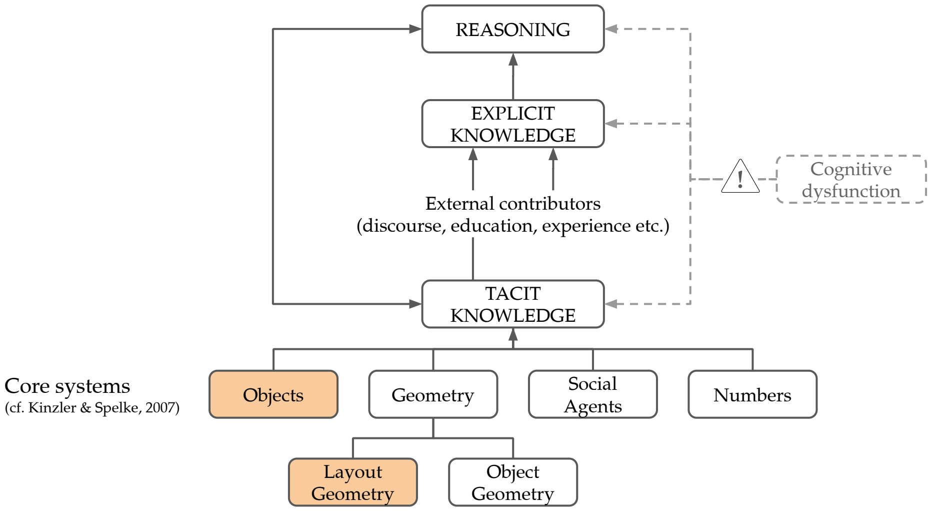


Figure 4.1: Situating the second case study in the broader study of cognition. Relevant parts of the framework marked in orange.

4.1.1 Introduction

The brain sciences have long been a source of inspiration for novel algorithms and AI architectures [LKC20; SK20]. For instance, the first connectionist models tried to mimic neural information structures in an abstracted way [Has+17]. Convolutional Neural Networks (CNNs) were inspired by the visual cortex's hierarchical processing of stimuli [YD16]. Reinforcement Learning (RL) has benefited from temporal-difference methods, which emerged from behaviorist psychology research on animal behavior. Deep Q-Networks utilize "experience replay," which mirrors episodic memory by randomly sampling and replaying past experiences to avoid catastrophic forgetting. LSTMs are designed to remember information over extended time intervals, similar to human working memory – the list could go on [Has+17].

Despite this rich history, the influence of neuroscience and psychology on AI has become rarer in recent years. As models have grown increasingly powerful and their adoption in industry has become ubiquitous, the incentive structure has shifted towards solving the problems that arise when training and deploying NNs at scale [Mom23]. This engineering-focused approach has led to neural network designs taking more freedoms and adopting solutions alien to mental processes in the name of pragmatism [PP20]. However, some researchers have argued against this trend, suggesting that we need to (re-)turn to the brain for inspiration and guidance if we want to improve AI models [Lak+17; BMM19; GB22a]. For all the progress DL systems have made over the years, there still exists a gap between state-of-the-art neural networks and biological minds. One of the aspects in which NNs continue to lag behind human intelligence is compositional generalization [LB23].

Compositionality refers to the ability to understand and generate complex expressions or ideas by combining simpler parts according to systematic rules [LB23]. It is what allows us to produce infinite combinations of known linguistic, visual, or motor concepts, including combinations we have never encountered before. For instance, if we understand English grammar and the words "blue," "car," and "fast," we can comprehend the sentence "The blue car is fast" even if we are hearing it for the first time [PA16]. The ability to reason compositionally is a desirable property for AI models, as it could lead to more human-like, robust generalization on out-of-distribution data and increased sample efficiency.

Compositionality in neural networks has thus been the subject of numerous empirical investigations – with mixed results. Several studies using a variety of NN architectures have found that models either failed on compositional tasks or succeeded given enough data but did so without relying on systematic compositional rules [Bar20; LB18; LBL18; SSG19; Key+19; Hup+20; And+19; Cha+20]. Others found that such architectures could reach compositional solutions without being explicitly constrained to do so but that this ability varied dramatically across random initializations of the same model [LKB18; MML20; WSB18].

This work addresses the challenge of compositional reasoning by taking inspiration from the cognitive science concepts of selective attention and joint attention. Selective attention helps humans focus on specific parts of their sensory input [Tyn+17]. The ability to concentrate on certain stimuli while ignoring others is crucial for learning and reasoning compositionally [BTT15]. Joint attention is a social form of selective attention. It refers to a shared focus of two individuals and plays an important role in children's development [TC07]. By looking or pointing at specific stimuli, caregivers direct children's attention to certain inputs [LS20]. This guidance helps children parse their environment into relevant components and scaffolds their learning process [DMP16]. In the following study, we embed selective and joint attention processes into the training setup of a neural network to investigate the effect of this inductive bias on the model's ability to reason compositionally and its sample efficiency.

Our dataset of choice for this investigation is gSCAN, a challenge benchmark for systematic generalization in grounded language understanding. The model we use is a hybrid architecture, containing some weights trained with gradient descent, some optimized with an evolutionary strategy, and some initialized randomly and left frozen. A detailed justification of these design choices is given in Section 4.1.4. The architecture has around 60 times fewer trainable parameters than models previously tested on gSCAN, which allows us to run extensive ablation studies and error analyses to investigate factors contributing to generalization performance. Our findings indicate that including selective and joint attention mechanisms helps the model break down gSCAN tasks into simpler, reusable parts and to combine them compositionally. The model achieves accuracies comparable with previously proposed, more complex models on most test splits – even when trained on as little as 2% of the full dataset. On adverb-to-verb generalization, it outperforms previous proposals by 65 to 86%.

4.1.2 Related Work

Compositional generalization A number of works have addressed the challenge of building AI systems that generalize compositionally. Neural Module Networks were designed for visual question answering, and they achieve systematicity by dynamically assembling question-specific models out of trainable reusable components [And+16a; And+16b; Bah+18]. Other approaches explore ways of encouraging compositional representations in commonly used state-of-the-art models without major architectural changes. In this vein, Hupkes et al. [Hup+18] and Baan et al. [Baa+19] find that attentive guidance during training helps develop small functional groups of neurons that yield more compositional solutions by seq2seq models on lookup table tasks. Andreas [And20] and Akyürek et al. [AAA20] propose data augmentation schemes that promote compositional learning in instruction following and morphological analysis. Ontañón et al. focus on the effect that design decisions such as position encodings, weight sharing, or model hyper-parameters can have on the compositional generalization abilities of Transformer models [Ont+22]. Finally Power et al. identify weight decay as being particularly effective at improving generalization on a binary operation table task [Pow+21].

Grounded instruction following Several datasets have been proposed in recent years for training embodied agents to follow instructions in simulated 2D or 3D environments [Her+17; Yu+18; Mis+18; Cha+18; YZX18; Der+19; Che+19; Shr+20]. One such dataset is gSCAN, which was specifically introduced as a benchmark for compositionality in grounded language understanding and contains nine test splits for assessing different kinds of out-of-distribution generalization [Rui+20]. Previous approaches to solving gSCAN include language-conditioned message passing [GHM20], compositional networks [KKB21], neuro-symbolic, dual-system models [Nye+21], and the introduction of auxiliary tasks [JB21; HB20]. The most successful model to date uses a general-purpose Transformer architecture with cross-modal attention and solves five out of nine tasks [Qiu+21].

Neuroevolution Evolutionary Algorithms (EAs) are stochastic, gradient-free methods that explore multiple areas of a search space in parallel. Our work was particularly inspired by Tang et al., who combine EA techniques with neural networks to solve vision-based RL tasks [TNH20]. Their model extracts relevant patches from input images through a hard (non-differentiable) attention mechanism optimized via an EA rather than more commonly used techniques like RL. The most attended-to patches are then passed on to an LSTM controller, which determines the agent's action. The authors find that this approach significantly reduces the number of model parameters needed compared to previous methods, as well as offering increased interpretability and higher robustness to out-of-distribution modifications [TNH20].

4.1.3 Background

Following Tang et al., we make use of the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) to implement our selective attention mechanism. However, we use an Echo State Network (ESN) for the agent's memory component rather than an LSTM to reduce the number of learnable parameters needed (see Section 4.1.4). As neither technique is commonly used in DL, we provide some background on these topics here.

CMA-ES CMA-ES is a black-box optimization algorithm that has been empirically shown to perform robustly on a range of tasks [Han+10]. It also requires minimal parameter tuning, which made it the optimization method of choice in the work of Tang et al. that inspired our architecture. CMA-ES begins by sampling λ individual solutions $x_1^{(g+1)}, ..., x_\lambda^{(g+1)}$ from a multivariate Gaussian distribution $\mathcal{N}\big(m^{(g)}, {\sigma^{(g)}}^2 C^{(g)}\big)$ with mean $m^{(g)}$, step size $\sigma^{(g)}$ and covariance matrix $C^{(g)}$. The initial mean, step size, and covariance matrix are then adapted iteratively to increase the likelihood of successful solutions as evaluated by some function f. Mean adaptation is done by shifting m by the weighted average of the μ best solutions of generation g [Sha+20]:

$$m^{(g+1)} = m^{(g)} + c_m \sum_{i=1}^{\mu} w_i \big(x_{i:\sigma}^{(g+1)} - m^{(g)}\big), \tag{4.1}$$

where c_m is a learning rate. The new step size σ is determined as follows [Sha+20]:

$$\sigma^{(g+1)} = \sigma^{(g)} \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\left\|\mathbf{p}_\sigma^{(g+1)}\right\|}{E\|\mathcal{N}(0,I)\|} - 1\right)\right), \tag{4.2}$$

where c_σ is a separate learning rate, d_σ is a damping parameter, and $\mathbf{p}_\sigma^{(g+1)}$ is the next generation's conjugate evolution path computed as [HMK03]:

$$\mathbf{p}_\sigma^{(g+1)} = (1 - c_\sigma) \cdot \mathbf{p}_\sigma^{(g)} + \sqrt{c_\sigma \cdot (2 - c_\sigma)} \cdot \frac{\sqrt{\mu}}{\sigma^{(g)}} (x_\mu^{(g+1)} - x_\mu^{(g)}). \tag{4.3}$$

Finally, the covariance matrix is updated [HMK03]:

$$C^{(g+1)} = (1 - c_{cov}) \cdot C^{(g)} + c_{cov} \cdot \mathbf{p}_c^{(g+1)} (\mathbf{p}_c^{(g+1)})^T, \tag{4.4}$$

where c_{cov} is another learning rate. For a more in-depth description of the CMA-ES algorithm please see Hansen et al. [HO01a].

Echo State Networks A basic ESN consists of an input layer W_i^r , a Recurrent Neural Network (RNN) or so-called reservoir, and an output layer W_o. The reservoir's state is updated at each discrete time step as follows:

$$\mathbf{x}[n+1] = (1 - \alpha)\mathbf{x}[n] + \alpha f\big(W_i^r \mathbf{u}[n] \ + W_r^r \mathbf{x}[n]\big), \tag{4.5}$$

where α is a leak rate, $\mathbf{x}[n]$ is the current reservoir activation state, f is a hyperbolic tangent function, $\mathbf{u}[n]$ is the external input, and W_r^r is the reservoir's internal weight matrix. The ESN's output is computed as

$$\mathbf{y}[n+1] = g(W_o \mathbf{x}[n+1]), \tag{4.6}$$

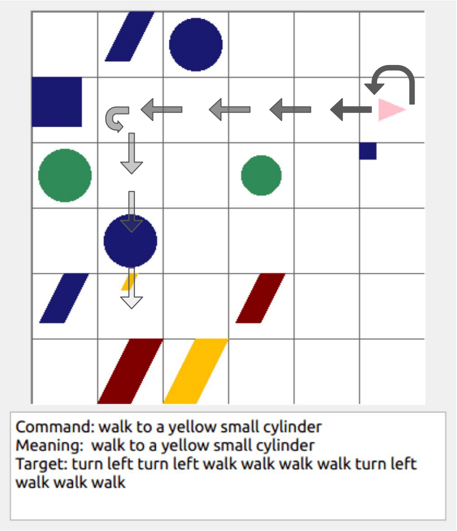
where g is an activation function. Crucially, W_i^r and W_r^r are randomly initialized and left untrained. Only W_o is optimized. This leads to considerably faster training times than for conventional RNNs where all weights are learned [Gau+21]. ESNs' main areas of application, therefore, include resource-constrained contexts like robotics and edge computing [Nak20].

4.1.4 Methods

gSCAN Benchmark The gSCAN environment is a grid with objects of various shapes, sizes, and colors. It is represented as a $16 \times 6 \times 6$ array, where 6 is the grid size and 16 is the dimension of the feature encoding for each grid cell. The agent receives synthetic English language instructions, which it must carry out using a combination of six output actions. These actions include walking, turning left or right, pushing, pulling, and staying put. The agent may be asked to navigate to or interact with a specified object. Objects can be moved by pulling or pushing, once for light objects and twice for heavy objects.

Instructions may omit attributes if they can be deduced from the context. For instance, the same object could be referred to as "the small yellow circle", "the yellow circle", "the yellow object", etc., depending on whether there are other small, yellow, or circular distractor objects. "Small" and "large" are not descriptions of absolute sizes but are always to be inferred relative to other objects. All commands can be modified by adverbs, such as "cautiously" or "while spinning.". This means the agent may have to look both ways before moving or turn around four times before each step. Figure 4.2 shows a sample task.

Some input combinations are withheld from the training set. Out-of-distribution generalization is then assessed on nine separate test splits containing only examples with unseen combinations,



Command: walk to a yellow small cylinder
Meaning: walk to a yellow small cylinder
Target: turn left turn left walk walk walk walk turn left walk walk walk

Figure 4.2: Example of a gSCAN task. Agent represented as a pink triangle.

Table 4.1: Overview of gSCAN's compositional test splits.

Test Split	**Held-out Examples**
A: Random	Random (in-distribution)
B: Yellow Squares	Yellow squares as targets if referred to as *yellow*
C: Red Squares	Red squares as targets
D: Novel Direction	Targets south-west of the agent
E: Relativity	Circles of size 2 referred to as *small* (references are relative to other grid objects, not tied to absolute sizes)
F: Class inference	Pushing squares of size 3 (*heavy* objects are pushed/pulled twice)
G: Adverb k = 1	All except *k* mentions of *cautiously* (looking both ways before each step)
H: Adverb to verb	Commands containing both *pull* and *while spinning* (turning 4 times)
I: Length	Action sequences of length ≥ 15

listed in Table 4.1. Performance is measured using exact match accuracy of predicted action sequences. The full dataset has $\approx 370,000$ training and $\approx 20,000$ test sequences.

Hupkes et al. propose to distinguish between five types of compositionality, namely, the systematic recombination of known parts and rules (*systematicity*), the extension of predictions beyond lengths seen during training (*productivity*), robustness to synonym substitutions (*substitutivity*), dependence on local vs. global structures (*localism*), and the preference for rules vs. exceptions (*overgeneralization*) [Hup+20]. Following this taxonomy, split G tests the model's one-shot learning capabilities or overgeneralization. Split I tests for productivity. We mainly consider splits B, C, D, E, F, and H, which focus on systematic generalization and substitutivity.

Model To solve a task, the agent requires knowledge of the instruction, the grid state, and its own past actions. The latter is needed to keep track of, e.g., the number of turns completed when "spinning".

To create the representation of the language command, we chose an ESN due to its ability to capture information about all input words and their order in a single vector without requiring any weight optimization. This fits our goal of keeping the number of trainable parameters low for ease of analysis. The instruction to the agent is tokenized, one-hot encoded, and input sequentially to a reservoir with 400 hidden neurons, which is updated after each token according to Equation 4.5. All reservoir neurons are randomly connected to an output layer W_o of size 64, yielding a 64-dimensional command embedding $\mathbf{x}_{\text{lang}} \in \mathbb{R}^{1\times 64}$.

The model's selective attention component is responsible for extracting task-relevant information from the input grid. Note that here, we are not referring to the widely used attention mechanism of Transformer models. Our "attention" is a much simpler form of relevance computation. We pass the command embedding $\mathbf{x}_{\text{lang}}$ through a layer $W_{\text{vis}} \in \mathbb{R}^{64\times 16}$. The resulting vector is convolved with the input grid at each position to obtain a heatmap over a grid $G \in \mathbb{R}^{16\times 6\times 6}$. Thus, W_{vis} can be understood as a kind of association matrix or learned filter, which maps between language and visual inputs. The x- and y-coordinates and the 16-dimensional feature vector for the most-attended grid cell $\mathbf{g}^*$ are then extracted:

$$\mathbf{g}^* = \arg\max\big((\mathbf{x}_{\text{lang}} \cdot W_{\text{vis}}) * G\big) \tag{4.7}$$

Because this arg max operation is non-differentiable, we follow Tang et al.'s approach of using CMA-ES to optimize W_{vis}. However, in contrast to Tang et al., we are working with feature vectors rather than image patches, and we do not evolve all learnable parameters in our model. This is because our model has significantly more parameters than that of Tang et al., and the time and space complexity of CMA-ES is quadratic in the dimensionality of its objective function. This restricts its application to problems with no more than a few hundred variables [Var+18]. Therefore, only the selective attention part of the model is optimized using CMA-ES. The rest is trained using gradient descent. Inspired by joint attention mechanisms and parental guidance during child learning, the CMA-ES receives auxiliary feedback on whether the correct target object was most attended to. We also test and report the results for a version where the CMA-ES only receives as feedback the cross-entropy loss produced by the agent's final prediction outputs (see Section 4.1.5).

The action attention part of the model serves as the agent's "memory" of past outputs. The command embedding $\mathbf{x}_{\text{lang}}$ undergoes a self-associative step, where it is passed through a layer W_{lang} and multiplied element-wise with the original $\mathbf{x}_{\text{lang}}$, yielding a weighted embedding $\mathbf{a}_{\text{lang}} \in \mathbb{R}^{1\times 64}$. This is then passed through another layer $W_{\text{act}} \in \mathbb{R}^{64\times 200}$ and multiplied element-wise with a vector $\mathbf{x}_{\text{act}} \in \mathbb{R}^{200\times 1}$ containing the agent's one-hot encoded past 20 actions and orientations:

$$\mathbf{a}_{\text{act}} = (((\mathbf{x}_{\text{lang}} \cdot W_{\text{lang}}) \odot \mathbf{x}_{\text{lang}}) \cdot W_{\text{act}}) \odot \mathbf{x}_{\text{act}} \tag{4.8}$$

As there is no arg max operation involved, W_{act} is trained with conventional gradient descent.

The outputs of the selective and action attention modules are concatenated with the agent's current x- and y-coordinates and orientation, as well as the original command embedding. Coordinates of the target cell extracted via the selective attention module are encoded as row- and column-wise relative distances to the agent. This is motivated by the fact that egocentric spatial encoding supports navigation in animals and humans and is sometimes used in RL goal navigation tasks [VC22]. The concatenated inputs are passed to the agent's controller to predict the agent's next step. The controller consists of a normalization layer, a linear layer with 100 hidden ReLU units, and an output layer of size 6, as the agent has six output options. In total, the model has a little under $5 \cdot 10^4$ trainable parameters (see Table 4.2), less than 2% of the $3 \cdot 10^6$ for models previously tested on gSCAN [Qiu+21]. A schematic overview is shown in Figure 4.3.

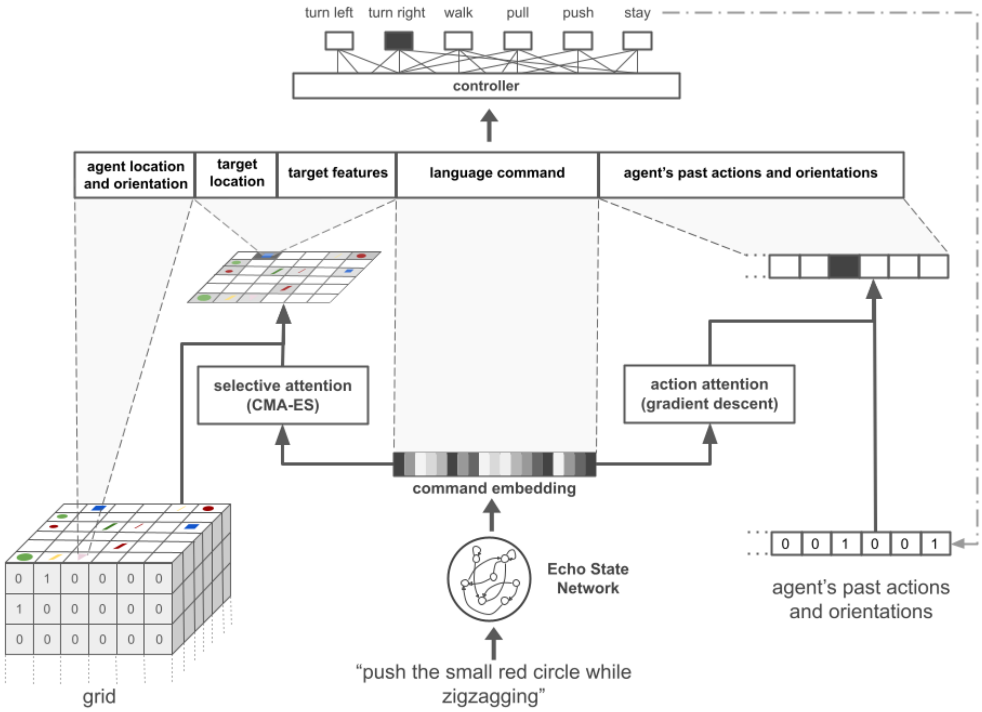

turn left
turn right
walk
pull
push
stay
controller
agent location and orientation
target location
target features
language command
agent's past actions and orientations
selective attention (CMA-ES)
action attention (gradient descent)
command embedding
Echo State Network
0 0 1 0 0 1
agent's past actions and orientations
0 1 0 0 0 0
1 0 0 0 0 0
0 0 0 0 0 0
grid
"push the small red circle while zigzagging"

Figure 4.3: Schematic visualization of our model.

Training details The weights of the ESN were initialized with a spectral radius of 0.99 and a density of 10^{-2}. The leaking rate was set to 10^{-1}. For the CMA-ES, we used a population size of 8 and an initial normal distribution with standard deviation 10^{-1}. For the part of the model trained via gradient descent, we used the Adamax optimizer and a learning rate cycle with an upper boundary of 10^{-2}. Weight decay was set to 10^{-4}, and models were trained with batch size 4,096 for 100 epochs unless otherwise specified. All performance results are based on 10 runs. Each run used a different random seed for model weight initialization. However, the same 10 seeds were used for all tested modified or ablated architectures so that all compared models started with the same 10 sets of weights. Experiments were implemented in Pytorch [Pas+19]. The training time for one model was approximately 1.3 hours on the full dataset. For reference, the gSCAN authors report the training time for their baseline model as less than 24 hours [Rui+20].

Table 4.2: Overview of our model's trainable parameters (biases were only used in layer normalization).

Parameter	**Size**
Hidden layer	28,800
Layer normalization weights	100
Layer normalization biases	100
Output layer	600
Selective attention key matrix	1,024
Self-attention key matrix	4,096
Action attention key matrix	12,800
Total	**47,520**

Table 4.3: Exact match accuracy on gSCAN compositional splits. For our model (trained with auxiliary attention feedback), we report both the performance of models trained on the full dataset and of those trained on a 10% subset.

	A	B	C	D	E	F	G	H
Seq2Seq (2020)	97.7 ± 0.2	55.0 ± 39.4	23.5 ± 21.8	0.0 ± 0.0	35.0 ± 2.4	92.5 ± 6.8	0.0 ± 0.0	22.7 ± 4.6
GECA (2020)	87.6 ± 1.2	34.9 ± 39.3	78.8 ± 6.6	0.0 ± 0.0	33.2 ± 3.7	86.0 ± 0.9	0.0 ± 0.0	11.8 ± 0.3
Heinze (2020)	94.2 ± 0.7	86.5 ± 6.3	81.1 ± 10.1	-	43.4 ± 7.0	-	-	-
Gao (2020)	98.6 ± 1.0	99.1 ± 0.7	80.3 ± 24.5	0.2 ± 0.1	87.3 ± 27.4	99.3 ± 0.5	-	33.6 ± 20.8
Kuo (2020)	96.7 ± 0.6	94.9 ± 1.3	67.7 ± 10.8	**11.5 ± 8.2**	76.8 ± 2.3	98.7 ± 0.1	1.1 ± 0.3	21.0 ± 1.4
Qiu (2021)	**100 ± 0.0**	**99.9 ± 0.1**	99.2 ± 0.9	0.0 ± 0.0	**99.0 ± 1.2**	**100 ± 0.0**	0.0 ± 0.0	22.2 ± 0.01
Jiang (2021)	-	-	-	-	-	-	**4.9**	28.0
Nye (2021)	74.7	81.3	78.1	0.0	53.6	76.2	0.0	21.8
Ours (100%)	99.7 ± 0.1	73.5 ± 25.4	99.4 ± 0.4	2.2 ± 1.5	97.4 ± 2.0	99.1 ± 0.6	0.0 ± 0.0	**98.4 ± 1.1**
Ours (10%)	99.5 ± 0.1	81.6 ± 14.3	**99.5 ± 0.2**	3.5 ± 2.7	96.8 ± 1.9	98.3 ± 1.7	0.0 ± 0.1	94.2 ± 3.7

Table 4.4: Sequence and attention match accuracies on additional held-out verb-adverb and shape-color combinations. Original verb-adverb and shape-color splits H, B, and C shown for reference. All models trained with auxiliary feedback on the full dataset.

	Attention match	Exact match	Exact match if attention match
Custom split: Pull while spinning + push while zigzagging + walk hesitantly	0.98 ± 0.01	0.99 ± 0.01	0.996 ± 0.00
H: Adverb to verb	1.00 ± 0.00	0.93 ± 0.06	0.943 ± 0.06
Custom split: Yellow squares + red squares + green cylinders + blue circles	0.99 ± 0.01	0.99 ± 0.02	1.00 ± 0.00
B: Yellow squares	0.86 ± 0.14	0.83 ± 0.17	1.00 ± 0.00
C: Red squares	1.00 ± 0.01	0.99 ± 0.01	1.00 ± 0.00

4.1.5 Results

Performance As shown in Table 4.3, the model with auxiliary attention feedback reaches competitive accuracy on splits A, C, E, and F. On split H, it outperforms previous proposals by 65 to 86%. To see if generalization extended to other combinations, we also tested two custom splits. The first is a variation of task C, where not only red squares but also yellow squares, green cylinders, and blue circles never appear as targets during training. The second is an extension of split H, where in addition to "pull while spinning", the agent is never told to "push while zigzagging" or to "walk hesitantly" during training. The model generalized to test sets containing only held-out shape-color and verb-adverb combinations, reaching 98.7% ± 1.5 and 98.9% ± 0.5 accuracy, respectively (see Table 4.4).

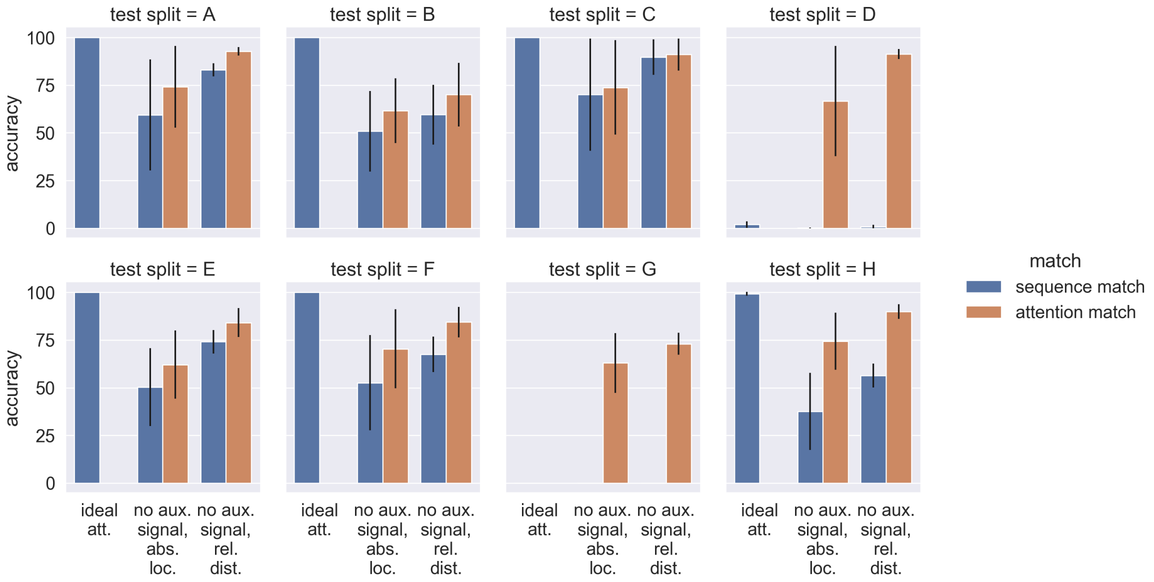
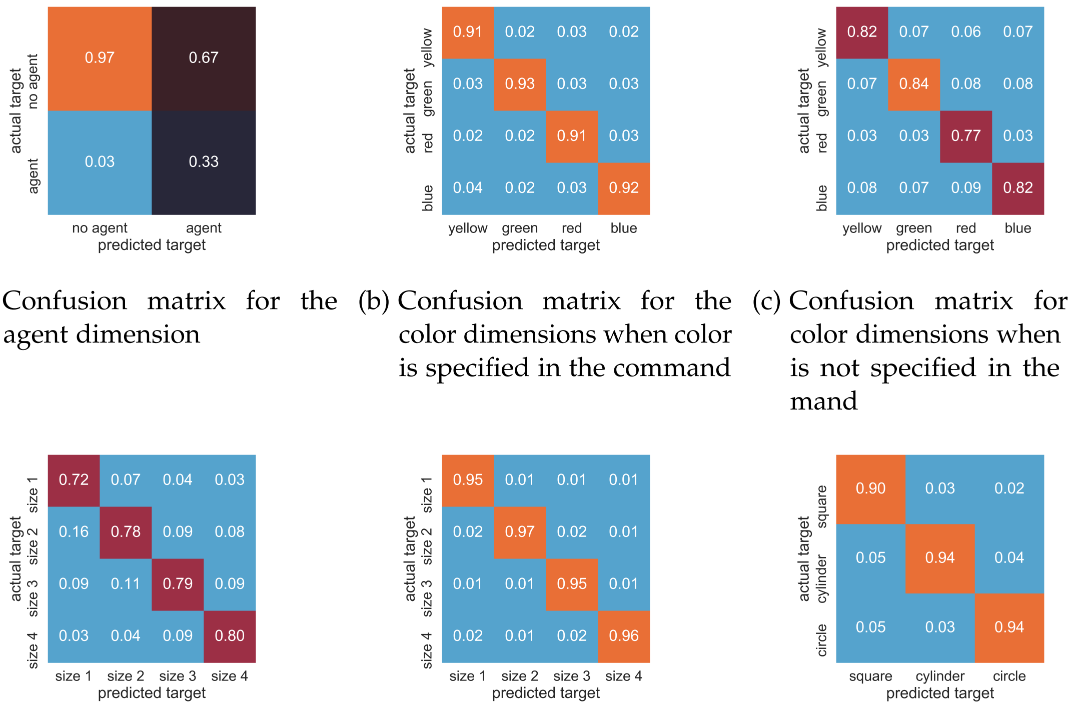


Figure 4.4: Exact match accuracy and attention match accuracy on gSCAN compositional splits for models with idealized selective attention, selective attention optimized without auxiliary feedback and absolute target locations, and selective attention optimized without auxiliary feedback and relative (egocentrically encoded) target locations.

Figure 4.4 analyzes the performance of models trained without the joint attention-inspired auxiliary signal. Instead of feedback on its target choice, the selective attention module only receives the cross-entropy loss of the model's final output. Figure 4.4 also compares models using absolute locations vs. relative target distances. We show both sequence and attention match accuracy. Sequence match refers to whether an instruction was carried out correctly. Attention match denotes whether the selective attention module picked out the right target. We include sequence match accuracy for idealized models receiving perfect target location inputs for reference. As the attention match accuracy shows, models without an auxiliary signal do learn to focus on the target in some cases, but their performance exhibits high variation. Egocentric encodings outperform allocentric ones in both sequence and attention match.

We analyze the mistakes made by the models trained without auxiliary feedback by treating the task of focusing on the correct target as a classification in its own right and analyzing the feature-wise confusion matrices of the models (see Figure 4.5). This reveals an accumulated false discovery rate of 66.5% for the “agent" dimension of the grid cell feature vectors (Figure 4.5a), compared to 0% for the models trained with feedback. Thus models without attentive guidance tend to overly focus on the agent. The location of the agent does coincide with the target object's location around 18% of the time, which might lead to an over-reliance on this dimension. We also find that models trained without attention supervision struggle more with under-specified commands. For example, the models focus on an object of the correct color in ca. 92% of cases when the color is explicitly mentioned in the command (Figure 4.5b). When the target object is only referred to by its shape or size, the accuracy drops to about 81%.

In the case of split B (yellow squares), performance exhibits a large variation (see Table 4.3). Out of 10 runs, approximately half always achieve accuracies in the range of 90 - 99% while the others only reach 35 - 55%. A look at the confusion matrices (not shown) reveals that, on average, models correctly identify a square as their target object in 97% of test cases. However, their color accuracy is only around 75%. Taken together, this suggests that the models overfit



(a) Confusion matrix for the agent dimension

(b) Confusion matrix for the color dimensions when color is specified in the command

(c) Confusion matrix for the color dimensions when color is not specified in the command

(d) Confusion matrix for the color dimensions when size is specified in the command

(e) Confusion matrix for the color dimensions when size is not specified in the command

(f) Confusion matrix for the shape dimensions (always specified in the command)

Figure 4.5: Confusion matrices of the agents' selective attention module, trained without an auxiliary signal.

to the absence of yellow squares. Depending on the random initialization of W_{vis}, a model may be more or less predisposed to generalization on this task. In the absence of any samples with yellow squares that could cause a course correction, this predisposition may be exacerbated with each update and thus deteriorate performance in higher-data regimes. This would explain why performance is better for the 10% subset than for the full dataset.

Regarding the low performance on split D (targets southwest of the agent), the problem seems to be related to navigation rather than to the identification of the correct target. The model's attention match accuracy is 100%. However, it cannot find its way to the identified cell successfully. On average, it ends up in the correct row in 44% of cases, in the right column in 23% of cases, and never both.

Sample efficiency Besides the model's accuracy, we are also interested in its sample efficiency. As shown in Figure 4.6, it achieves around 90% accuracy on splits A and C when trained on only 1% of the dataset and 90 - 97% accuracy on splits A, C, E, and F with 2% of the data. This is well below the 40% data requirement threshold identified by L. Qiu et al. for their cross-modal Transformer model [Qiu+21]. In accordance with our discussion of split B in the previous section, the exact match accuracy for shape-color splits B and C peaks at the 10% subset and declines slightly when given more data. Performance on task H increases more slowly than on other splits and requires at least 10% of the dataset to surpass 90% accuracy.

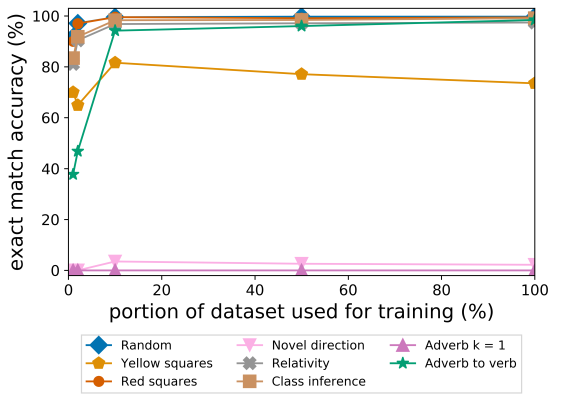


Figure 4.6: Sample efficiency on test splits for models with selective attention and auxiliary feedback.

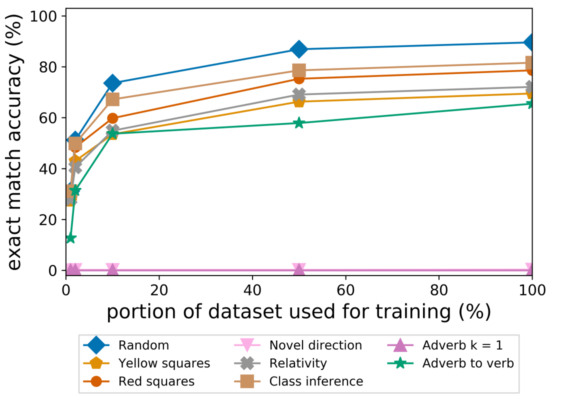


Figure 4.7: Sample efficiency on test splits for models without selective attention.

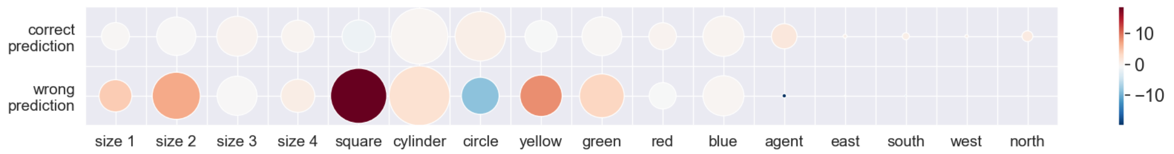


Figure 4.8: Standardized residuals of a χ^2-test comparing the wrong predictions of models trained without vs. with selective attention, on in-distribution data. Circle color represents absolute value. Red indicates that a feature is over-represented, blue indicates it is under-represented. Circle size represents the number of occurrences.

Much of this sample efficiency seems to be due to the selective attention mechanism. To demonstrate this, we train a model without selective attention. Instead of the isolated feature vector of the most attended-to grid cell, this model's controller receives the flattened, attention-weighted whole grid as input. To account for the added dimensionality, we increase the number of neurons in the controller to 500. Figure 4.7 shows that performance-wise, this causes a drop-off, but the model still achieves around 90% accuracy on split A (in-distribution data) when trained on the full dataset. However, the models need to have seen more combinations to start generalizing. This is also supported by a comparison of the confusion matrices for models with and without selective attention via a χ^2-test on split A. Figure 4.8 shows the strength of the difference between observed and expected values. Squares, the color yellow, and small object sizes, which are under-represented in the training data, are over-represented in the incorrect target predictions of models trained without selective attention.

Ablations As shown in Figure 4.9, ablating weight decay or attention over past steps causes the steepest performance drops in splits E, F, and H. To compare structural differences between the ablated models, we perform a neuron pruning experiment. We record the activation of each neuron in the controller's final layer, multiplied by the neuron's outgoing weight. We then sort neurons based on their accumulated contribution to the final model output, and we test exact sequence accuracy on the gSCAN validation set with the top X% of neurons active, where X is a variable. The remaining neurons are disabled by setting outgoing weights to 0. The result is shown in Figure 4.10. All full models require only 13 hidden neurons to solve all tasks. Without action attention, 16 neurons are needed to reach the final accuracy. For models without weight decay, pruning any of the 100 neurons leads to decreased performance.

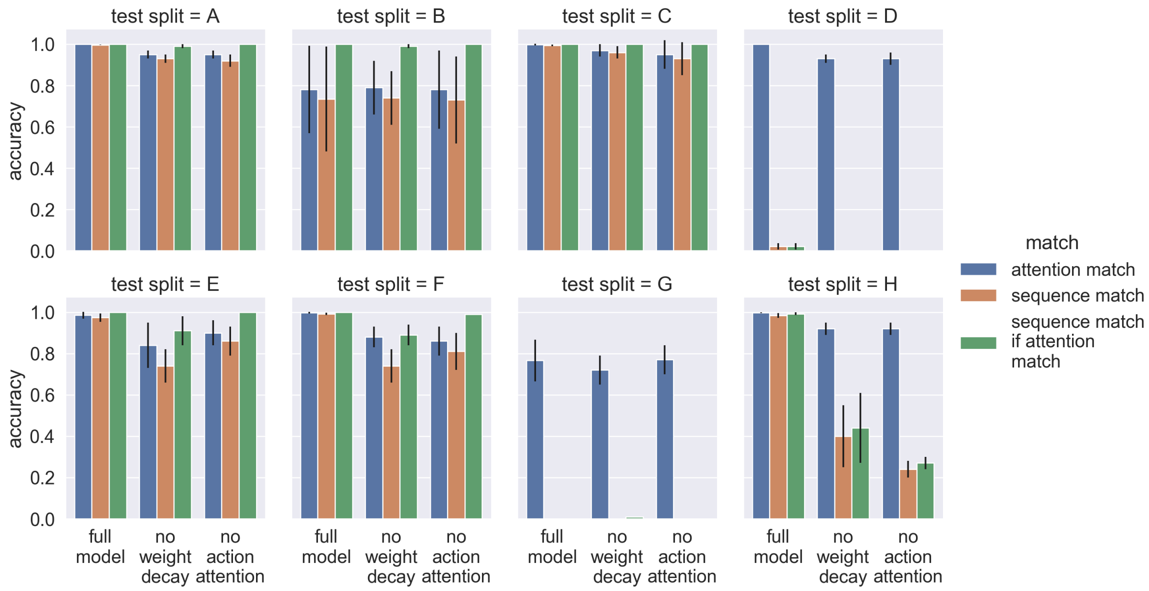


Figure 4.9: Sequence and attention match accuracies on gSCAN compositional splits with the full selective attention model, weight decay ablation, and action attention ablation (all trained on the full dataset).

This difference in learned representations is also illustrated in Figure 4.11, which shows the weights between the agent's past actions and the hidden layer of three identically initialized models with different ablations applied. The model with weight decay and action attention learns the most sparse weights and focuses on recent steps. The hidden model without action attention has a similarly sparse hidden layer, but a longer "memory", i.e., it takes into account past actions from further back in the step sequence. The model without weight decay is very densely connected.

“Spontaneous" generalization During our ablation studies, we observed that generalization on the “adverb to verb" split does occur frequently in models without weight decay and action attention, but only sporadically. As shown in Figure 4.12, performance on split H may spike on one training batch, then fall again. Higher systematic generalization ability is not necessarily evident from looking at the performance on in-distribution data – two models can have the same training loss or test accuracy, but very different out-of-distribution accuracies. Such spurious generalization behavior may also explain the variation in performance on split H observed by Gao et al. [GHM20] and Jiang et al. [JB21].

One reason often cited for unstable generalization is sharp local minima [Kes+17]. However, a visualization of model loss landscapes at various points during training shows relatively flat planes. The landscapes for training and split H data are simply well aligned for some model-batch combinations and less so for others (see Figure 4.13). We also investigated whether the batches used to update the models immediately before out-of-distribution performance spikes had any special properties that would facilitate generalization. We saved batches that preceded an increase in split H accuracy of at least 5%, injected them randomly into the training of other models, and recorded the difference in performance caused. However, we found no statistically significant improvement over random batches and no statistically significant differences in feature or label distributions of such “spike" batches.

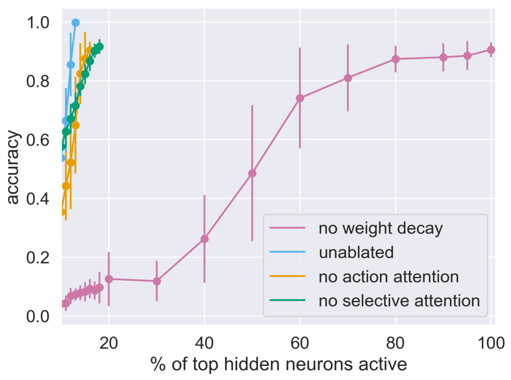


Figure 4.10: Sequence accuracy on in-distribution data for ablated and unablated models with different percentages of disabled top contributing neurons.

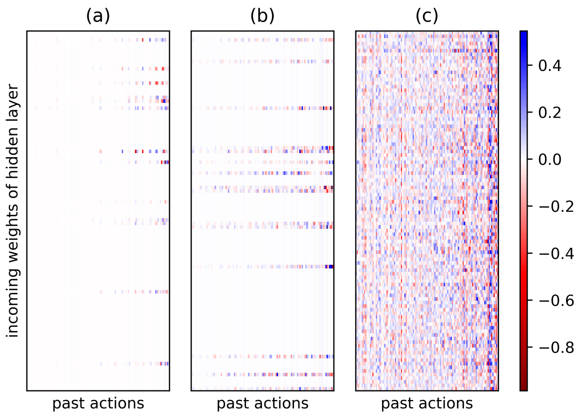


Figure 4.11: Weights between past agent actions and the controller's input layer, as learned by the full model (a), model without action attention (b), and model without weight decay (c).

We did find that batch size had an impact on the likelihood of generalization spikes. We trained 10 models without weight decay on a 2% subset of the training data with batch sizes 256, 512, 1024, 2048, and 4096. All models were initialized with the same random seeds and trained for the same number of absolute updates. We then sampled the models' performance on split H at 50 points in regular intervals during training. As shown in Figure 4.14, generalization performance with smaller batches was higher but more volatile. Comparing the distribution of sampled split H accuracies across batch sizes yielded statistically significant Z-scores > 2 between batch sizes ≤ 512 and ≥ 2048. This is consistent with previous findings that smaller batch sizes facilitate better generalization [SL18; Kes+17; Smi+18; HHS17; ML18].

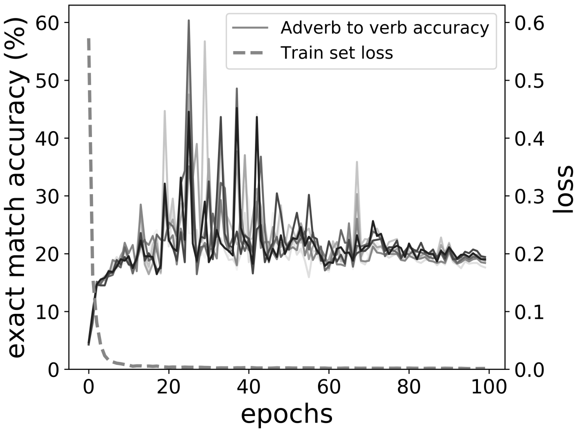
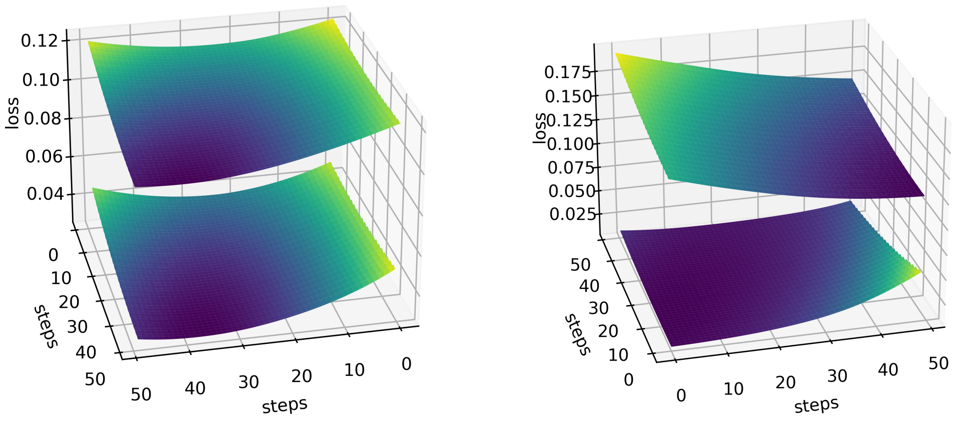


Figure 4.12: Performance on split H during training for a model without action attention.



(a)

(b)

Figure 4.13: Examples of loss landscapes for models trained without weight decay. Lower planes show the landscapes for a random training batch of size 256. Upper planes show the landscapes for the entire “adverb to verb" split. For some model-batch combinations, the two align well (left). For others, less so (right).

4.2 Discussion

In summary, I take inspiration from the cognitive sciences, specifically the concepts of selective attention, joint attention, and egocentric spatial encoding, and ask whether these mechanisms can improve an AI model’s performance on the gSCAN benchmark for systematic generalization. It is a common complaint that neural networks may produce high-accuracy outputs but do not necessarily rely on compositional rules to do so. I would argue that neural networks are not incapable of systematic generalization, i.e., the flexible composition of known parts. However, they need to receive atomic input units that are as separated from irrelevant context as possible. Otherwise, they may overfit to that context and learn solutions that only perform well on in-distribution data. Factors identified as helpful to generalization, both in the literature [GB22a] and in this study, all facilitate separating the signal from the noise – singling out important parts of the input and avoiding the memorization of unimportant parts.

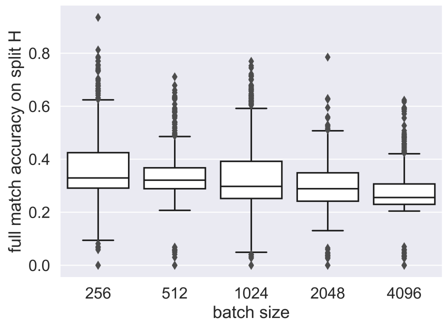
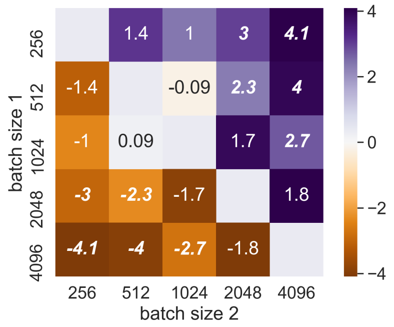


Figure 4.14: Distributions of split H accuracy sampled during training for different batch sizes.



	256	512	1024	2048	4096
256		1.4	1	***3***	***4.1***
512	-1.4		-0.09	***2.3***	***4***
1024	-1	0.09		1.7	***2.7***
2048	***-3***	***-2.3***	-1.7		1.8
4096	***-4.1***	***-4***	***-2.7***	-1.8	

Figure 4.15: Pairwise comparison of split H performance sampled during training for different batch sizes. Statistically significant z-scores $\geq |2|$ marked bold.

Concretely, the setup's selective and joint attention mechanisms help the agent focus on relevant inputs, dramatically improving its sample efficiency. The action attention part of the model, which essentially constitutes a filtering step, encourages sparser representations in the agent's controller. The egocentric location encoding saves the model from having to extract information relevant to its action prediction, namely relative position, and from potentially overfitting to absolute positions. Finally, weight decay supports generalization by serving as a kind of inductive simplicity bias [Pow+21; Kir+23].

I also find that, even without weight decay and attention, generalization performance may improve sporadically during training independent of in-distribution accuracy, especially with smaller batch sizes. However, compositional generalization encompasses a wide range of skills, and even within systematic generalization, solving one task, e.g., recombining shapes and colors, may not translate to another, e.g., recombining directions. Several gSCAN tasks remain unsolved and likely require different inductive biases than the ones presented here.

Returning to the criticisms against NNs, this work would seem to be an improvement over the first case study: It takes inspiration from the cognitive sciences and tries to go into more explanatory depth. On the other hand, the model is kept so minimal that it is perhaps even less biologically plausible than the more complex DL model used to detect patient relapses. Furthermore, I used methods like ablation studies, confusion matrices, and loss landscape visualizations to try to understand the factors that help or hinder systematic generalization – but are these the "right" techniques for explaining a model? To gain a better understanding of how this case study relates to my guiding questions, it is worthwhile taking a closer look at the notions of explanation and idealization.

4.2.1 The notion of explanation

Defining what constitutes an explanation is not entirely straightforward, as many disciplines hold different views on the matter. Broadly speaking, however, we can distinguish between five notions of explanation: deductive-nomological, inductive-statistical, causal mechanistic, unificationist, and pragmatic [CK19].

The deductive-nomological notion characterizes scientific explanation as deductive arguments where a phenomenon (explanandum) logically follows from certain premises and initial conditions (explanans). The explanans must be true and contain at least one law of nature. This aligns with the instantial view of models outlined in section 2.1.1. The inductive-statistical model is more lenient in this regard – it only requires the explanandum to follow probabilistically, rather than logically, from the explanans [Hem98]. Harkening back to the discussion of data-driven inquiry in section 3.2, an explanation need not be based on laws of nature – statistical regularity is enough. This notion contrasts with the causal-mechanistic account, which maintains that an explanation must enumerate the causal processes and interactions leading up to it [Sal20].

The unificationist view poses a more high-level requirement: an explanation should unify diverse phenomena under a joint, simple, and coherent framework [Kit89]. Finally, much like the pragmatic view of models, the pragmatic notion of explanation emphasizes that one needs to take the epistemic agent's interests, goals, and beliefs into account [Fra80]. For example, a physicist may need a deductive-nomological account, a molecular biologist may want to uncover causal-mechanistic processes, and a historian may look for unificationist explanations. But what kind of explanations might a cognitive scientist be interested in?

A common framework for understanding complex information processing systems, such as the brain or a computer, is David Marr's levels of analysis account [Mar82]. Marr distinguishes between explanations at the computational, algorithmic, and implementational level. At the computational level, the focus is on what a system is doing and why. "What"-questions may be answered by describing a system's behavior in terms of a high-level input-output mapping. "Why"-questions call for interpreting the system's behavior as a response to its surroundings. Rather than seeking answers in the "black box" itself, a computational-level "why"-explanation looks to the environment in which a behavior is learned and performed [Zed21].

The algorithmic level is concerned with how a system does what it does. "How"-explanations aim to uncover the states and transitions that govern a given behavior. Finally, the implementational or "where" level addresses where these states and transitions are physically realized – which neurons, synapses, or transistors perform a computation. Thus, the behavior of any cognizer, biological or artificial, affords explanations at different levels [Zed21; KMK19].

What all explanations, be they at the computational, algorithmic or implementational level, have in common, however, is that they must be wrong [Rud19]. They cannot be completely faithful to their target system's computations. Otherwise, the explanans would equal the explanandum and become superfluous [RL18]. To draw the connection to this chapter's opening quote, a map that is the size of a country is not very useful. In this sense, explanations bear a close resemblance to models.

4.2.2 Understanding from false models

Much like explanations, most scientific models can only ever describe their target systems imperfectly. They depart from reality in significant ways, omitting vast amounts of detail and simplifying relationships between entities. By the standards of faithful representation and causal accuracy, this would make models appear deficient as epistemic tools. At best, they are grossly distorted caricatures of a small part of reality. At worst, they are entirely fictional constructs [Gel19]. Nevertheless, scientists keep building and employing these "false" representations. What is it that enables us to derive insights from models even though they are – unavoidably – inaccurate depictions of reality?

According to one school of thought, this question rests on a flawed premise: Models are useful, not *although*, but precisely *because* they idealize and approximate. Simplification allows researchers to break down and systematically address their questions about the world [KV03]. Idealized models can be categorized into different types, of which I will here briefly discuss three: adjustable models, template models, and non-denotative models.

Adjustable models and Galilean idealization Adjustable models incorporate simplifications that can be systematically adjusted to increase their realism or accuracy [Dié15]. The kind of idealization usually involved in adjustable models is sometimes referred to as Galilean idealization [Gel19]. Galilean idealization simplifies by distortion, altering features of the target system to make the model more mathematically or conceptually manageable [Wei07]. For example, Galileo assumed perfectly round spheres and frictionless surfaces in his investigations of the behavior of bodies in motion [McM85]. This assumption constitutes a distortion of the real world, where relevant factors like drag forces are at play.

Although adjustable models start with a simplified version of the phenomenon under study, they should be amenable to de-idealization to be considered successful. There must be a realistic hope that even though they are literally "false", they can be linked to the world by adding details back in [Gel19]. For example, thanks to scientific advances like the formulation of Stokes' law, we can include frictional forces in models of bodies in motion and no longer need to resort to Galileo's perfect spheres and surfaces. However, not all idealizations should be characterized as temporary simplifications to be "fixed" down the line.

Template models and minimalist idealization As the name suggests, template models function as templates against which real-world deviations can be measured and analyzed [Dié15]. Unlike adjustable models, the idea of a template model is not to "solve for" the simple case of a real system first before moving on to more complicated cases. Instead, the template model sets up an idealized situation that does not exist in reality and looks at how the real system's behavior differs from that of the model [Wim87]. Discrepancies between the two can help identify the presence and magnitude of factors that shape the actual target system's behavior. For instance, the Hardy-Weinberg model in population genetics assumes an infinitely large, randomly mating population without selection, mutation, or migration. No actual population meets these criteria. However, the model helps scientists understand how deviations from the predicted gene frequencies can be attributed to evolutionary forces like natural selection or genetic drift [Dié15]. Template models are related to the idea of minimalist idealization.

Minimalist idealization emphasizes that selective modeling can be part of a purposeful epistemic strategy [Knu11]. Similarly to physically sealing off intervening elements in experimentation, isolating causal factors of interest in a model can allow researchers to focus on a delimited set of interrelationships [Mäk05]. Including only the core features needed to give rise to a phenomenon is often the very thing that allows a modeler epistemic access to a problem. It can help make the implications of the model's central ideas more transparent and guide the direction of further inquiry [McC09]. In contrast to Galilean idealization, minimalist idealization is not an ad-hoc measure meant to be corrected as research progresses [Knu11]. Minimal models are not impoverished representations of a target system [Gel19]. Instead, they are intended to shine a spotlight on the "essential character of the phenomenon in question" [Wei07, p. 642]. Adding details back in would defeat their purpose.

Perhaps the best-known proponent of minimalist idealization in the cognitive sciences is Randall D. Beer. Beer introduced the notion of minimal cognition, or minimally cognitive behavior, to describe the exploration of the simplest forms of behavior that still raise cognitively interesting questions [Bee96]. This concept is rooted in the idea that even small neural networks, devoid of complex computational or representational processing, can exhibit intriguing behaviors [BM23]. Beer was particularly interested in embodied intelligence and the role of an agent's environment in the emergence of cognitive phenomena [Bee21].

Using dynamical systems theory and EAs, he demonstrated how agents modeled after simple organisms like C. elegans could engage in orientation, navigation, or memorization without relying on internal symbol manipulation [BM23]. His studies contrasted with the predominant view at the time, which saw cognition as a set of discrete computational tasks performed by the brain [BW15]. Crucially, it was not Beer's goal to create evermore accurate representations of C. elegans or to graduate to more complex organisms like cats or humans one day. Instead, his idealized models were intended to demonstrate with a small set of ingredients how cognition could be understood as an emergent property of dynamic interactions between agents and their environments [Bee21].

Non-denotative models Non-denotative models involve entities, properties, or mechanisms that do not correspond to anything in the real target system [Dié15]. Out of the three types of idealized models presented here, they are the "falsest." An example of a non-denotative model is the luminiferous ether, which was once postulated as a medium for the propagation of light waves but was later shown not to exist [Mas19]. Despite their fictional nature, non-denotative models can sometimes be useful in science. They may facilitate calculations, guide research, or serve heuristic purposes. However, their contribution to genuine understanding is debatable, as they fail to refer to any real-world counterpart [Dié15]. This begs the question: How do we know whether a simplification is of the "right" kind? What distinguishes a model based on legitimate Galilean or minimalist idealization from a non-denotative one?

According to Diéguez, a good idealization is one that, despite being an inaccurate representation of reality, serves as a useful tool in scientific inquiry. He outlines several criteria that can help determine whether an idealization contributes to genuine understanding. He suggests that a good idealization should not be an oversimplification that excludes relevant functional factors, which are essential to the target system's behavior. Idealizations should not be so far removed from reality that they fail to assist in understanding how a system behaves under various causal influences or manipulations. Additionally, they should not rely on a pseudoscientific ontology, meaning they should not postulate entities or processes incompatible with current scientific understanding. The mechanisms proposed by the model should offer analogies to those operating within the real system, and the model's predictions about related phenomena should not consistently fail. These criteria aim to ensure that the idealizations facilitate a better understanding of real phenomena and are not merely subjective or misleading representations [Dié15].

4.2.3 Relation to the guiding questions

With the points above in mind, let us return to the overall questions of this thesis and the criticisms against NNs as models of cognition – starting with their supposed lack of connection to theory. At first glance, this work resembles the first case study in that it is centered around a benchmark dataset provided by a third party and focuses on predictive performance. However, there are also some major differences in the modeling process (see Figure 4.16). The first difference is the starting point of the investigation. In the first case study, the impetus came from the dataset, which prompted the search for a suitable model. In this second case study, the motivation came from the cognitive science literature, which suggested that factors like selective attention play an important role in language acquisition, paired with the knowledge that NNs struggle with systematic generalization. This hypothesis prompted the search for a suitable dataset to test whether selective attention may be helpful in compositional learning. The model can thus be said to be more rooted in theory, as it is inspired by findings from cognitive science.

A second significant difference to the first case study stems from the dataset itself. The gSCAN benchmark was created by its authors based on their assumptions about human cognition, specifically compositional reasoning abilities, and how best to operationalize these skills as a set of tasks. Ruis et al. designed these tasks to test whether NNs could solve them, and if not, to encourage exploration of "missing ingredients" that might enable systematic generalization. As gSCAN is targeted towards the ML community, solving the tasks set out in this benchmark is framed by the authors as ultimately being in service of creating better, more powerful AI applications. However, in contrast to the first study, accuracy on the dataset is not the end goal in itself. Instead, gSCAN enables the comparison of a range of models whose performance serves as an indicator of their fit with a desired, human-like behavior. Thus, both the model and the dataset in this study were, at least partially, informed by theory and created in pursuit of questions relevant to cognitive science.

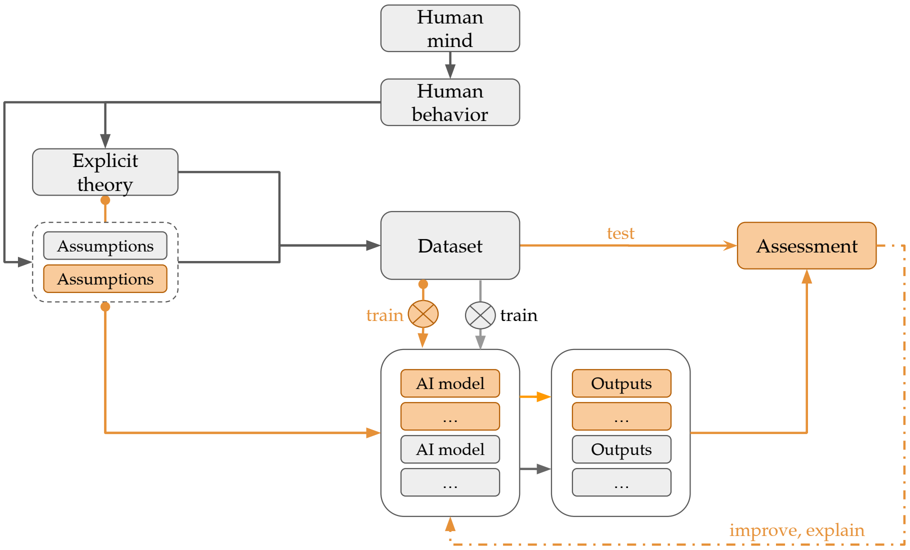


Figure 4.16: Overview of relations between human cognition, theory, assumptions, data, model, and outputs in the second case study. Relevant components performed or generated by us shown in orange. Components provided by third parties shown in gray. Circle at the beginning of an arrow indicates the starting point of investigation.

Of course, much like the contributions from the brain sciences to AI listed in the study's introduction, the translation of inspirations from cognitive science into computational form takes place at a highly abstract level. This brings us to the question of biological plausibility. The proposed model and the gSCAN dataset are highly simplified and nowhere near as complex as the brain or the environments in which we acquire language. However, as discussed in section 4.2.2, simplification need not be problematic. Minimalist idealization can be a purposeful epistemic strategy to isolate factors of interest. Previous studies on the gSCAN benchmark had focused on testing whether generic state-of-the-art DL architectures can learn compositionally. Here, I took a different approach in line with Beer's notion of minimal cognition. Namely, I identified minimal components that an agent needs to solve (most) gSCAN tasks, and I used this model to analyze when and how generalization occurs.

This bottom-up approach of building a task-specific minimal model has both advantages and downsides. On the one hand, it allowed me to conduct a wide range of ablation studies and error analyses, which would not have been computationally feasible with the more complex, bigger models proposed by others. On the other hand, it limits the study's applicability to broader contexts – the proposed model is very tailored to gSCAN. Still, although the architecture is specific to the dataset at hand, the factors contributing to its performance are consistent with related work on systematic generalization and likely to also apply to other situations. While the simplifications in this work were, in this first step, meant to be minimalist idealizations, they could conceivably be adapted to more complex problems in the spirit of Galilean idealization. Indeed, I make use of the idea of selective attention in later case studies.

To sum up, while the model in this work is biologically inspired at a very high level, it is fair to say that it is not biologically plausible. However, it is precisely *because* of the model's high degree of idealization that I can more easily isolate and systematically vary or ablate factors of interest, namely those that contribute to compositional generalization. This brings us to the third criticism against NNs: their inability to offer explanations.

As discussed in section 4.2.1, the notion of explanation is complex. It is difficult to say when something has been sufficiently explained because, at least according to the pragmatist perspective, an explanation's success depends on the individual epistemic agent and the level of explanation in which they are interested. However, this second work arguably offers more in-depth explanations than the first by any definition. If the first case study can be said to provide explanations, they are at the relatively abstract "why" and "what" levels of the Marr hierarchy.

This second work presents a broader range of explanations. On the "what" level, and related to the inductive notion of explanation, it looks at model inputs and outputs, including statistically analyzing error patterns. At the "why" level, the study tries to determine the influence of certain factors in the agent's learning environment, e.g., by devising custom train-test splits or running ablation studies of the auxiliary "joint attention" signal. It investigates some algorithmic or "how" level questions by ablating model components like action attention or selective attention. It additionally offers a "where" level analysis of the specific neurons in the agent's controller that contribute to its predictions.

Finally, it relates the results of these investigations to previous findings on out-of-distribution generalization, venturing into the direction of unificationist explanation. Specifically, it contextualizes previous proposals on how to enhance compositionality in NNs as methods to encourage the separation of relevant from irrelevant parts of inputs. On a meta-level, benchmarking different models on a dataset like gSCAN can also serve as a stepping stone toward explanation. Comparing different models may reveal relevant success factors and help select models that constitute promising candidates for further inquiry [CK19].

Overall, we can conclude that NNs can be explicitly connected to theory by taking inspiration from the cognitive science literature in designing training environments, architectural constraints, forms of regularization, or augmented loss functions. As shown in this case study, this not only allows for the investigation of questions relevant to cognitive science but may, in some cases, also improve NN performance [Has+17]. NNs can be explained at various levels, and benchmarking the predictive performance of different models can serve as a stepping stone to explanation. In the pragmatic view, whether an explanation is successful always depends on the individual interests of the epistemic agent.

Furthermore, NNs need not be biologically plausible to be of value. As long as we do not stray too far from reality or postulate entities contradicting current scientific understanding, minimalist or Galilean idealization can be a viable strategy to facilitate epistemic access to a model and isolate factors of interest [Dié15]. We should, however, transparently communicate how such simplifications impact what we can conclude from an investigation [McC09]. The success of specialized models on artificial tasks in a benchmark like gSCAN does not mean that systematic generalization has been "solved" by DL. To use a metaphor from this chapter's epigraph: maps help guide us by omitting irrelevant details, but it is important not to confuse the map with the territory.

As mentioned in the introduction of this thesis, there are two facets to the criticism that NNs do not offer explanations: The issue of understanding the models themselves and the issue of understanding a target system through these models. This second case study tried mainly to understand the model and its behavior. In the following section, we will look at how a NN can be used to explain a real-world phenomenon.

5 Modeling the emergence of letter shapes with drawing-based signaling games

> *The science fiction method is dissection and reconstruction. You look at the world around you, and you take it apart into all its components. Then you take some of those components, throw them away, and plug in different ones, start it up and see what happens. That's the method: restructure the world we live in in some way, then see what happens.*
>
> Frederik Pohl

5.1 Study

The third case study relates to core knowledge of object geometry (see Figure 5.1). According to Spelke, this innate system underlies our processing of visual properties, such as shape or symmetry [Deh+06]. This notion is supported by several studies that have used so-called "intruder tasks", where subjects must find which of a set of shapes is different, to show that intuitions of geometric regularity are present in all human groups regardless of age, education, and culture [Sab+21]. We approach the topic through the lens of analyzing letter shapes, which can be seen as cultural artefacts that reflect these human geometric preferences.

Specifically, we create artificially evolved writing systems by employing a drawing-based signaling game involving two AI models. We then explore how different design choices impact empirical regularities in the surface form of these artificial glyphs and their similarity to human-created visual signs. Our goal is to explore *in silico* factors that have been hypothesized to influence the shapes of letters in human writing systems, as their emergence cannot be studied *in vivo* except retrospectively. In our first experiment, we investigate the role of the models' perception system on glyph line orientation and symmetry. We find that these characteristics are impacted by the input statistics of data used to pre-train models and, to a lesser extent, canvas shape and architectural model properties. Our second experiment analyzes the grapho-phonemic mapping that emerges when we integrate representations learned by a DL model trained for speech conversion into our setup.

A version of this study was presented at the 46th Annual Meeting of the Cognitive Science Society and published in its proceedings [HD24b].

5.1.1 Introduction

Writing is an ancient cognitive technology that has been invented independently several times in the course of human history [Mor22b]. Even before fully-fledged writing systems, humans have produced geometric signs since at least the Paleolithic [Dut+20]. Curiously, graphic codes across times and cultures consistently share certain characteristics. Specifically, glyphs appear to reflect the input statistics to which our visual system has adapted. Letters tend to

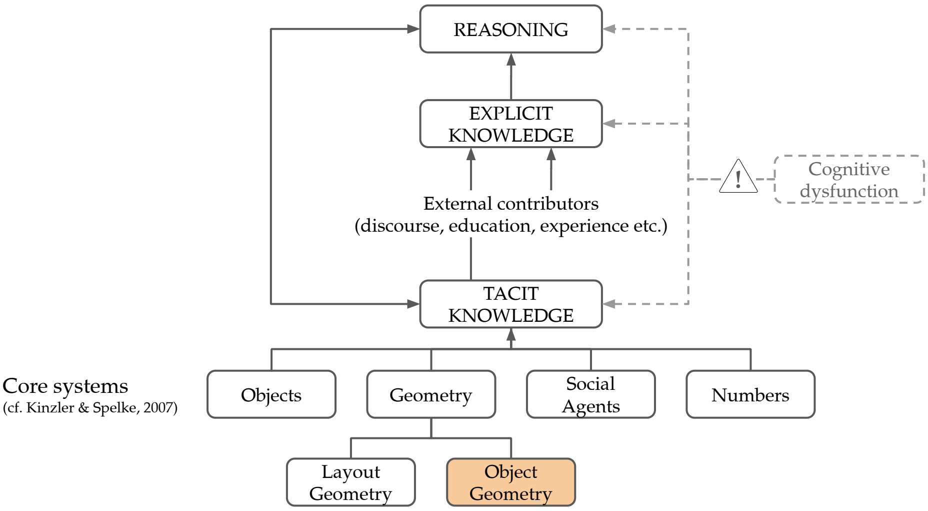


Figure 5.1: Situating the third case study in the broader study of cognition. Relevant parts of the framework marked in orange.

display a disproportionate rate of vertical symmetry, which is a feature of, e.g., human faces or standing bodies [Mor18] and they extensively comprise topological signatures found in natural scenes [Cha+06; TSZ17]. Furthermore, vertical and horizontal strokes are over-represented compared to obliques [Mor18]. This cardinality preference has been attributed to our visual acuity being better for vertical and horizontal lines than for other orientations – the so-called "oblique effect" [App72]. Finally, there is a tendency towards simplicity, as complex characters are more effortful to read and produce [Lin+19; MM21]. These findings support an ecological hypothesis that signs have evolved to accommodate human visual perception.

Several studies have investigated the emergence of graphical conventions using signaling games where participants communicate through drawing [Gal05; Gar+07; Fay+10; BDL13; RLG15; Fay+18; Fan+20; Haw+23]. Their primary focus has been on the role of social context in the construction of sign systems and the trade-off between iconicity and abstraction. A smaller number of works develop AI models for generating sketches. Most of these models are trained to convert images into simplified drawings [HE18; Muh+18; Son+18; Cao+19; MH21; Vin+22; Qiu+22] or to play collaborative or Pictionary-type games [FDH19; Bhu+20].

Similar to previous work, we follow a drawing-based signaling game setup involving two AI models. However, the sketches produced by the models represent pre-defined classes rather than aiming to depict a given image faithfully. Thus, our goal differs in that we do not focus on iconicity, i.e., the visual resemblance between a sign and its referent. Instead, we are interested in how different design choices impact empirical regularities in the surface form of artificially evolved glyphs and their similarity to human visual signs. We explore this question using two experiments. The first experiment investigates aspects of the receiver model's perception system that impact glyph stroke orientation and symmetry in abstract graphic codes. The second experiment takes a step towards modeling orthography by introducing an aural dimension and a notion of sender-side motor effort minimization.

Table 5.1: Summary of the sender model's architecture. C is the number of classes, which varies by dataset.

Layer	Type	Input Shape	Output Shape	Activation	Stride	Padding
1	Conv2d	$1 \times 64 \times 64$	$64 \times 32 \times 32$	LeakyReLU	2	1
2	Conv2d	$64 \times 32 \times 32$	$128 \times 16 \times 16$	LeakyReLU	2	1
3	Conv2d	$128 \times 16 \times 16$	$256 \times 8 \times 8$	LeakyReLU	2	1
4	Conv2d	$256 \times 8 \times 8$	$128 \times 4 \times 4$	LeakyReLU	2	1
5	Conv2d	$128 \times 4 \times 4$	$C \times 1 \times 1$	Identity	1	0
6	Flatten	$C \times 1 \times 1$	C	-	-	-
7	LogSoftmax	C	C	-	-	-

5.1.2 Approach

Our setup consists of a sender and a receiver. The receiver is a visual model – specifically, a five-layer CNN. An architectural overview can be found in Table 5.1. We use a kernel size of 4×4, batch normalization, no bias, and a leaky ReLU activation with negative slope 0.2 for the convolutional layers. The sender is a simple linear model that generates a graphic code consisting of a pre-defined number of glyphs. The criteria that these glyphs should fulfill vary by experiment. The sender can place up to three lines per glyph on a 64×64 canvas. We chose the number three because it is the average number of strokes per character across many writing systems [Cha+06]. Each line is a quadratic Bézier curve defined by the x and y coordinates of its start, control, and endpoint. The sender thus has to optimize six parameters per glyph and stroke.

As in the second case study, these parameters are optimized using the CMA-ES algorithm [HO01b]. CMA-ES has been found empirically to outperform other black box optimizations in a range of applications [Han+10], including activation maximization in CNNs [WP22], which is closely related to our experimental setup. It also has the benefit of being quasi parameter-free and allowing us to define non-differentiable loss functions, which is not the case for gradient estimation methods such as backpropagation.

Our approach is inspired by Park, who explores a similar setup of a CNN receiver and a sender drawing a set of abstract glyphs with Bezier curves [Par20]. However, we use different model architectures, loss functions, and optimization algorithms. Our work also diverges in scope in that Park presents a technical proof-of-concept mainly focused on aesthetics. We significantly expand on their proposal by systematically analyzing the generated codes, contextualizing them in cultural evolution research on human writing systems, and, in experiment 2, introducing an aural dimension.

5.1.3 Experiment 1

In our first experiment, we build a computational model of the hypothesis that letters evolved to reflect the statistics of natural visual inputs. We pre-train receivers on different image datasets and measure the effect on the symmetry and cardinality, i.e., propensity for vertical and horizontal strokes, of glyphs produced by the sender.

Table 5.2: Composition of the NAT dataset.

#	class	source
5,666	natural landscape	15-Scene [LSP06]: `coast`, `forest`, `mountain`, `open country` Flickr [CLL18]
5,500	face	CelebA 64x64 [Liu+15]
5,500	plant	ImageCLEF 2013 [Goë+13]
5,399	animal	Animal Image [Ban23]

Table 5.3: Composition of the H-M dataset.

#	class	source
2,200	urban landscape	15-Scene [LSP06]: `street`, `suburb`, `living room`, `office`, `industry`, `building`, `inside city`, `highway`
2,200	motorcycle	COCO 2017 [Lin+14]
2,200	airplane	COCO 2017 [Lin+14]
2,200	wine glass	COCO 2017 [Lin+14]
2,200	bowl	COCO 2017 [Lin+14]

Datasets Inspired by Changizi et al., we train one model on "natural" images (henceforth NAT) and one on images of urban landscapes and human-made objects (henceforth H-M) [Cha+06]. We also include a randomly initialized, untrained CNN for comparison. The composition of the NAT and H-M datasets can be found in Tables 5.2 and 5.3, respectively. We resize the shortest side of each image and apply centered crops to obtain 64×64 inputs, which we gray-scale and normalize. Any images containing text were removed to prevent exposure to human writing systems. We use 80% of the data for training and 20% for validation. We also create a dataset of human-made scripts to compare their visual characteristics to those of our evolved glyphs. This dataset is based on the collection of 116 writing systems analyzed by Morin [Mor18]. We generate one image per glyph using a consistent font (Noto Sans). Loma, Woleai, Kpelle, and Afak scripts were omitted as they are not yet part of the Unicode codespace.

Receiver Model All receiver models' architectures are identical (shown in Table 5.1), except for their output dimension C. C is four for the NAT and random models and five for the H-M dataset. NAT and H-M models were trained for 200 epochs on image classification, using the Adam optimizer [KB15a], negative log likelihood loss, and a batch size of 64. The final validation accuracy after early stopping was 86% for both models. Note that receiver models are not updated further during their interaction with the sender model to avoid overfitting to the produced glyphs.

Sender Model The sender is tasked with developing a graphic code with 25 glyphs, which should be perceptually distinct to the receiver. More specifically, it maximizes the distance between the activations elicited in the receiver by the different glyphs. The sender thus optimizes for what S. Qiu et al. term *symbolicity*, i.e., consistent separability in high-level visual embedding space [Qiu+22]. Let A represent the activations in the receiver's convolutional layers. We use embeddings from the receiver's last layer for most experiments. Each activation vector a corresponds to a different glyph. The sender's loss function aims to maximize the L2 norm between each activation and its closest neighbor:

$$\text{Loss} = \frac{1}{|A|} \sum_{a_i \in A} \min_{\substack{a_j \in A \\ a_j \neq a_i}} \|a_i - a_j\|_2 \tag{5.1}$$

For the CMA-ES optimization of sender models, we use a population size of 32, uniform random solution initialization, and an initial standard deviation of 0.05. We let models run for 1300 iterations and train ten models per setting. We report averages across these ten models.

Metrics To measure the orientation of glyph strokes, we use Histogram of Oriented Gradients (HOG) [DT05]. HOG is a computer vision feature descriptor that splits an image into a grid of cells. For each pixel in the cell, intensity gradients, i.e., edge directions, are computed, binned into orientations, and counted to obtain a histogram. Contrast normalization may be applied block-wise for better invariance to lighting changes. An example is shown in Figure 5.2. HOG is traditionally used for tasks such as object detection. We here re-purpose it as an automated alternative to the manual coding by which letter cardinality has previously been analyzed [Mor18]. We use 12 orientations, cells of size 16 × 16, three cells per block, and L2 normalization. Before applying HOG, we resize images to 128 × 128 pixels and apply a Gaussian blur of size 2 to avoid square pixelation artefacts that would artificially increase cardinal dominance.

To measure glyph symmetry, conceptually, we place an axis through an image at each angle between 0° and 179°, mirror it along that axis, and record the overlap for each angle. In practice, we rotate the glyphs in SVG space by each angle between 0° and 179°. We pad images to prevent parts of the glyph from rotating out of the picture at certain angles. We then flip the rotated image vertically, sum-normalize the rotated and flipped rotated images, and convolve the two via the fast Fourier transform method. Intuitively, this auto-correlation corresponds to moving the flipped rotated image over the rotated image and computing the overlap at each point. We use the maximum value of the convolution as a measure of the highest overlap, i.e., symmetry, at a given angle. Figure 5.3 shows an example of the process outlined above.

Note that our metric is a more continuous measure of symmetry than used by Morin [Mor18]. In their work, symmetry was coded manually, and only wholly symmetric letters were considered. Our Bézier curve-based glyphs are more akin to handwriting than standardized letters and contain a higher degree of noise, e.g., small shifts or rotations. We propose our automated measure as a way still to capture symmetric regularities in such cases.

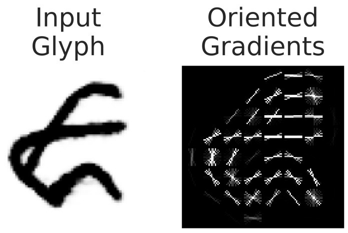


Figure 5.2: Oriented gradients example.

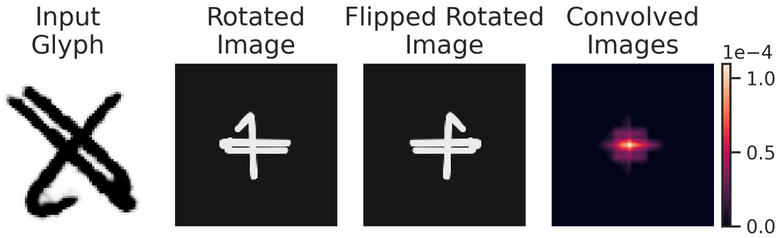


Figure 5.3: Illustration of the steps involved in our symmetry measure. The example shown is for an angle of 137°.

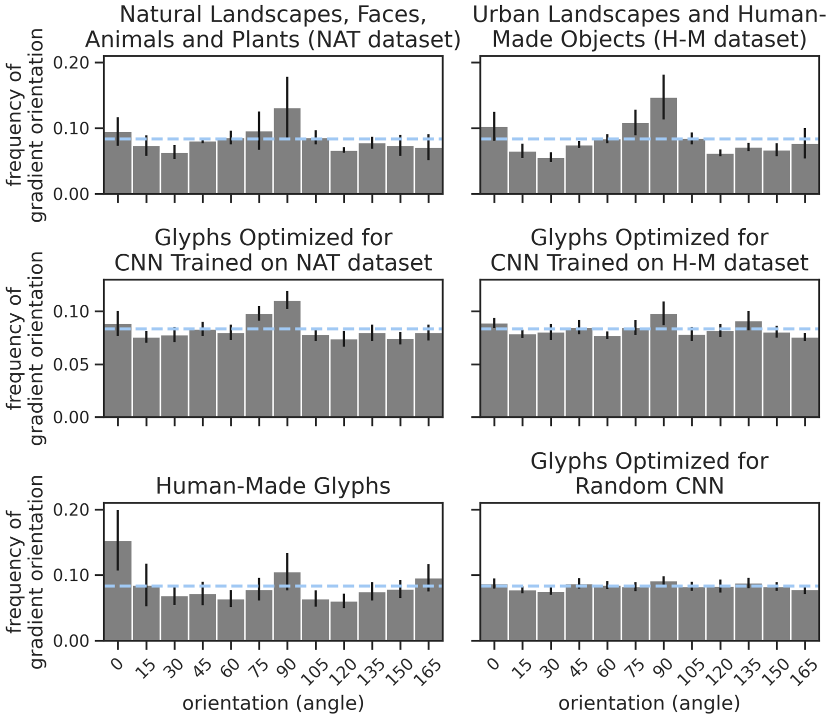


Figure 5.4: Histograms of oriented gradients for the pre-training datasets, artificially evolved graphic codes, and human writing systems. Blue dotted line marks expected frequency in the uniformly random case.

Results: Stroke Orientation Figure 5.4 shows that vertical and horizontal orientations are most common in both the NAT and H-M dataset, consistent with previous analyses [Cop+98; Cha+06; GLS11]. Pre-training on this data promotes a preference for cardinality: Particularly, gradients near 90° occur with above-average frequency in glyphs evolved for the pre-trained receivers. In contrast, glyphs optimized for a random CNN show a nearly uniform gradient distribution. The correlation between orientation statistics of the pre-training dataset and optimized glyphs is stronger for NAT ($R = 0.92$, $p < 2 \times 10^{-5}$) than for H-M ($R = 0.68$, $p < 2 \times 10^{-2}$), likely because NAT models were exposed to more training data.

Although the tendency towards cardinality is less pronounced than in human-made letters, we find a moderate correlation between orientation characteristics of evolved and human-made glyphs ($R = 0.44$, $p < 0.15$). Overall, the results support the notion that optimizing for a visual system that has been exposed to natural input statistics can give rise to the preferred line orientations observed in human writing systems. Analogously to the mechanisms thought to have shaped human letters, CNN units will be more attuned to common orientations in their training set [HS21], which the sender may, in turn, exploit to optimize discriminability.

Results: Symmetry We now turn our attention to another aspect of anisotropy: Glyph symmetry. Figure 5.5 shows that, consistent with human preferences, there is an above-average tendency towards vertical symmetry in our evolved glyphs, particularly for the NAT setup. This result is perhaps to be expected from our HOG analysis, as cardinal lines tend to be symmetrical. However, interestingly, even glyphs evolved for random CNNs show a slight above-chance symmetry along 45° and 135° angles despite not having been exposed to any training that could explain this preference.

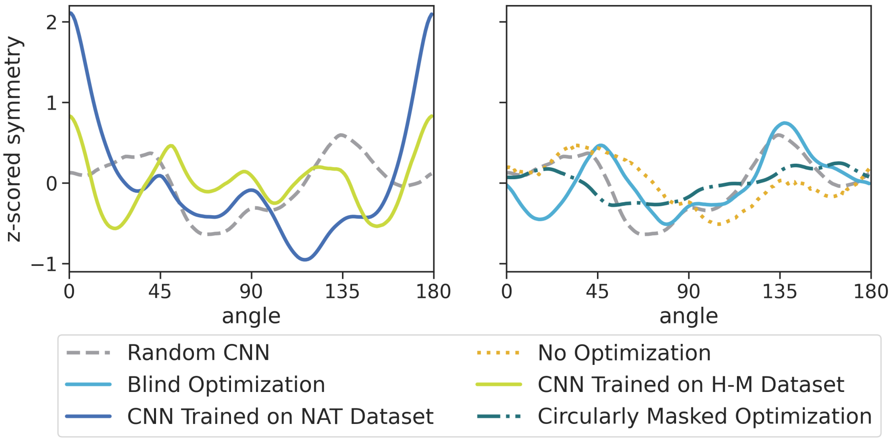


Figure 5.5: Z-scored symmetry distribution of artificially evolved graphic codes.

We considered three possible explanations for this phenomenon: 1) a bias introduced by the CMA-ES algorithm, 2) a bias introduced by the canvas shape, and 3) an inductive bias of the CNN architecture. To test each of these options, we plotted results without any optimization (creating glyphs via uniform random sampling), optimization with a circular mask (resampling any time CMA-ES suggests a solution containing points outside the circle), and blind optimization (setting loss to constant 0). Figure 5.5B shows that, without any optimization, there are still peaks at 45° and 135° angles. This speaks against option 1. Blind optimization closely resembles that for the untrained receiver, suggesting that the random CNN's feedback, rather than containing some hidden preference, is basically arbitrary. This contradicts option 3.

However, the symmetry preferences disappear when using circularly masked optimization, confirming option 2. Considering that the square canvas is uniformly sampled, a slight over-representation of points in the four corners will implicitly promote symmetry at 45° and 135° angles as measured by our "overlap" metric. This result relates to the role of physical constraints imposed by writing materials on the evolution of human scripts. E.g., rectangular canvases have been considered as a potential cause of cardinal dominance in paintings [Mil07], and Indian and Southeast Asian scripts are thought to be less angular because they were written on flexible leaves [Wat94].

Besides pre-training and canvas shape, we identified an additional factor influencing symmetry: the layer of the CNN receiver used to calculate equation 5.1. Figure 5.6 shows examples of how optimized glyphs change qualitatively across layers. When we compare the symmetry of the produced glyphs (Figure 5.7), it is interesting to see that, even in glyphs optimized for random CNNs, there is a higher level of symmetry at 0°, 90°, 180°, and to a lesser degree at 45° and 135° angles. The tendency is less pronounced for the highest layer. We hypothesize that this symmetry emerges due to the square nature of the receiver's convolutional filters, which partition the input image into overlapping patches. From the sender's perspective, each patch essentially represents a separate noisy channel [Sha48].

Given that the sender is limited to contiguous strokes in utilizing these channels, it will tend to connect grid neighbors, resulting in increased cardinal and oblique symmetry for a grid of squares. Symmetry is less pronounced in the last layer because this layer's receptive field spans the whole image (see Table 5.1), limiting the sender to a single, global channel. To further illustrate the described effect, we create a setup that mimics the mechanism of CNN filters in a

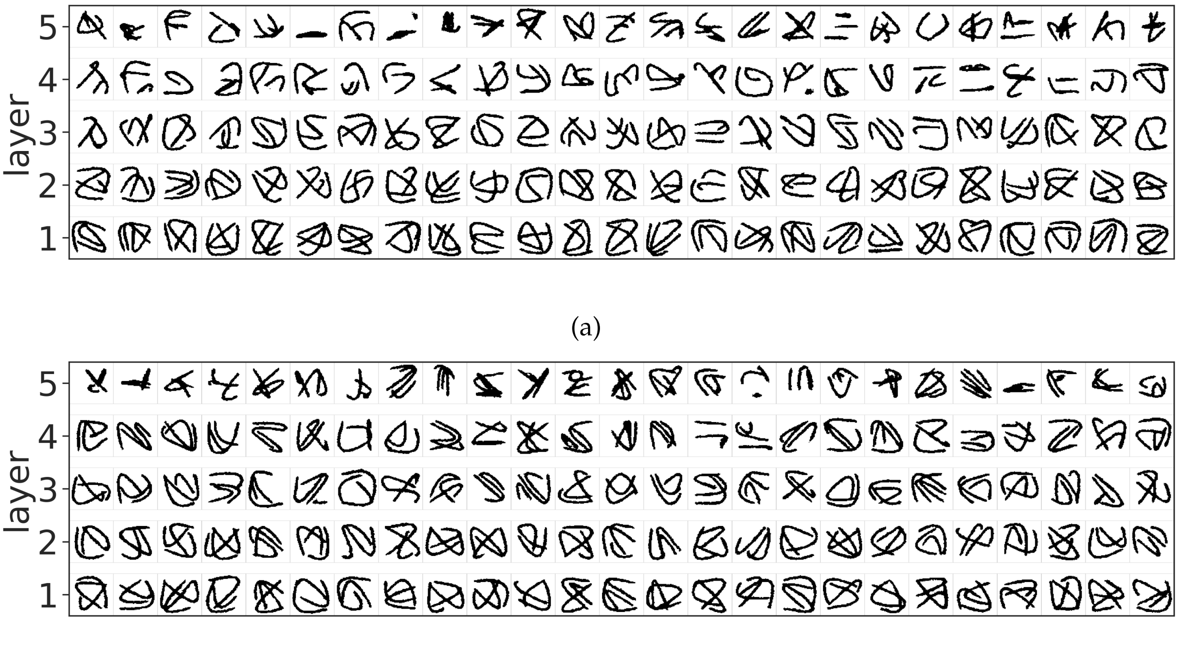


Figure 5.6: Exemplary graphic codes evolved for the different convolutional layers of receivers pre-trained on "natural" images (a) and images of urban landscapes and human-made objects (b).

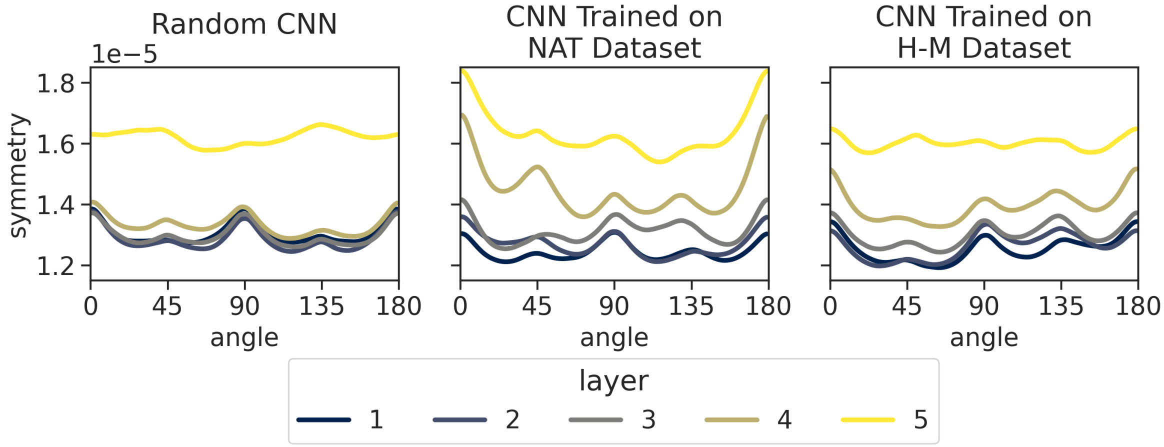
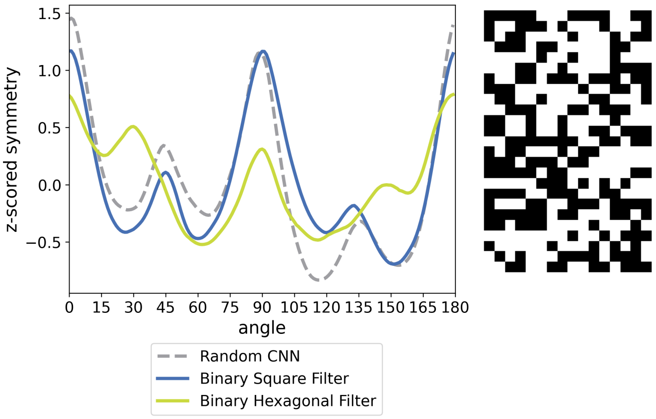


Figure 5.7: Comparison of glyph symmetry for codes evolved for different receiver layers.



(a) (b)

Figure 5.8: Symmetry of graphic codes evolved in binary grid set-ups (a) and binary representation of an evolved code (b).

simplified way. We split canvases into 4 × 4 patches and define the signal communicated to the sender as a 16-dimensional vector with binary entries: 0 if a patch contains no ink in a square and 1 if it does. We then compare the effect of using a square and a hexagonal grid, the latter implemented as proposed by Steppa et al. [SH19].

As can be seen in Figure 5.8a, the resulting symmetry distribution for the square grid setup highly correlates with that of the random CNN ($R = 0.91$, $p < 8 \times 10^{-72}$). In the hexagonal case, the angles of a cell's neighbors change, increasing symmetry at 30° and 150°. Thus, although CNNs carry the added complexity of overlapping filters and multiple channels per filter, the results from our simplified setups lend credence to the proposed explanation for the symmetry behavior in lower CNN layers.

5.1.4 Experiment 2

The glyphs produced in experiment 1 are far from a full-fledged writing system. They can be considered closer to emblems, i.e., graphic codes that do not encode a language and lack the productivity of specialized codes such as musical notation [MKW20]. A graphic code is only considered writing if it represents words, morphemes, or phonemes [Mor22b]. By encoding spoken units of meaning, writing vicariously inherits the generality of language: We can express anything we could say in written form (glottographic principle). In our second experiment, we take a step towards modeling the development of such a glottographic code representing language. Our setup differs from experiment 1 in two regards: The first is that the semantics of the evolved glyphs change from abstract classes to representing linguistic information. The second is that we add constraints designed to mimic the evolutionary pressure towards reduced effort in producing and processing writing.

Speech Model Linguistic information is incorporated using a DL model trained for speech conversion, i.e., transforming source speech to a target voice without changing the content. This task resembles writing in that the goal is to communicate content in a standardized form, abstracting away speaker-specific acoustic features. The specific model we use was proposed and implemented by Niekerk et al. [Nie+22]. It works by extracting features from HuBERT [Hsu+21], a widely used Transformer-based speech model. HuBERT is pre-trained in a self-supervised manner on LibriSpeech-960 [Pan+15].

Niekerk et al. apply k-means ($k = 100$) to the intermediate representations of HuBERT's 7th layer and train an acoustic model to decode the resulting clusters into output speech. However, we ignore this decoder here and simply use the 100 clusters, which we will henceforth refer to as *units*. Note that these units are not explicitly trained to map to phonemes, morphemes, or syllables. They are simply representations learned by the model to optimally fulfill its speech conversion task. Note also that the model has only been trained to predict speech in spectrogram space, not to transcribe it. Its representations have thus not been shaped by exposure to English orthography.

Sender Model As in experiment 1, the sender model must optimize a graphic code, this time containing 100 glyphs representing the speech model units. However, instead of creating maximally distinguishable glyphs, it is tasked with using as few strokes as possible while still ensuring what S. Qiu et al. term *semanticity*, i.e., visually preserving the topology of the speech model's latent space [Qiu+22]. These constraints reflect two competing pressures on writing systems: transmission efficiency and referential efficiency [Mor16; Mor18; Kel+21]. We model them with a frequency penalty and a similarity constraint, respectively.

We allow the sender to draw up to $L_{\max} = 3$ strokes. To model transmission efficiency, we add a penalty P. P calculates how many strokes l were used for a glyph, multiplied by the relative frequency f with which the unit it represents occurs in natural speech. This term reflects the finding that, across writing systems, more frequent characters consistently have a lower degree of complexity [MM19; KMM23]. We collect frequency statistics by applying the speech conversion model to the Flickr 8k Audio Caption Corpus [HG15] and recording how many times each speech unit occurs. The penalty is weighted by a factor α, here set to $\frac{1}{2}$.

$$P_i = \alpha \times \left(1 - \frac{l_i}{L_{\max}}\right) \times f_i$$

For the similarity constraint, we calculate the pairwise L2 distance d between the centers of the $N = 100$ units in the speech model's activation space A. We do the same for the images of the 100 glyphs in the visual receiver model's embedding space. We normalize each row in the distance matrices and subtract it from 1 to obtain a measure of similarity s:

$$s_{ij} = 1 - \frac{d_{ij}}{\max_i d_{ij}}, \quad d_{ij} = \|A_i - A_j\|_2$$

We then minimize the mean absolute distance between the two similarity matrices S^{vis} and S^{speech}. The reasoning behind this is that characters that look similar tend to have similar canonical pronunciations across orthographies [JTS22]. The combined loss function reads as follows:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}|S_{ij}^{\text{vis}} - S_{ij}^{\text{speech}}| \times (1 + P_i)\right)$$

Given the added complexity of the problem, we increase our CMA-ES population size to 64 and the number of iterations to 30,000. We initialize solutions to 0.5 with a standard deviation of 0.1 and use a NAT CNN from experiment 1 as our receiver, with activations taken from layer 4. To mimic the variability inherent to handwriting, we add Gaussian noise to the solutions with a standard deviation of $0.005^{\frac{1}{2}}$.

Association Rule Mining In this experiment, we are interested not only in the surface form of the glyphs but also in the kind of grapho-phonemic mapping the models produce. We, therefore, analyze the co-occurrences of glyphs and speech units using TIMIT [Gar+93], a standard dataset used to evaluate automatic speech recognition systems. TIMIT is designed for broad phoneme coverage and includes rich word and phoneme-level annotations. For each sentence in the corpus, we collect the phoneme annotation and the units produced by the speech model during processing. Having collected co-occurrences, we apply the Apriori algorithm [AS94] to identify association rules. We generate all rules with a minimum support of 10^{-4}, minimum confidence of 0.2, and minimum lift of 3. Support here refers to the relative frequency of a unit-phoneme pairing, confidence refers to the conditional probability of a unit given a phoneme or another preceding unit, and lift refers to the ratio between confidence and support [AS94].

Results We show co-occurrences between phonemes and glyphs, representing HuBERT speech units, in Figure 5.9. Table 5.4 lists the phonemic and phonetic notation used in the graph. The visualization illustrates that the similarity constraint of our loss function works as intended: phonetic similarities are reflected visually. For instance, glyphs correlated with the phonemes s (sea), sh (she), and z (zone) look alike. Similar phonemes often share a glyph, see, e.g., y (yacht) and iy (beet), or b (bee) and p (pea). Interestingly, some glyphs co-occur with common phoneme combinations, e.g., "ix ng/ix n", as used in the English present participle form, or the word "sh iy" (she). This mapping has some correspondence with human scripts in that no writing system consistently follows a single organizing principle [Mor22b]. I.e., while some scripts may be predominantly syllabic, alphabetic, or logographic, no system falls into purely one category [Mat92].

Table 5.4: Phonemic and phonetic symbols from the TIMIT lexicon [Gar+93].

	Symbol	Description
	pau	pause
Others	epi	epenthetic silence
	h#	begin/end marker (non-speech events)

	Symbol	Example word	Possible phonetic transcription
	iy	beet	bcl b IY tcl t
	ih	bit	bcl b IH tcl t
	eh	bet	bcl b EH tcl t
	ey	bait	bcl b EY tcl t
	aa	bott	bcl b AA tcl t
	aw	bout	bcl b AW tcl t
	ay	bite	bcl b AY tcl t
	ah	but	bcl b AH tcl t
Vowels	ao	bought	bcl b AO tcl t
	oy	boy	bcl b OY
	ow	boat	bcl b OW tcl t

Continued on next page

Table 5.4: Phonemic and phonetic symbols from the TIMIT lexicon [Gar+93]. (Continued)

	uw	boot	bcl b UW tcl t
	ux	toot	tcl t UX tcl t
	er	bird	bcl b ER dcl d
	ax	about	AX bcl b aw tcl t
	ix	debit	dcl d eh bcl b IX tcl t
	axr	butter	bcl b ah dx AXR
Semivowels and Glides	l	lay	L ey
	r	ray	R ey
	w	way	W ey
	y	yacht	Y aa tcl t
	hh	hay	HH ey
	hv	ahead	ax HV eh dcl d
	el	bottle	bcl b aa tcl t EL
Nasals	m	mom	M aa M
	n	noon	N uw N
	ng	sing	s ih NG
	en	button	b ah q EN
	nx	winner	w ih NX axr
Affricates	jh	joke	DCL JH ow kcl k
	ch	choke	TCL CH ow kcl k
Fricatives	s	sea	S iy
	sh	she	SH iy
	z	zone	Z ow n
	f	fin	F ih n
	th	thin	TH ih n
	v	van	V ae n
	dh	then	DH e n
Stops	b	bee	BCL B iy
	d	day	DCL D ey
	g	gay	GCL G ey
	p	pea	PCL P iy
	t	tea	TCL T iy
	k	key	KCL K iy
	dx	muddy, dirty	m ah DX iy, dcl d er DX iy
	q	bat	bcl b ae Q

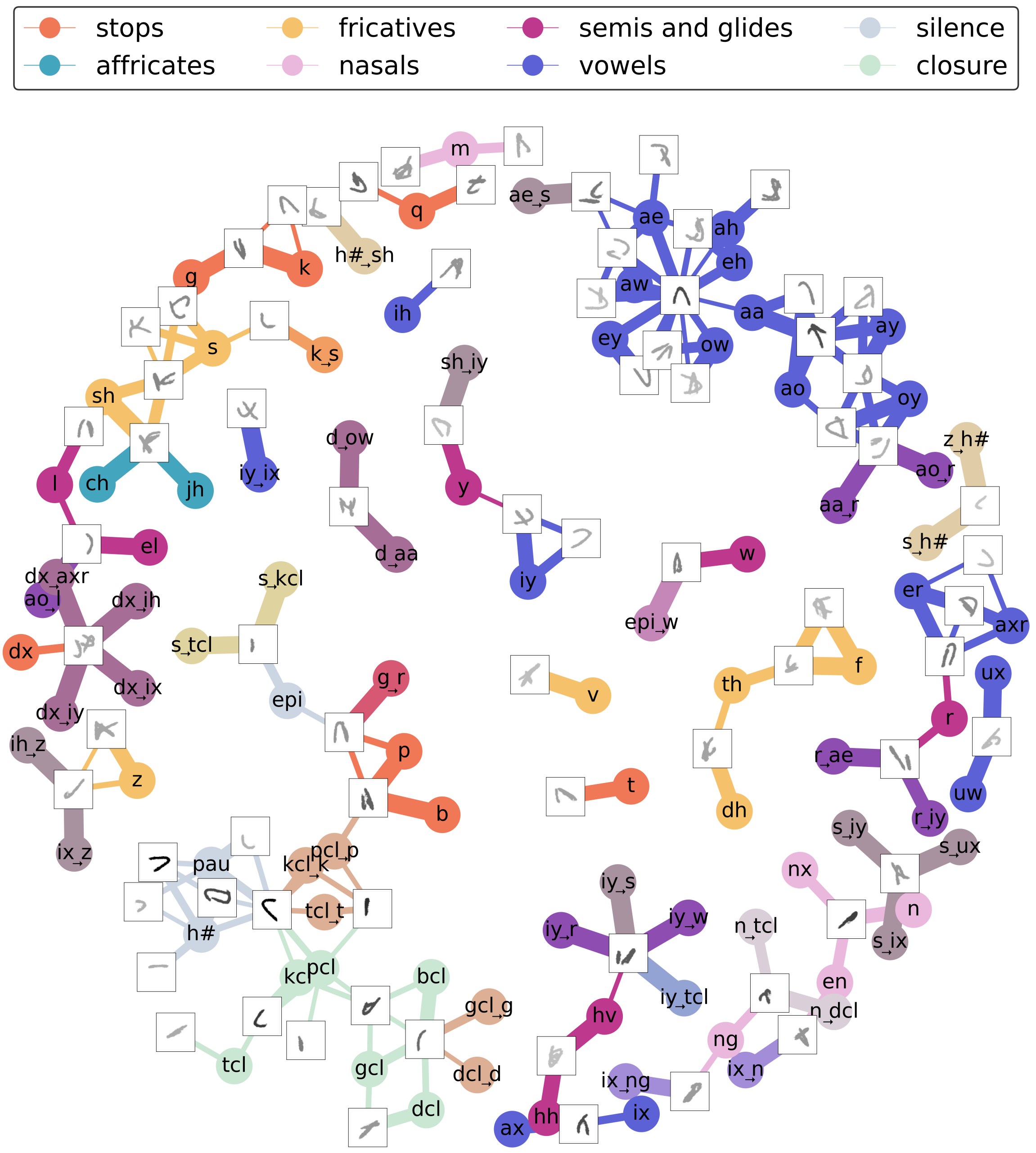


Figure 5.9: Association rules for glyphs and phonemes, encoded using TIMIT notation. For nodes representing transitions (→), colors corresponding to the individual phonemes' categories were combined. Edge width represents rule confidence. Glyph opacity represents frequency of occurrence.

In addition to preserving phonetic similarity, the evolved glyphs successfully reflect transmission effort. Frequent glyphs, such as those representing closures, pauses, or the sentence marker h#, tend to be simple, often consisting of a single line (see Figure 5.10). The bulk of the glyphs evolve to contain between one and two strokes, with only a few low-frequency glyphs encoded using three. One high-frequency glyph is effectively a space, i.e., it has zero strokes. Note that this glyph is not contained in Figure 5.9 as we only show rules with a minimum confidence of 0.2. Consistent with "Zipf's law of meaning" [Zip49], the high-frequency glyph in question co-occurs with many different phonemes, diluting its association rules' confidence.

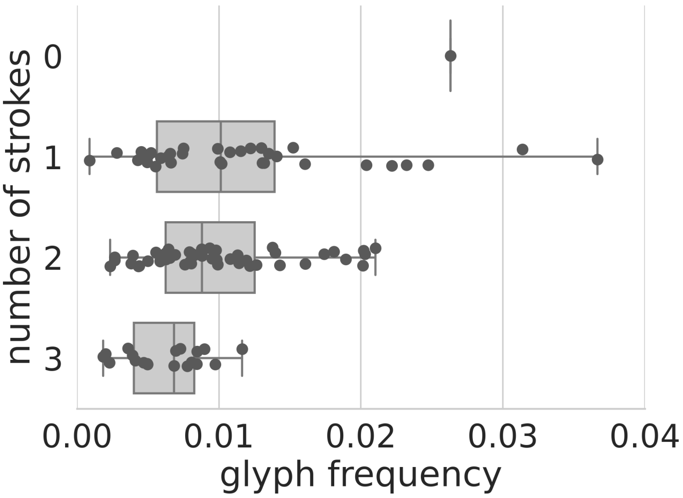


Figure 5.10: Glyph frequency vs. complexity for speech-based graphic code.

5.2 Discussion

In summary, I find that, as predicted by the ecological hypothesis of letter shapes, glyphs that are evolved for models pre-trained on images reflect the statistics of their input data and display anisotropy consistent with human-made glyphs. I also observe that the square nature of the canvas and the receiver's convolutional filters impact glyph symmetry, albeit to a smaller extent. I then integrate representations learned by a pre-trained speech model as well as efficiency pressures into my setup. The resulting code yields a hybrid orthography and shows a Zipfian effect for glyph complexity.

Compared to the second case study, predictive performance plays a very minor role in this work, and I make almost no attempts at "how" or "where" level explanations of the models I use. Instead, I focus on analyzing model outputs and try to draw conclusions about a real-world phenomenon, namely, the emergence of statistical regularities in the shapes of human-created glyphs. This approach raises several questions: Is it possible to understand a phenomenon through a model without a detailed understanding of the model itself? And if so, given that areas like cultural evolution research seldom lend themselves to provable, definitive answers, what kind of explanation could we hope to gain from models such as the one in this study? In the following sections, I want to provide some background on a few topics that I consider relevant to answering these questions: how-actually vs. how-possibly explanations, the modal dimension of modeling, and the idea of toy models.

5.2.1 How-actually and how-possibly explanations

As mentioned in the introduction to this thesis, there are two aspects to the explainability problem: The explanation *of* NNs and the explanation *by* NNs. Often, explaining the models themselves is seen as a prerequisite for explaining a target phenomenon through them. However, Sullivan, for instance, argues against the common notion that the "black box" nature of NNs inherently limits our ability to understand the phenomena they model [Sul22]. Instead, she posits that the primary obstacle to understanding is not a model's opacity but the level of "link uncertainty" – the extent to which the model is empirically supported and adequately connected to the target phenomenon.

If this link is missing, even simple models can fail to provide understanding [Gel16d]. On the other hand, a complex model that is sufficiently connected to a phenomenon of interest can help us comprehend that phenomenon without needing to know low-level details. Unless the explanatory question concerns the implementation itself, a high-level grasp of the model is enough [Sul22]. This distinction between understanding a model's workings and using a model to understand a phenomenon is akin to the difference between knowing how a car's engine operates and using it to navigate a city. Both are valuable, but they serve different purposes. If we accept Sullivan's argument, this leads us to the question of what kind of explanations NNs might be able to provide about phenomena of interest.

In section 4.2.1, I briefly discussed different notions of explanation that focused on an explanation's form or level of granularity. However, there is another, more high-level way of categorizing scientific explanations according to the purpose they serve – namely, the distinction between how-actually and how-possibly explanations.

How-actually explanations aim to describe why a particular real-world phenomenon occurs in the way that it does. They are concerned with identifying the actual causes or mechanisms behind an event. These explanations are often tied to the idea that there is a correct account of the phenomenon grounded in the causal history and laws of nature that govern the behavior of the system under investigation. In the context of scientific models, how-actually explanations typically involve models that are intended to be empirically accurate representations of real-world systems. These models strive to capture the relevant causal factors and interactions responsible for the phenomenon being explained. The success of a how-actually explanation is judged by how closely the model aligns with the actual behavior of the target system and by its ability to account for the observed phenomena [Gel19].

How-possibly explanations, on the other hand, are concerned with exploring the ways in which a phenomenon *could* occur. They do not purport to describe the actual causes or mechanisms at work. Instead, they seek to demonstrate that a particular event or pattern is possible given certain conditions. These explanations are particularly useful when there is uncertainty about the actual causes. How-possibly explanations are often employed in the early stages of scientific inquiry, where there may be a lack of detailed empirical data or a well-established theoretical framework. They serve as a means to probe the space of possibilities, to generate hypotheses, and to guide further investigation. In this sense, how-possibly explanations can be seen as a form of speculation, where the goal is to identify potential explanations that are consistent with what is known and to rule out those that are not [Gel19]. How-possibly explanations relate more broadly to the so-called "modal" dimension of modeling.

5.2.2 The modal dimension of modeling

Recall from section 2.1 that in a prevalent view of models, they should represent some real-life entity as accurately as possible. We have seen in the previous chapter that epistemic agents necessarily (and often felicitously) employ idealization in this process. However, some scientists seem entirely uninterested in representing existing real-life targets at all [Gel19]. Instead, they construct models of non-actualized or non-existent systems. E.g., Einstein famously used a *Gedankenexperiment* to begin building his theory of general relativity [Eli21].

These models are not failed attempts at representing the real world. They are deliberately hypothetical systems designed to probe what kinds of causal processes could underlie our observations [Knu11]. Thus, in addition to representing what we *know* to be the case, models can be tools for finding out what *might* be the case [RK22] – they can serve as embodied how-possibly explanations [Gel19]. This kind of inquiry is particularly important in areas like cognitive science, where many subjects are still poorly understood. It enables researchers to explore a range of scenarios and test which ones are compatible with our empirical reality [Gel19]. The capacity to "navigate the possibility space" [Gri+16, p. 122] in this way is referred to as the modal dimension of modeling. One problem with using models to navigate the possibility space is that this space is, in theory, infinitely vast. If we accept that a model need not represent a target system, how do we avoid the problem of link uncertainty? What stops the model from becoming a "freely floating subject of inquiry, unconstrained by any concern as to how it might be connected to the real-world facts" [Mäk09, p. 36]?

The answer is that the construction of models is, in practice, constrained by being designed for a particular purpose. This purpose is informed by existing theories or empirical findings, which means that much knowledge is built into a model from the outset – often implicitly [Knu11]. Every modeler brings with them their own set of values, assumptions, and expertise. This context determines the kinds of questions they ask and the choice of modeling ingredients they employ [Kit14; Sul22]. Therefore, even models of hypothetical systems rarely float about in a scientific vacuum, waiting to be linked to real-world entities. They are usually already linked to what we know of the world through the theoretical and empirical considerations that motivated and enabled their construction [Knu11].

5.2.3 Toy models

When researchers engage in modal modeling, they often employ what are commonly called "toy models". Toy models are characterized by two main features: simplification and idealization [Gel19]. Simplification refers to reducing model complexity by focusing on a small number of causal or explanatory factors that are responsible for the target phenomenon. Idealization, as discussed in section 4.2.2, involves making assumptions that may not hold in the real world but allow for clearer insights into the system's behavior or underlying principles. These idealizations can be minimalist (ignoring certain aspects), Galilean (altering certain aspects), or, often, a combination of both. Despite their simple and idealized nature, toy models are not trivial; they allow scientists to cut through the complexity of real-world systems to focus on the underlying principles that govern behavior [RHH18]. Nobel Prize-winning economist Paul Krugman, e.g., has been vocal on the need for toy models alongside econometric simulations [Kru96].

Simplification and idealization are, of course, involved in most types of modeling, and Reutlinger et al. are keen to point out that there is no sharp boundary between toy models and other models [RHH18]. Instead, models exist on a continuum, with toy models at the extreme ends of the simplicity and idealization dimensions. They are kept so simple and idealized because their primary aim is usually not to predict or explain a phenomenon. Instead, they are used to gain a qualitative or even a quantitative understanding of how a system might behave under certain simplified conditions. We can distinguish between two types of toy models: embedded and autonomous.

Embedded toy models are closely tied to an existing, empirically well-confirmed framework theory. They are derived from the principles of this theory and are used to explore specific questions within its domain. Embedded toy models can provide how-actually explanations by demonstrating how a phenomenon occurs according to the principles of the embedding theory. For example, an embedded toy model based on Newtonian mechanics to describe the orbit of a planet around the sun may simplify the solar system by limiting its focus to the gravitational interaction between two bodies. Despite this simplifying assumption, the model can provide a how-actually explanation for the elliptical orbits of planets by highlighting the core causal factors at play [RHH18].

Autonomous toy models, in contrast, are not directly derived from a well-confirmed theoretical framework. Instead, they are constructed based on more speculative or idealized assumptions to explore phenomena in areas where a comprehensive theory may not yet exist or is a matter of dispute. Toy models are particularly well-suited for providing how-possibly explanations [RHH18]. By creating a simple model world where they can freely manipulate rules and variables, researchers can test out "what-if" scenarios and map out the landscape of possibilities [Gel19]. Schelling's model of racial segregation in urban areas is a classic example of an autonomous toy model. It consists of a grid world populated by agents, each belonging to a group and slightly preferring to be surrounded by a certain percentage of their own race. Agents move to new locations until everyone's preferences are met. The unfolding dynamic often results in separated groups despite the initial mild preferences, providing a how-possibly explanation of segregation as due to individual choices [RHH18].

Much like the distinction between toy models and other models, the classification into embedded and autonomous toy models is more of a heuristic than a strict dichotomy. In practice, models may have varying degrees of connection to established theories [RHH18], and models may change status depending on the context in which they are used.

5.2.4 Relation to the guiding questions

Against the background outlined above, we can now address the questions from the beginning of this section and put the present case study into the context of the overall thesis. The modeling process of this work (see Figure 5.11) is similar to the last case study in that it is informed by the cognitive science literature. However, this inspiration takes different forms in the two studies. In the gSCAN case, I was loosely inspired by things we *know* to be the case: that object attention is limited, joint attention is helpful in learning, and the brain encodes locations egocentrically. In contrast, the model in this study is a computational implementation of a hypothesis of how letters get their shapes – i.e., something we *think* might be the case.

Writing systems have arisen over millennia under different constraints, such as the availability and characteristics of writing technologies, but also idiosyncratic choices of a few individuals that hardened into systems over time because of network effects: Having an accepted writing system within a culture is better than not having one. As we cannot run controlled experiments in which we let Sumerian-era humans repeatedly rediscover writing, there is value in running similar experiments *in silico*. Similarly, an astrophysicist cannot create black-hole collisions and must either look for them as they occur or create simulation models.

Thus, another major difference between this case study and the last one lies in the two studies' goals. The aim in the gSCAN case was to better understand the model so as to ultimately create higher-performing AI. The present study, on the other hand, has little interest in predictive performance or even inner model workings. Instead, the model serves to investigate a theory about a real-life phenomenon. This approach represents a departure from traditional uses of NNs, which is reflected in the fact that it is not centered on a pre-existing benchmark. Turning to what this case study can tell us more broadly about the epistemic affordances of NNs, let us once again consider the criticisms against them one by one, beginning with the concern that NNs are "unscientific".

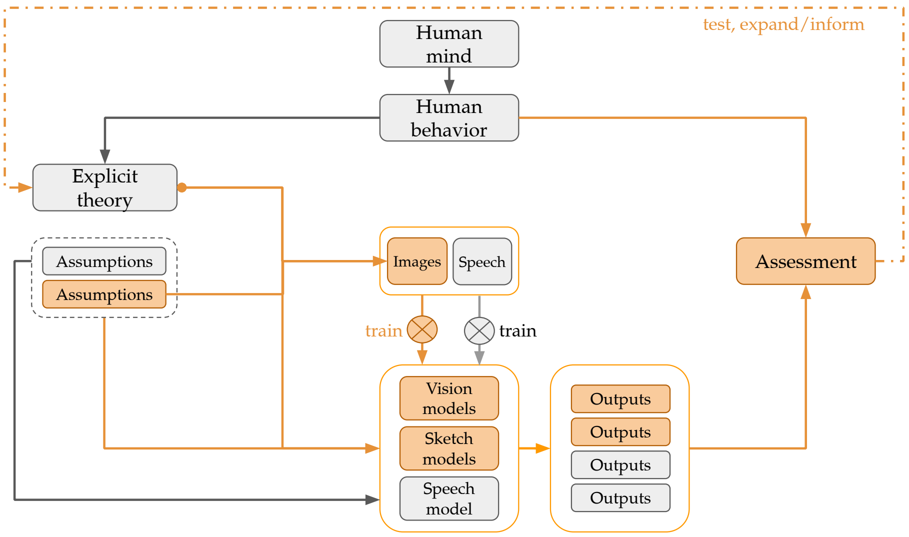


Figure 5.11: Overview of relations between human cognition, theory, assumptions, data, model, and outputs in the third case study. Relevant components performed or generated by us shown in orange. Components provided by third parties shown in gray. Circle at the beginning of an arrow indicates the starting point of investigation.

As discussed in section 5.2.2, models, even models that do not accurately depict a real-world target system, are often linked to theory through the questions they are designed to answer. In the case of this study, all aspects of the modeling setup, including the choice of pre-trained speech model, the architecture and optimization of sender and receiver models, and the composition of the image datasets, were driven by my research question of how statistical regularities in letter shapes may arise. The connection between model and theory is particularly explicit here in that the model is a computational implementation of an empirically supported hypothesis. Thus, the model can be considered an embedded toy model that is situated within a theoretical framework.

By definition, this toy model is highly simplified and idealized, making no pretense of being biologically plausible. Besides the differences between brains and the AI models used, the scenario in my setup is not necessarily realistic. For instance, the sender receives a direct loss signal derived from the internal representation of the receiver. This kind of instant feedback is more akin to synchronous face-to-face communication than the asynchronous, slow processes involved in the emergence of most actual writing systems [MKW20; Mor22a]. In future work, the setup could conceivably be made more complex and cognitively plausible. However, as discussed in the previous case study, simplifications need not be amenable to Galilean de-idealization to be of worth. Modeling minimalist, hypothetical scenarios can be helpful in its own right by allowing us to expand our understanding of the possible as well as the actual.

To draw the analogy to the "science fiction method" from this chapter's opening quote, I identify a real-world phenomenon's core components and mechanisms and create a simplified version of it by including only the most essential elements I am interested in. Then, I can use that simplified model to gain a qualitative understanding of the system's behavior under the new set of assumptions. For instance, I observed the effect of switching out rectangular convolutional filters with hexagonal ones.

In future work, the experimental setup could also be used to investigate the trade-off between referential and transmission efficiency more in-depth by varying the penalty factor α or code size N. For example, a code size of 100 is so large that there is room for symbols that encode composite sounds, as seen in Figure 5.9 – this would presumably not be the case for smaller codes. Furthermore, HuBERT could easily be replaced by speech models pre-trained on other languages or augmented with visual input of speakers' mouth areas [SMH22] to test how this influences the organizing principles of artificially evolved graphic codes.

Finally, we come to the criticism that NNs do not offer explanations. As mentioned, we should distinguish between two issues: "can NNs be explained?" and "can NNs help explain real-world phenomena?" Compared to the gSCAN case, the current study is less concerned with the first question. In a broad sense, the signaling game can be seen as relating to the interpretability technique of activation maximization for CNNs [Yos+15] and probing studies of self-supervised speech models [JPS22]. Similarly to how writing, as a cultural artefact, provides insights into human cognitive constraints, the evolved graphic codes can be considered a projection of high-dimensional model representations to a form more familiar to us – providing a window into these artificial systems of visual and speech perception.

However, the main focus of this case study is on the second question, namely, how can we use NNs to understand a target system? As discussed above, toy models can provide how-possibly and, in the case of some embedded toy models, how-actually understanding of real-world phenomena. In some fields, fully-formed, widely accepted theories and how-actually explanations may be difficult to attain. In cultural evolution research, for example, there are few opportunities to directly observe the development of socially shared behaviors [Gal05], much less manipulate them. In such cases, toy models can highlight potential mechanisms or patterns that could serve as the starting point for further empirical investigation.

The purpose of the model was not to have the English alphabet or the Indus script re-emerge as faithfully as possible as a function of societal and technological factors. Instead, I aimed to explore in some quantitative depth the effects of different constraints on the characteristics of writing systems that arise. Given that we know which writing systems did, in fact, arise and the statistical similarities they exhibit, we may intuit or abduce that the constraints placed on the corresponding model share similarities, in some regards, to those that shaped the development of human-created glyphs. The model in this study can thus be seen as an embodied how-possibly explanation, demonstrating how features of "cultural attractors" in writing systems [Kel+21] could emerge without pre-supposing universal, innate aesthetic preferences.

Crucially, when using NNs to answer these types of questions, looking inside the "black box" may not be necessary. The relevant elements of the system can be at the "what" or "why" level of the Marr hierarchy. In this case study, for example, explanations are mainly based on higher-level building blocks of the modeling setup, such as the input statistics, network structure, functional objective, and learning algorithm [KMK19]. Investigating individual weights or neurons would not have been conducive to the study's overall goal.

In summary, models, including NNs, are often connected to theory by virtue of the research questions that motivate their construction and the knowledge built into them. This link can also take a more explicit form in the case of embedded toy models. Furthermore, NNs can be of epistemic use without faithfully representing the brain – in fact, they need not represent any existing target system at all. Models have a modal dimension, meaning they can be used as tools for exploring hypothetical systems or scenarios. Finally, NNs can not only be the target of explanations but can serve as (how-possibly) explanations of real-world phenomena in their own right. For NNs to serve this function, it may not be necessary to understand the model itself in detail – high-level “what” and “why” explanations may be enough. The next case study demonstrates this point more starkly.

6 Modeling the emergence of intuitions about agents’ goals, preferences and actions with Video Transformers

The purpose of models is not to fit the data but to sharpen the questions.

Samuel Karlin

6.1 Study

The fourth case study relates to the core system of agents and their actions (see Figure 6.1). Although AI has made large strides in recent years, state-of-the-art models still largely lack these “commonsense psychology” abilities that emerge early on in infant development. The Baby Intuitions Benchmark (BIB) was explicitly designed to compare these aspects of social cognition in humans and machines. RNN-based models previously applied to this dataset were shown not to capture the desired knowledge. Here, we apply a different class of DL-based model, namely a VT, and show that it quantitatively more closely matches infant intuitions. We demonstrate through qualitative analyses that the model seems to learn to implicitly track relevant semantic categories, such as agents, goals, and subgoals. However, we also find that the VT is prone to exploiting particularities of the training data for its decisions.

The model proposed in this case study placed first in the Machine Visual Common Sense Challenge’s BIB track at the 2022 European Conference on Computer Vision (ECCV) [HD22c]. Versions of the following text were also presented at the Shared Visual Representations in Human & Machine Intelligence Workshop at the 2022 Conference on Neural Information Processing Systems (NeurIPS) [HD22b] and included in the proceedings of the 45th Annual Meeting of the Cognitive Science Society [HD23].

6.1.1 Introduction

The foundations of “commonsense psychology" emerge early on in a human’s development: Even pre-verbal infants have expectations about agents’ goals, preferences, and actions [Sto+23]. However, despite the tremendous progress DL has made in recent years, this core component of human cognition is still lacking in many state-of-the-art models [Lak+17]. When tested on BIB, a dataset designed to compare the social cognitive abilities of infants and machines, Behavioral Cloning (BC) and video prediction models based on RNNs failed to show infant-like reasoning [Gan+21]. We here evaluate a different class of DL model, namely a VT, on BIB.

Recent years have seen the rise of Transformers in various areas of AI, including tasks adjacent to social cognition, such as trajectory prediction for cars or pedestrians [Yua+21; Li+20; CWS21; Sui+21; Giu+21; Yu+20] and spatial goal navigation [DYZ21a; CPM21; Fuk+22]. As the

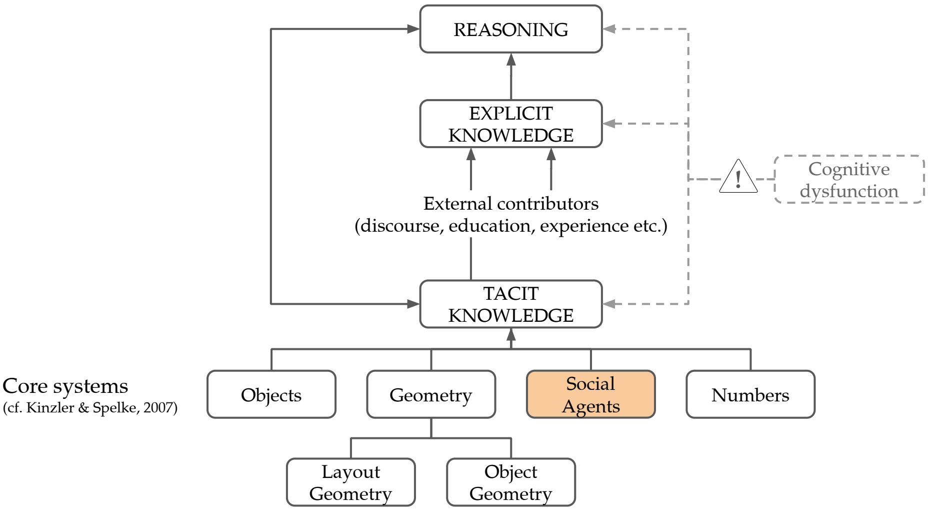


Figure 6.1: Situating the fourth case study in the broader study of cognition. Relevant parts of the framework marked in orange.

Transformer attention mechanism is based on computing pairwise interactions [LL22], it constitutes a promising approach for capturing the relations between, e.g., agents and goals in the BIB dataset. However, Transformer-based video prediction models require many costly pairwise computations. They are usually trained and evaluated on datasets like *Kinetics-400* [Kay+17] or UCF101 [SZS12], where video clip lengths range from 7 to 10 seconds – much shorter than those used in BIB, which may be up to 2 minutes long. We, therefore, implement some modifications to allow a VT to process BIB episodes and evaluate the resulting model.

We find that the VT quantitatively more closely matches infant intuitions about agents' goal preferences and efficient actions than previously tested DL baselines. However, qualitative error analyses show that the model fails to generalize systematically on some test tasks when agent or environment dynamics differ slightly from background training observations.

6.1.2 Baby Intuitions Benchmark

BIB is a dataset designed to test whether ML systems can discern the goals, preferences, and actions of others [Gan+21]. It consists of videos in the style of Heider and Simmel's animations [HS44], where agents, represented by simple shapes, carry out actions in a 2D grid world. BIB follows the Violation of Expectation (VoE) paradigm, i.e., each video has a familiarization and a test phase. The familiarization phase consists of eight successive trials during which an agent consistently displays a certain behavior, allowing the observer to form an expectation of future actions.

The test phase includes an expected outcome (perceptually similar to the previous trials but involves a violation of expectation) and an unexpected outcome (perceptually less similar but conceptually more plausible). BIB contains six types of test tasks, outlined in Table 6.1. It also contains background training episodes with four types of training tasks, which share the same structure as the test set. However, Gandhi et al. designed the BIB dataset such that only expected trials are provided in the background training episodes, and only isolated tasks are trained [Gan+21]. Therefore, the systematic combination of acquired knowledge is needed to generalize to the test tasks.

Table 6.1: Overview of BIB tasks.

	Familiarization trials	Test trial	Expected outcome	Unexpected outcome
Preference	Agent consistently chooses one of two goal objects and moves to it efficiently	Identical to a familiarization trial, but object positions are switched	Agent moves to preferred object at new location	Agent moves to non-preferred object at familiar location
Multi-agent		New agent appears	New agent moves to object not preferred by familiar agent	Familiar agent moves to previously not preferred object
Inaccessible goal		Preferred goal becomes inaccessible	Agent moves to other goal	Agent moves to other goal, even though both are accessible
Efficient agent	Agent moves efficiently around a barrier towards goal	Barrier is removed	Agent moves efficiently	Agent moves inefficiently
Inefficient agent	One agent moves efficiently, one moves inefficiently	Both agents move inefficiently	Previously inefficient agent moves inefficiently	Previously efficient agent moves inefficiently
Instrumental action	Agent removes a green barrier (inserts key into lock), then moves to goal	Green barrier gone or inconsequential	Agent moves directly to goal	Agent still moves to key

Goal-directed actions The *preference* task (1,000 episodes) tests whether an observer represents agents as having a preference for goal objects rather than locations. The setup consists of two goals and an agent whose starting position is fixed. In the familiarization trials, the agent consistently moves toward the same object. Goal locations and identities are correlated, such that preferred and non-preferred goals have a similar position across trials. In the test phase, the two objects appear in positions previously seen during familiarization. However, goal identities are switched. In the expected outcome, the agent moves to the preferred object. In the unexpected trial, the agent follows the same trajectory as seen during familiarization and moves to the non-preferred object (see Figure 6.2).

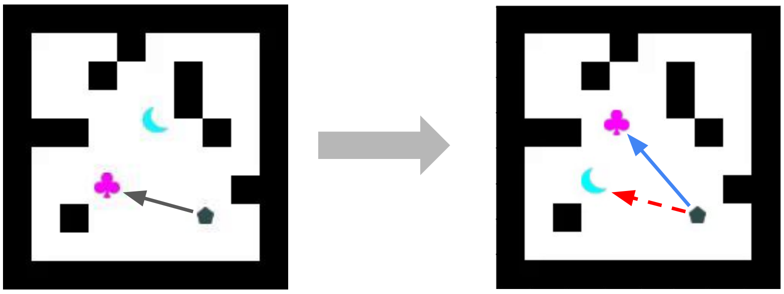


Figure 6.2: Example of a *preference* task. Agents move repeatedly to the same goal during familiarization (left). Goal locations are switched for testing (right). In the expected outcome (blue solid line), the agent still chooses the same object. In the unexpected outcome (red dashed line), the agent instead follows the familiar path to its non-preferred goal.

The *multi-agent* task (1,000 episodes) tests whether an observer attributes specific goal preferences to specific agents. The setup consists of two goal objects appearing at different positions across trials and an agent with a fixed starting position. Again, the agent moves repeatedly to the same object during familiarization. In the unexpected test outcome, the agent moves towards its non-preferred goal. In the expected outcome, a new agent replaces the previously seen one and moves toward the familiar agent's non-preferred object. The unfamiliar agent choosing a new goal should be less surprising than a familiar agent switching preference (see Figure 6.3).

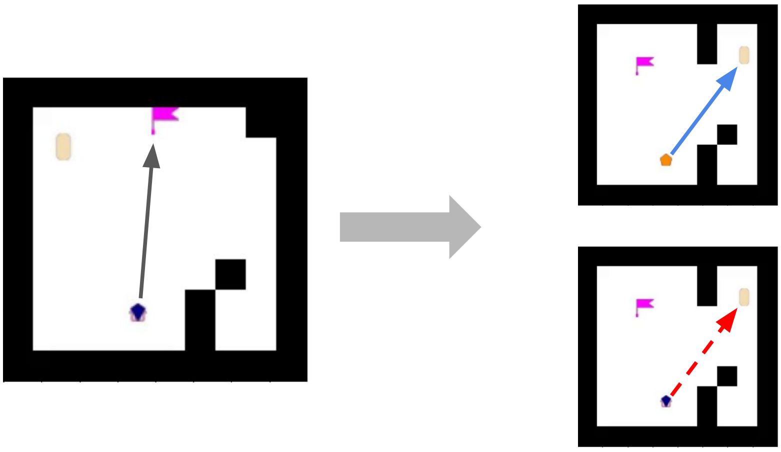


Figure 6.3: Example of a *multi-agent* task. Agents move repeatedly to the same goal during familiarization (left). A new agent appears in the test trial (right). This new agent choosing the other agent's non-preferred object (top right) should be less surprising than the familiar agent doing so (bottom right).

The *inaccessible-goal* task (1,000 episodes) tests whether an observer understands the principle of solidity and that physical obstacles may restrict agents' actions. The familiarization trials are identical to the *multi-agent* task. In the expected test trial, a black barrier makes the previously preferred object inaccessible, and the agent moves to the other goal. In the expected test trial, the agent switches goal preference despite both objects staying accessible (see Figure 6.4).

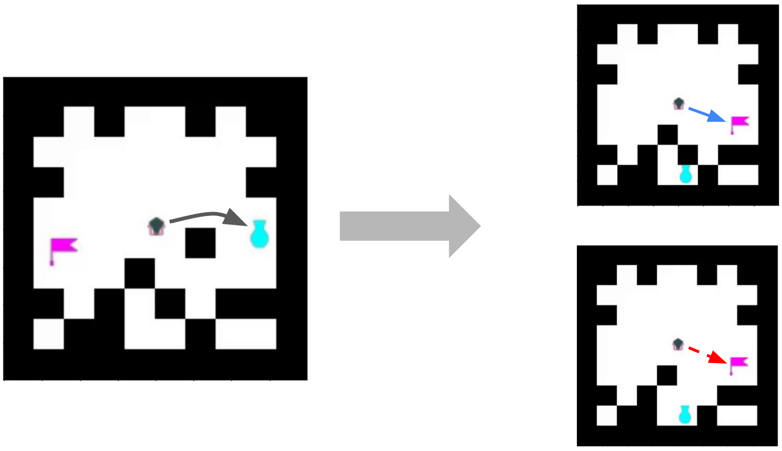


Figure 6.4: Example of an *inaccessible-goal* task. Agents move repeatedly to the same goal during familiarization (left). The agent switches goals in the test trial (right). This should be expected if the preferred object is inaccessible (top right) but unexpected if both objects are accessible (bottom right).

Efficient actions The *efficient-agent* task tests whether an observer expects agents to move efficiently towards their goal. It consists of two subtasks: path control (1,500 episodes) and time control (1,000 episodes). In both subtasks, the setup consists of one goal object and one agent. During familiarization, the agent moves efficiently towards the object but must navigate around a barrier to reach it. This obstacle is removed in the test phase. In both subtasks, the expected outcome consists of the agent moving efficiently towards its now-unobstructed goal. For the path control task, a previously seen combination of agent and goal location is used, and the unexpected outcome consists of the agent moving along the familiar, but now inefficient, trajectory (see Figure 6.5). For the time control subtask, the goal object is placed closer to the agent, and the unexpected outcome consists of the agent following a path that is inefficient but takes the same amount of time as the efficient one.

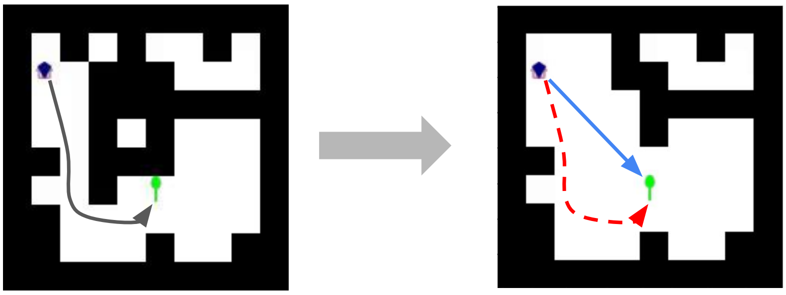


Figure 6.5: Example of an *efficient-agent* task. During familiarization (left), the agent navigates efficiently around an obstacle to reach its goal. The barrier is removed during testing (right). The agent is now expected to move efficiently (blue solid line) rather than following the same path as before (red dashed line).

The *inefficient-agent* task (890 episodes) tests whether an observer forms expectations about the actions of irrational agents. During familiarization, an agent is shown either moving efficiently or inefficiently. In the test phase, the agent is shown moving inefficiently to the goal. This should be an unexpected outcome if the agent previously behaved rationally and unsurprising if the agent previously behaved irrationally (see Figure 6.6).

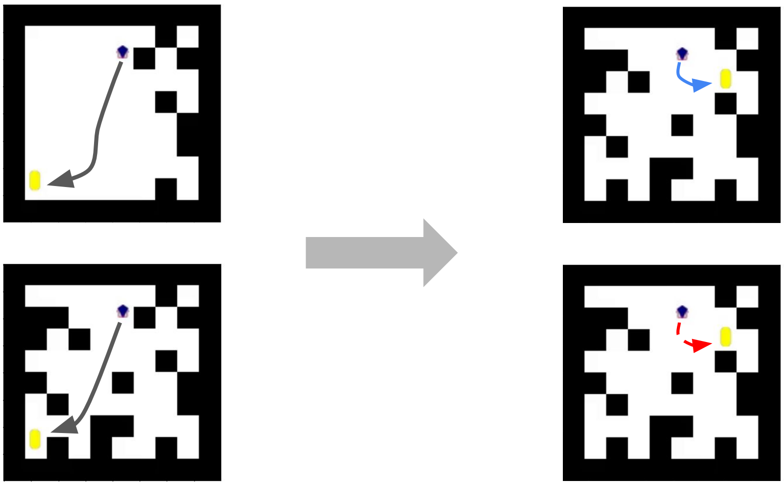


Figure 6.6: Example of an *inefficient-agent* task. Familiarization trials shown on the left, test trials on the right. An agent that moves inefficiently during familiarization (top) is expected to continue doing so during testing, whereas an efficient agent (bottom) beginning to move inefficiently should be surprising.

Instrumental actions The *instrumental-action* task (987 episodes) tests whether an observer can recognize an agent's action sequences as instrumental and directed towards higher-order goals. The setup consists of a goal, an agent, a removable green barrier with a lock, and a key, represented by a red triangle. During familiarization, the goal is obstructed by the green barrier. The agent collects the key, inserts it into the lock, removes the barrier, and moves to the goal. In the test phase, a key is still present, but the green barrier is either absent or no longer blocking the goal. In the expected outcome, the agent moves directly towards the goal, whereas it still moves towards the now-obsolete key in the unexpected outcome (see Figure 6.7).

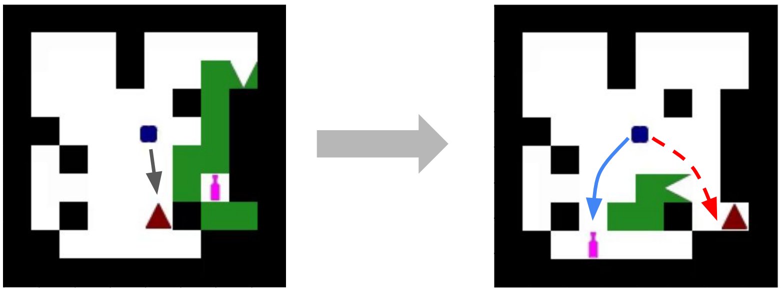


Figure 6.7: Example of an *instrumental-action* task.

Background training episodes To facilitate the training of ML models, BIB includes many background episodes that share the same structure, agents, and goal objects as the test set. However, only expected trials are provided during training. The training set is split into four tasks. In order to generalize systematically on the test trials, the model needs to combine knowledge acquired from all four training tasks.

In the *single-object* task (10,000 episodes), an agent navigates efficiently to a goal object (see Figure 6.8a). In the *preference* task (10,000 episodes), the agent consistently chooses one object over another across trials (see Figure 6.8b). In contrast to the *preference* test task, both objects are located very near the agent, so navigation is not trained. In the *multi-agent* task (4,000 episodes), the agent moves to a very close-by single goal object (see Figure 6.8c). At some point in the episode, the agent is replaced with a new agent. This differs from the *multi-agent* test task, which has two goals placed farther away, and the new agent only appears in the test trial. In the *instrumental-action* task (4,000 episodes), the agent is initially confined by a green barrier, which it removes with a key to access its goal. It differs from the *instrumental-action* test task in that the barrier surrounds the agent rather than the goal.

Because BIB adopts its tasks and paradigm from developmental cognitive science and provides sufficient data to train DL-based models, it allows for the direct comparison of human and machine performance [Gan+21]. A critical first step in this direction was taken by Stojnić et al., who collected infants' responses on a representative selection of BIB episodes and compared them with three state-of-the-art DL models from two classes: BC and video modeling [Sto+23]. Recently, Zhi-Xuan et al. proposed a principled alternative to DL approaches based on a Hierarchically Bayesian Theory of Mind (HBToM) [Zhi+22]. The results of both works serve as a comparison in this paper. Note, however, that HBToM requires access to symbolic states and is specifically engineered to solve BIB-like social cognition tasks. In contrast, the data-driven baselines and VT model have weaker inductive biases in this regard.

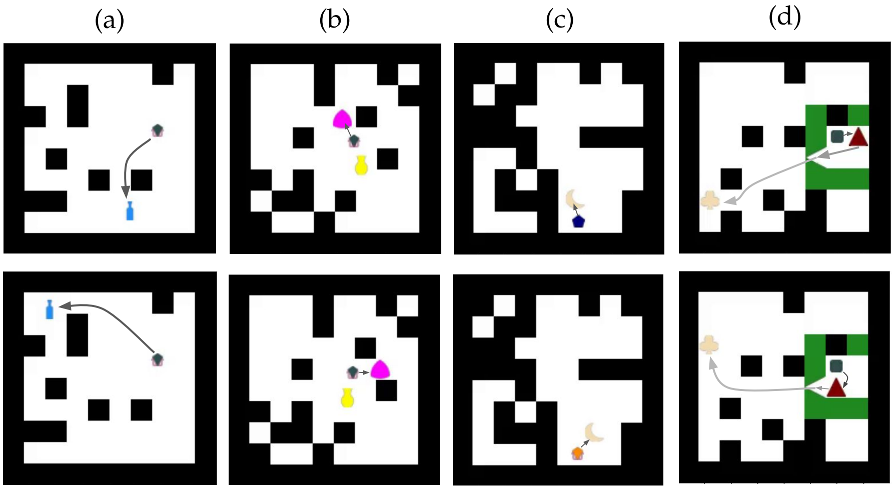


Figure 6.8: Examples of training trials, consisting of *single-object* (6.8a), *preference* (6.8b), *multi-agent* (6.8c), and *instrumental* (6.8d) tasks.

6.1.3 Methods

Our model consists of a CNN encoder, a Transformer component, a CNN decoder, and a linear output layer. A schematic visualization is shown in Figure 6.9. The CNN encoder (Figure 6.9 Ⓐ) has two convolutional layers and two max-pooling layers. For each 3 × 84 × 84 input image, it produces a 30 × 21 × 21 representation, which we concatenate with x- and y-position encodings, yielding 32 × 21 × 21 patches. As attending over every pixel would be computationally prohibitive, the CNN encoder was designed to reduce the frame's resolution by extracting higher-level features while retaining sufficient spatial detail.

After encoding all the frames of an episode in this way, we extract the top-n patches per frame that display the most significant change compared to the previous frame (Figure 6.9 Ⓑ). We do this for each frame of the familiarization trials. The reason we only use n patches is that attending over every patch, frame, and trial would be extremely computationally expensive. N was set to 3, as using a higher number would have exceeded the memory resources in our training setup, even with our very small batch size. However, it is unlikely that a choice of $n > 3$ would have led to substantially better performance, as BIB trials are mostly static. The only movements stem from the agent and, in *instrumental-action* tasks, the green barrier. Therefore, it is rare that more than three patches exhibit a change from one frame to the next. The extracted patches are fed into the first of three blocks of the Transformer component.

Each block has five layers with eight heads of input dimension 32 and hidden dimension 256. The first block (Figure 6.9 Ⓒ) performs cross-attention over the test trial's encoded first frame and previous familiarization trials, effectively "priming" the model by calculating the influence of previous observations on the current input. The results of attending over each trial are averaged and passed through a self-attention block (Figure 6.9 Ⓓ). We then extract n patches for each frame in the test trial (Figure 6.9 Ⓔ) in the same way as we did for the familiarization trial frames. The patches serve as input to the third attention block (Figure 6.9 Ⓕ), which attends over past steps in the test trial. In the final step, the outputs of the Transformer component are passed through an output layer (Figure 6.9 Ⓖ), which produces a 1 × 21 × 21 prediction of the agent's next position and a CNN decoder (Figure 6.9 Ⓗ), which produces a 3 × 84 × 84 prediction of the next frame.

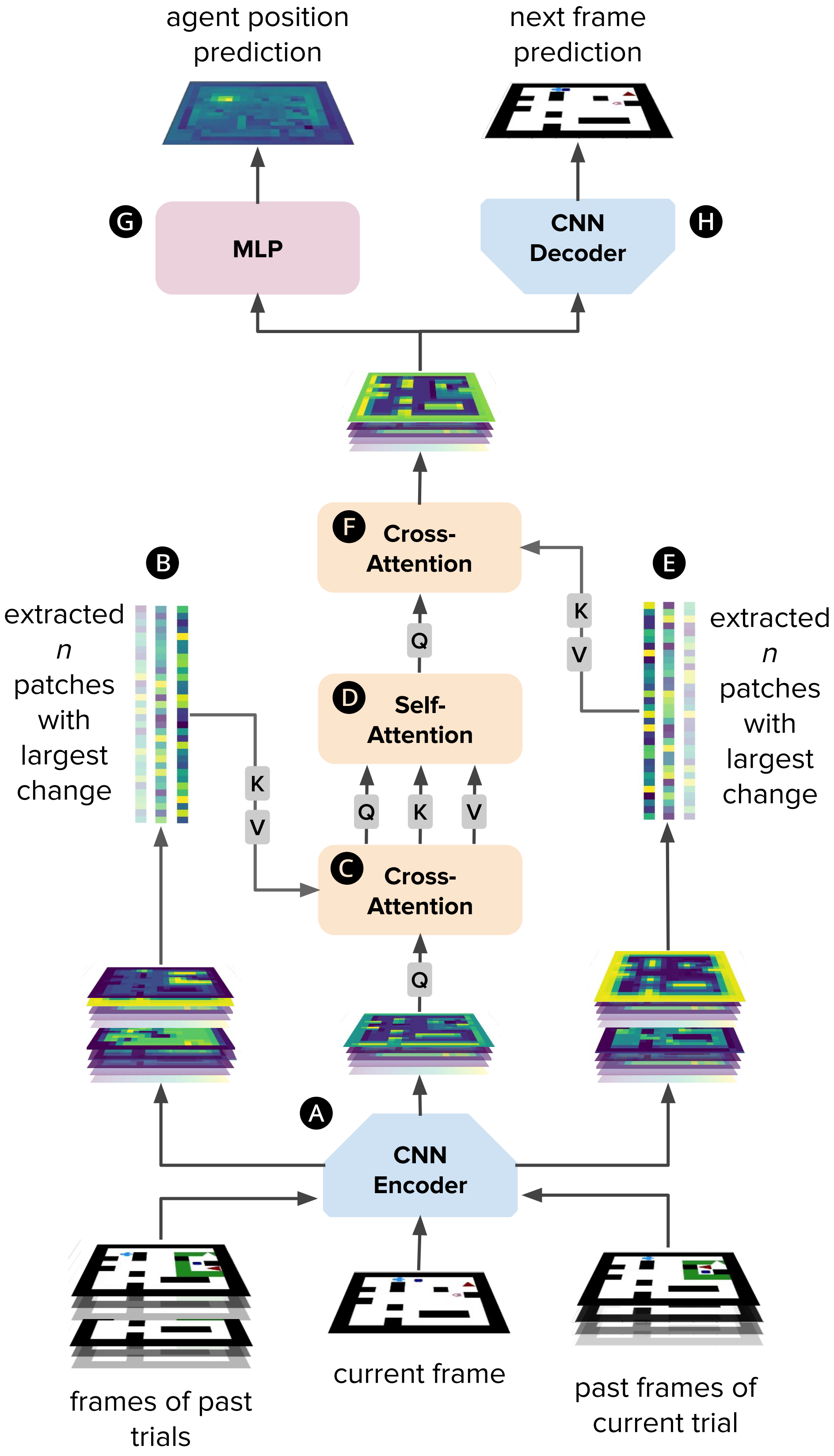


Figure 6.9: Schematic visualization of the Video Transformer architecture.

Given the model's two prediction targets, our loss function consisted of the sum of two terms. The first term was the Binary Cross-Entropy (BCE) loss between the prediction of the agent's next step and the actual agent position. To address the imbalance between the “agent" and “no-agent" class, we employed a weighted version of the BCE loss, which is widely used in instance segmentation [Jad20]. The second term was the Mean Squared Error (MSE) between the prediction of the next frame and the actual next frame, upweighted by a constant factor so that both loss terms were scaled evenly. This second term was introduced because Transformers may disregard agent identities unless incentivized otherwise [Yua+21]. For tasks like *preference*, which relies on the preservation of agent shapes and colors, we thus found that it improved performance to include an auxiliary reconstruction loss. During evaluation, only the main BCE loss was used.

As in Gandhi et al., the videos' frame rate was downsampled by a factor of 5 [Gan+21]. We used a maximum sequence length of 90. Frame rates of longer sequences were interpolated to fit the maximum length. Of the BIB background episodes, we used 80% for training, 15% for testing, and 5% for validation. Models were trained using the Adamax optimizer for six epochs, after which we saw no further improvement on background training tasks. The batch size was set to 6 because of the VT's high memory requirements. We tested the models on the validation set in five evenly spaced intervals per epoch and saved the model with the lowest validation loss to avoid overfitting. The total number of trainable parameters in the VT is $772,162$. For comparison, the two publicly available baseline BC methods by Gandhi et al. contain $925,666$ and $986,306$ trainable parameters, respectively [Gan+21]. On a 16-core AMD EPYC 7282 server with six GeForce RTX 2080 GPUs, training time was around 3 hours per epoch.

Our model shares some commonalities with the BIB baseline DL models but also differs in several aspects. Both the VT and baseline models use CNNs to encode frames and average embeddings across familiarisation trials to obtain context vectors. However, we use attention mechanisms to obtain these embeddings, whereas the baselines used RNNs or MLPs. In contrast with the BC baselines, we also do not pre-train our CNN encoder separately, and we do not add the agent's actions as inputs – only the video frames. Finally, we predict both the next frame and the agent's position, while Gandhi et al.'s video modeling approach predicted only the next frame, and their BC approach predicted only the agent's next action [Gan+21].

6.1.4 Results

We trained five models on the BIB background training tasks, each with a different random weight initialization. We report the models' average performance on the test set of the background training tasks in Table 6.2 and the performance on BIB evaluation tasks in Table 6.3. The baseline DL models previously tested on BIB used the prediction error of the frame with the highest loss as their metric of "surprise," as this provided better results compared to the mean error over entire trials [Gan+21]. In our case, the mean error yielded a higher performance on most tasks, so we report both metrics here. However, binary VoE accuracies include no information about the magnitude of the difference in surprisal scores between expected and unexpected trials. We, therefore, also show z-scored means of both the models' average prediction error and infants' looking times in Figure 6.10, as reported by Stojnić et al. [Sto+23].

Table 6.2: Mean squared error of the frame prediction and weighted binary cross-entropy loss of the agent prediction on the test split of the BIB background training tasks, averaged over the five trained models.

Training task	**MSE**	**BCE**
Single object	$7.05 \times 10^{-}4$	$1.58 \times 10^{-}2$
Preference	$7.07 \times 10^{-}4$	$1.38 \times 10^{-}2$
Multi-agent	$5.94 \times 10^{-}4$	$1.32 \times 10^{-}2$
Instrumental actions	$1.42 \times 10^{-}3$	$1.33 \times 10^{-}2$

Table 6.3: VoE Accuracy on BIB evaluation tasks. VoE Accuracy denotes whether model error is higher on expected trials than on unexpected trials. VT (Mean) uses the average error over all test trial frames as the "surprise" metric, whereas VT (Max) uses the error for the frame with the highest loss. For the VT models, we report the average accuracy and standard deviation over five models trained on the same data but with different random initializations. Baselines and VT are data-driven computer vision models, whereas HBToM uses a principled Bayesian solution that requires access to symbolic states. Chance level accuracy is 50%.

		Baselines			VT (ours)	
Task	**HBToM**	**BC-MLP**	**BC-RNN**	**Video-RNN**	**VT (Mean)**	**VT (Max)**
Goal-directed						
Preference	99.7	26.3	48.3	47.6	82.1 ± 0.0	80.8 ± 0.0
Multi-agent	99.2	48.7	48.2	50.3	49.1 ± 0.0	49.2 ± 0.0
Inaccessible goal	99.7	76.9	81.6	74.0	89.8 ± 0.0	85.5 ± 0.0
Efficiency						
Efficient agent	95.8	96.0	95.3	99.5	98.3 ± 0.0	98.4 ± 0.0
Inefficient agent	96.6	73.8	56.5	50.1	29.5 ± 0.1	34.1 ± 0.1
Instrumental actions						
Instrumental action	98.5	67.0	77.9	79.9	92.6 ± 0.0	84.7 ± 0.0

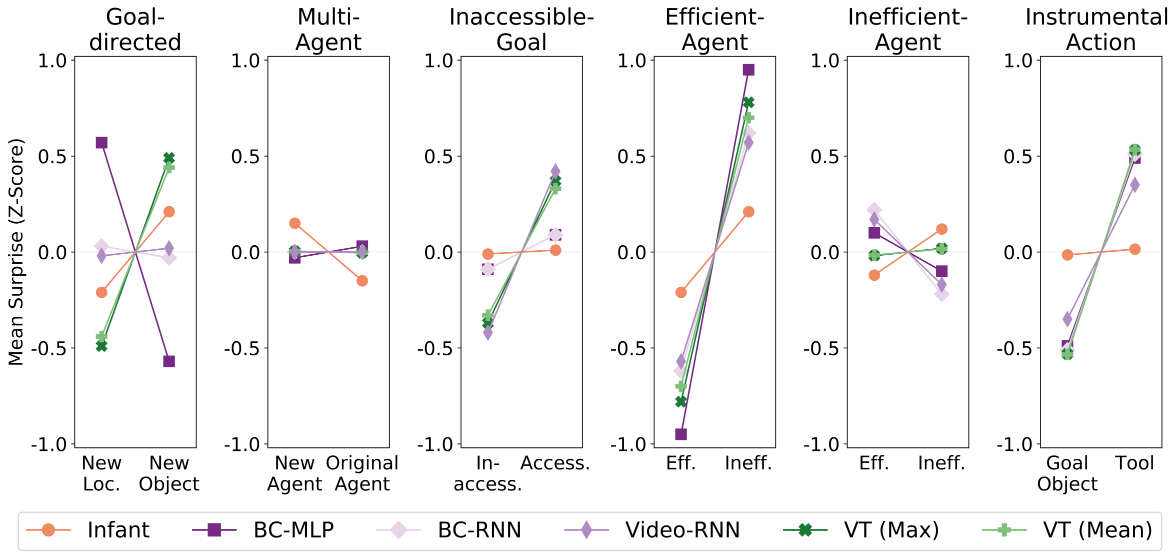


Figure 6.10: Z-scored means of the models' average surprisal scores and infants' looking times to the expected and unexpected outcomes in the BIB test episodes.

Preference In contrast to the DL-based baselines, the VT seems, at least to some degree, to associate agents with certain goal preferences in the *preference* task (see Figure 6.10). To investigate which parts of the familiarization trials the model relied on most, we performed a form of occlusion analysis. We used only one trial as the familiarization input (performance was almost identical when using one vs. the full eight trials) and dropped each patch fed into the first Transformer block in turn. We recorded the z-scored difference in prediction error between the expected and unexpected outcomes for each patch. An example result is shown in Figure 6.11. Models tended to rely on either the agent's last or first step. Averaged over all models and episodes, the patch with the largest impact on the final prediction was part of the last two frames of the familiarization trial in 52.6% of cases.

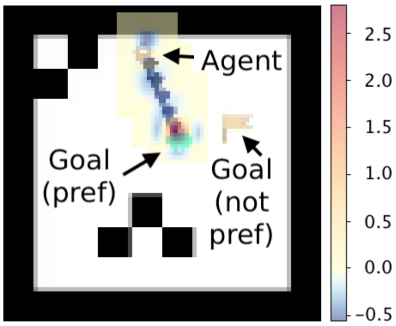


Figure 6.11: Z-scored impact of omitting a patch from the *preference* familiarization trial.

Multi-Agent Similarly to the other DL models, the VT does not acquire the desired knowledge from the *multi-agent* background training tasks, which feature both agents moving towards the same single goal across trials. Note that the infants tested on BIB were, in fact, more surprised at the supposedly "expected" trials (see Figure 6.10). Stojnić et al. hypothesize that this may be because of the increased novelty of the new agent. A closer look at the frame predictions produced by the VTs hints at some confusion regarding the agents' identity: In some cases, the model reconstructs the familiar agent in the unexpected trial rather than the new agent present in the input (see Figure 6.12 for an example). Averaged over all models and episodes, this was the case 27.9% of the time.

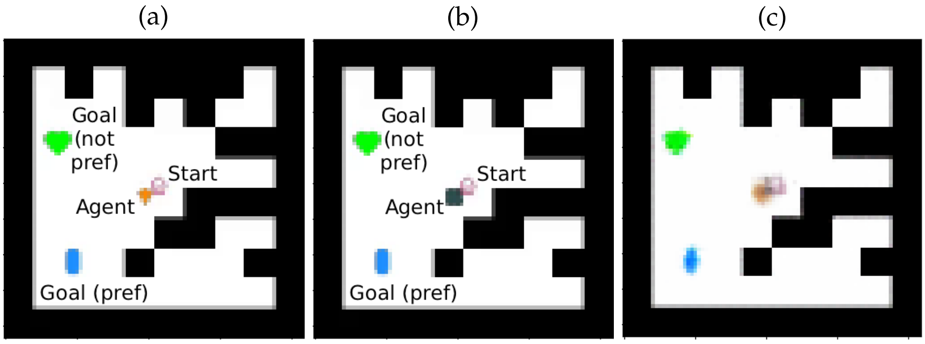


Figure 6.12: Unexpected *multi-agent* outcome involving familiar agent (a). Expected outcome involving new agent (b). Model prediction for expected outcome (c).

Inaccessible In the *inaccessible-goal* task, the VT model achieves a higher accuracy than previous DL models. It exhibits a stronger deviation in surprise than the infants, who were indifferent on this task (see Figure 6.10). Stojnić et al. posit that infants may have considered the new barrier in the expected outcome as indicative of a new environment and not carried over any goal preference expectations from the familiarization trials. Although the VT has a lower prediction loss on the expected outcome in most cases, it is more "split" than in the single-object case (see Figure 6.13 for an example prediction). Averaged over all models and episodes, the entropy of the models' predictions on the test trial's last frame was 1.10 for the expected and 1.47 for the unexpected outcome. For comparison, the average entropy for the last frame of the *single-object* background training task was only 0.58.

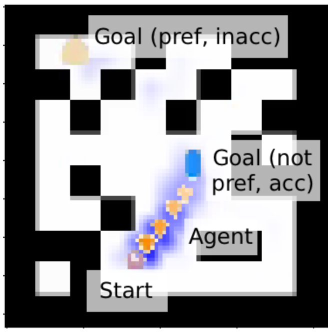


Figure 6.13: *Inaccessible goal* task. Predicted agent positions marked blue.

Efficiency As in previous models, the VT's VoE accuracy on the *efficient agent* tasks is nearly perfect – it strongly expects agents to move towards their goal efficiently. This is in accordance with infants' intuitions (see Figure 6.10). On the *inefficient-agent* task, the VT tends to be more "surprised" at the previously inefficient model moving inefficiently than at the previously efficient agent doing so. Although not necessarily a desired outcome, this is actually more in line with the intuitions of infants tested on BIB, who attributed rational action both to previously efficient and inefficient agents in a new environment (see Figure 6.10).

When we compare the impact of the familiarization trials featuring the efficient vs. inefficient agent on the VT model (see Figure 6.14), we see that a similar mechanism is at work: The lowest levels, which attend over past familiarization trials, show differences in activation. However, these differences all but disappear throughout the higher layers. This leads to the inefficient agent being treated the same way as the efficient one, which explains the mean surprise score being almost the same in both cases. The slightly larger error for the inefficient agent most likely stems from inefficient agents not being seen during training, leading to higher prediction uncertainty.

Instrumental Actions Compared with the other DL models, the VT performs similarly on episodes with no barrier and better on episodes with inconsequential or blocking barriers. Again, infants were indifferent on this task (see Figure 6.10). Stojnić et al. note that they may have failed to recognize the instrumental actions because they were causally opaque. Although the VT is correct in most cases in terms of VoE accuracy, it, too, seems to not have entirely understood the causal mechanism.

The frame predictions show that the model usually expects the disappearance of the key on the first step, even though the agent has not collected and inserted it. Averaged over all models and episodes, the VT at least partly predicts the key's position as the agent's first step in 47% of cases, even though the key is mostly far away from the agent. This is most likely because

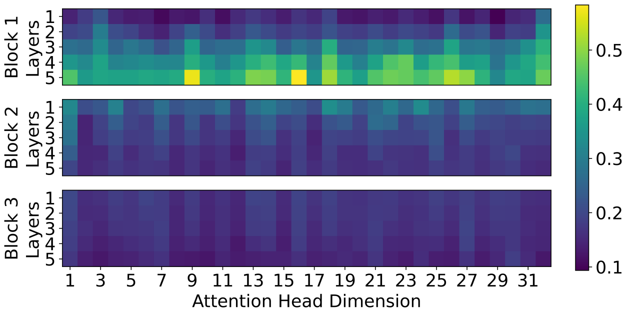
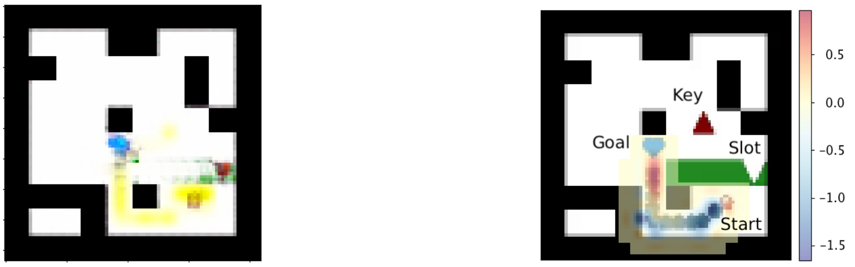


Figure 6.14: Avg. difference in the model layers' activations when processing the episodes' unexpected vs. expected familiarization trials, featuring an efficient or an inefficient agent, respectively.

the key is always right next to the agent in the background *instrumental-action* tasks, and thus constitutes its first step. The VT also often predicts the disappearance of the green barrier towards the end of the episode, even though the key was not inserted. This is most likely because the green barrier has always disappeared by the time the agent reaches the goal in the background tasks. Occlusion analyses support this hypothesis: The parts of the test trial that most contribute to the z-scored MSE prediction error on expected *instrumental-action* outcomes were usually the agent's first and last steps (see Figure 6.15 for an example).



(a) Predicted last frame and agent trajectory (yellow).

(b) Z-scored impact of each test trial patch on final MSE error.

Figure 6.15: Prediction on an *instrumental-action* task.

Decoding experiment Inspired by probing analyses of pre-trained language models [Cla+19], we trained linear regression models to predict the current position of the agent, goal, and sub-goals (keys and locks) based on the activations in each layer of each VT block. Each linear model had an input dimension of 256 (8 attention heads per layer, each with dimension 32) and an output dimension of 4 (one for each prediction target). The models were trained with the Adam optimizer [KB15b] set to default parameters, using the same epoch number and batch size as the main experiments described in the Methods section. We used the background training set for optimization and display the results for the background validation set in Figure 6.16.

In general, errors decrease in the deeper layers of the attention blocks, indicating more focused attention heads. The heads in the first block, which attend over familiarization trials, do not display much specialization regarding the analyzed categories. However, at least in the higher layers, the agent, key, and lock categories have a comparatively lower decoding error than the goal category. Note that the agent's position often corresponds with the key and lock position for long stretches of instrumental action trials, as the agent waits for the green barrier to

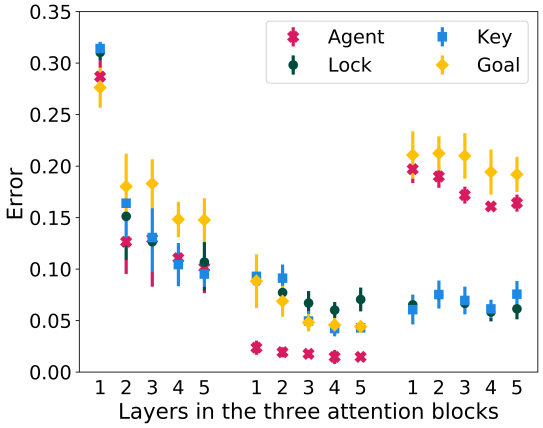


Figure 6.16: Weighted BCE loss of linear probes trained on decoding the current position of goals, agents, and sub-goals from attention head activations in each layer. Error bars indicate standard deviation across the five trained models.

disappear after having inserted the key into the lock. The second block, which self-attends over the test trial's first frame, has the lowest decoding error across categories and a particularly low error for the agent's current position. The third block shows a clear separation between categories, with locks and keys displaying a much lower decoding error than goals and agents. This is presumably because the third attention block autoregressively predicts the agent's next step, which, as mentioned, often coincides with the key and lock position while the agent is waiting in place for the barrier to disappear.

To sum up, the VT seems to have learned to implicitly keep track of relevant semantic categories, such as agents, goals, and subgoals, which are usually modeled as explicit variables in Bayesian approaches.

6.2 Discussion

In conclusion, the proposed VT model outperforms previous DL-based baselines on the *preference*, *inaccessible-goal*, and *instrumental-actions* BIB tasks in terms of VoE accuracy. Its surprisal scores are also more in line with infants' expectations than previous DL models, in that it tends to represent agents' actions as directed towards goals rather than locations, and it defaults to expecting rational actions. This suggests that the Transformer's attention mechanism can be helpful in acquiring intuitions about agents' goals, preferences, and actions purely from predicting the next step in videos.

However, a qualitative analysis of the VT's errors also demonstrates the pitfalls of this approach: Models may exploit the particularities of a training dataset in an unintended way [Gar+20; Gei+20], e.g., by associating the disappearance of the green barrier in the *instrumental-actions* task with the agent's first and last step rather than with the key mechanism. This may be mitigated with a more realistic data setting, where models can gain experience with diverse agents and disambiguate causes and effects of instrumental mechanisms interactively in a manner closer to human infants. The findings also support the benefit of investigating hybrid architectures that incorporate methods that explicitly model human intuitions, such as HBToM, to take advantage of both the flexibility of DL-based approaches and the data efficiency and robustness of principled Bayesian models.

On a broader level, this case study demonstrates the benefits of employing a plurality and diversity of models in the study of cognition. It also highlights how the epistemic aims and success criteria of cognitive science and ML may diverge and sometimes even conflict with one another. I will elaborate on both of these points in more detail in the following sections. As we encountered a Bayesian model in this chapter for the first time, I will begin by placing the Bayesian paradigm within the overall context of modeling approaches in cognitive science.

6.2.1 Approaches to modeling in cognitive science

Cognitive science has undergone significant transformations since its inception, particularly in how researchers model cognitive processes. This evolution of modeling paradigms has been deeply intertwined with the technological advancements of the times. The "tools-to-theory" heuristic suggests that the tools available to researchers not only facilitate the exploration of cognitive phenomena but also shape the theoretical frameworks that emerge. As computational tools evolve, they inspire new analogies and metaphors for mental processes, leading to paradigm shifts in modeling [Gig20].

The early days of cognitive science were marked by the development of symbolic models, influenced by the first digital computers. These computers inspired scientists to consider cognitive processes as discrete, serial computations akin to the operations of a computer program [Gig20]. In the 1950s and 1960s, researchers like Allen Newell and Herbert Simon developed the first computer programs that could perform logical reasoning and problem-solving tasks, such as the Logic Theorist and General Problem Solver. These models conceptualized thought as the manipulation of discrete symbols according to formal instructions. This approach emphasized the structured, rule-based nature of cognition, suggesting that the mind operates by manipulating internal representations of the external world [McC09].

Symbolic models have the benefit of being highly interpretable because they operate through explicit rules and symbols that can be directly inspected. They naturally support modularity, allowing for the construction of complex systems from simpler components, and they allow for precise definitions and formal proofs of model properties. However, symbolic models can struggle with noisy data and tasks that require generalization beyond the exact conditions they were programmed for. They can be brittle, i.e., small changes in input or rules can lead to large, unexpected changes in output. Furthermore, verifying and maintaining symbolic rule bases is difficult, and incorporating new knowledge typically requires significant manual rule crafting [IK20]. Despite early successes, these limitations of symbolic AI soon became apparent, leading to a period of reduced funding and interest known as the first "AI winter [Too+21]."

In the 1980s, connectionist or neural network models emerged as a significant challenge to the symbolic paradigm. This development was driven by the advent of more powerful computers and the introduction of the backpropagation algorithm, which enabled the training of multi-layer networks [PP20]. Connectionism represents a departure from the discrete and rule-based processing of symbolic models. It posits that cognitive processes emerge from the interactions of many simple, interconnected processing units. These units, loosely inspired by the neurons in the brain, work in parallel to process information. Learning in connectionist models occurs through the adjustment of the strengths of the connections between units based on experience. Rather than discrete symbols, knowledge takes the form of distributed representations [Has+17]. This approach has been particularly effective in modeling perceptual processes and certain aspects of memory.

In contrast to symbolic models, neural networks excel at learning patterns and generalizations directly from data, making them more dynamic and adaptable than symbolic models. They are also more robust to noise and better able to handle ambiguous or incomplete information effectively. However, connectionist models are often criticized for being "black boxes," as it can be challenging to understand how they arrive at a particular decision. Furthermore, their performance heavily depends on the quantity and quality of their training data. Although their distributed nature lends itself to leveraging efficient parallel processing, neural network optimization usually also requires significant computational resources [MP20].

Bayesian models represent a probabilistic approach to understanding cognition. These models are based on a method of statistical inference in which Bayes' theorem is used to update the probability of a hypothesis as more evidence becomes available. They have their origins in the 18th-century work of Thomas Bayes and Pierre-Simon Laplace. However, it was not until the late 20th century that these methods began to be widely applied in cognitive science. The rise of Bayesian models in cognitive science can be attributed to advances in computational methods, such as Markov chain Monte Carlo, which made it feasible to perform Bayesian inference on complex models. Bayesian models view the mind as a rational, probabilistic inference engine that integrates prior knowledge with new evidence to make predictions, decisions, and interpretations about the world [McC09].

Because Bayesian models naturally incorporate uncertainty, they are well-suited for modeling decision-making. They can seamlessly integrate prior knowledge with observed data, reflecting an essential aspect of human cognition: the capacity to deal with the inherently uncertain nature of sensory information by leveraging prior knowledge and probabilistic reasoning [KD18]. In contrast to neural networks, Bayesian models can generalize from limited data by leveraging prior distributions and are well-suited for continuous learning scenarios where data arrives incrementally [DS23]. However, performing Bayesian inference can be computationally intensive, especially in complex models. Furthermore, specifying prior distributions and likelihood functions can be challenging and introduces subjectivity. Scaling Bayesian models to large datasets or highly complex problems can also be difficult [MP20].

Symbolic, connectionist, and Bayesian models each offer valuable insights into the nature of cognitive processes, albeit from different perspectives. Symbolic and Bayesian methods are often used to model so-called "as-if" theories. These theories are characterized by their highly idealized nature and describe cognitive processes *as if* they were following a set of formal rules or statistical principles. As-if models are not necessarily concerned with actual psychological processes but rather with the outcomes that would result if the mind operated according to certain rational principles [Gig20]. These models are helpful in exploring how ideal solutions to problems could be computed but often do not account for the complex, frequently suboptimal processes used by humans. For example, a large body of behavioral science suggests that humans are rather poor at estimating conditional probabilities [Pol+87].

To connect this to the discussion in section 4.2.2, as-if models can serve as templates. In contrast, process models, often implemented as neural networks, are less concerned with idealized rationality and more with how cognition actually operates, including its limitations and biases [McC09]. In practice, modelers seldom specify whether a proposal is meant to be an as-if or a process theory, and the distinction may not always be clear. The same model can contain as-if components, e.g., for tractability reasons, and other parts meant to model actual cognitive processes [Gig20].

The tools we use to model cognition also influence the kind of data we collect and the experiments we design [Gig20]. If we adopt a symbolic or Bayesian approach, we must turn our problem into a simplified, tractable task and manually define discrete variables with which the model can work [MP20]. If we use connectionist models, we usually need to cast the problem as a differentiable task that can be learned from and evaluated with available datasets. In each case, there is a risk of overlooking aspects of cognition that do not fit neatly into one's chosen framework. Cognitive science can mitigate the impact of an individual paradigm's limitations and inherent biases to some degree by employing a plurality of models [CK19]. Different models can complement each other, with the strengths of one model addressing the weaknesses of another [McC09; MP20].

6.2.2 Conflicting goals in engineering and cognitive science

As touched upon in the previous section, there is a certain element of opportunism to modeling [Knu11]. When new technologies become available, they may be integrated into a discipline's modeling practices, shaping how a field frames and answers its research questions [Gig20]. Similarly, when a model has proven successful in reproducing features of some phenomenon, it will often be applied to other phenomena, including phenomena within entirely different disciplines [Knu11]. This can be said to have happened with NNs, which were taken up by the cognitive science community after having shown success in practical applications like image recognition or Natural Language Processing (NLP). However, some scholars have argued that the wholesale adoption of AI models in cognitive science is inadvisable because the goals, priorities, and success criteria in engineering and cognitive science do not necessarily align [McC09; MP20; SK20; Mom23].

Engineering goals are often driven by commercial interests, focusing on short-term, specific, practical applications [PP20]. They typically prioritize performance optimization, often without regard to human-like resource constraints [Has+17]. The focus is on creating the most efficient and powerful systems possible, which may involve leveraging vast computational resources beyond human capabilities [MC23]. In cognitive science, on the other hand, the goal is often to create models that accurately reflect the nuances of human cognition, including its limitations and inefficiencies [SK20]. Cognitive scientists usually engage in basic research that not only may not have immediate practical applications but indeed conflicts (or may conflict) with engineering goals. For instance, researchers may want to model human cognitive constraints, such as limited working memory, attentional bottlenecks, and slower processing speeds, all of which appear undesirable from an engineering perspective [KMK19].

Therefore, effectively adopting NNs for the purposes of cognitive science arguably requires researchers to change their own "reward function" for what constitutes a good model, which may, in turn, cause important changes to the traditional ML pipeline. Specifically, it is important to design ecologically relevant tasks that are informative about underlying cognitive mechanisms and allow for testing hypotheses about mental processes. Performance assessments on these tasks should not be restricted to typical ML metrics like accuracy. Instead, models may, for instance, be evaluated based on their ability to replicate human behavioral patterns, including error patterns, reaction times, and learning trajectories [MP20]. Besides task performance, another relevant assessment criterion may be how well the way a model internally represents and processes information aligns with neuroscientific findings and psychological theories [SK20].

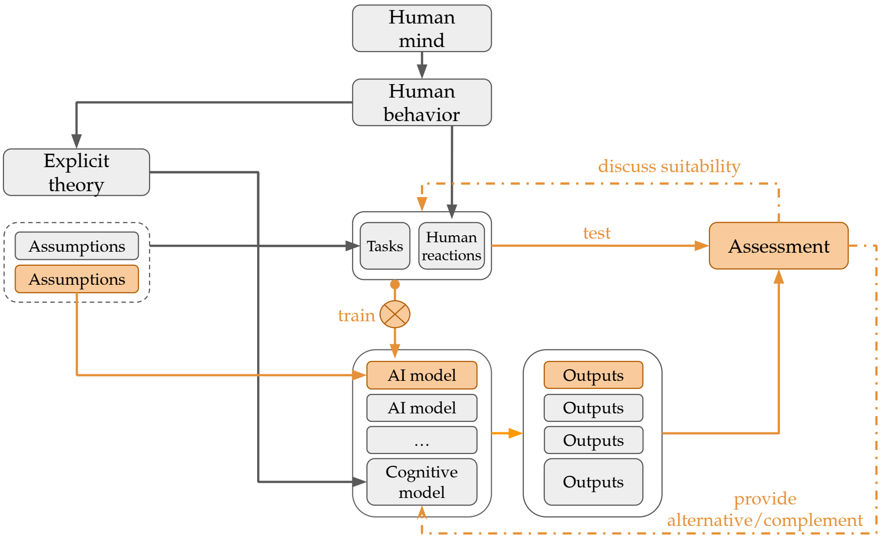


Figure 6.17: Overview of relations between human cognition, theory, assumptions, data, model, and outputs in the fourth case study. Relevant components performed or generated by us shown in orange. Components provided by third parties shown in gray. Circle at the beginning of an arrow indicates the starting point of investigation.

6.2.3 Relation to the guiding questions

Putting this fourth study into perspective, the modeling process (see Figure 6.17) is similar, in a sense, to the gSCAN case. My starting point is a pre-existing benchmark dataset inspired by certain human behavior patterns. In this case, these are not related to systematic generalization but to infants' intuitions about social agents and their actions. However, there are some major differences to the gSCAN case. One difference is that BIB provides reference data from human reactions. In contrast, in gSCAN, the "correct" answers were based on the authors' assumptions about desired behavior.

Another difference is that the human behavior in question has produced not just a collection of empirical findings but inspired an explicit theoretical framework that informed the design of the BIB tasks and one of the baseline models, HBToM. This framework views humans as approximately rational actors and tries to formalize the processes underlying their behavior using Bayesian inference [BST09; JST20]. It can thus be characterized as an as-if theory, which sometimes contrasts with the actual human reference data. For example, infant reactions in the *multi-agent*, *inaccessible-goal*,*inefficient-agent*, and *instrumental-action* tasks contradicted the predictions of the rational Bayesian as-if view. Crucially, this theory does not form the basis of my NN model.

These observations bring us back to the first criticism against NNs: their lack of theoretical grounding. In the case of this study, it is certainly fair to say that the NN is much less rooted in theory than the HBToM model. However, this is, in a way, precisely the point. The NN serves as a proof of concept for how the variables and mechanisms that a cognitive model like HBToM manually encode could emerge on a sub-symbolic level purely from end-to-end training. As discussed in section 6.2.1, modeling paradigms in cognitive sciences have different strengths and limitations, and there is value in fostering a plurality of frameworks that complement each other.

Whereas bottom-up, as-if theories like the one embodied in HBToM have the benefit of being interpretable and data-efficient, they are also relatively rigid and involve manual coding that introduces subjectivity. As-if models are our way of making a phenomenon graspable to our human brains; in this way, they perhaps often reveal more about us than about their target system (which may, confusingly, also be our own cognition). A process model like my NN is less interpretable and requires more data but is not as constrained by pre-defined hypothesized structures. It takes a more descriptive than normative approach and thus can broaden the scope of investigation by revealing unexpected patterns or relationships that may not have "fit into" existing theories.

The fact that the NN model is, in a way, detached from the theory that inspired BIB also allows it to raise questions about the benchmark's suitability for assessing the skills it sets out to measure. As discussed in section 6.2.2, adapting NNs from an engineering context into cognitive science requires changes to the ML pipeline. The BIB dataset is an excellent example of how this can be done. It proposes a set of diagnostic tasks based on cognitive science theory. It also expands on the typical set of evaluation metrics by including VoE accuracy and comparisons to human reference data. As with the gSCAN dataset, comparing different approaches on these tasks can allow for pre-selecting models that show promise in capturing the essence of the phenomena under study. In this way, benchmarking can contribute to developing models that are, if perhaps not biologically accurate, at least more cognitively plausible.

However, as revealed in the error analyses in this case study, the dataset has certain particularities to which the NN adapts during training. For example, the model seems to associate the disappearance of the barrier with specific time steps rather than, as intended, the causal key mechanism. This finding raises the question of what constitutes a desired behavior. If certain time steps always correlate with barriers disappearing in training, should we want the model to learn this correlation or not? Is the dataset implicitly favoring a paradigm that is less sensitive to these peculiarities, perhaps without the authors even realizing this? Would we expect an infant to react similarly if only exposed to this exact training data? If so, what important factors about an infant's experience are missing from the training data? And if not, what critical inductive biases are missing from the model?

By prompting these kinds of questions, an "unprincipled" NN model like the one presented in this case study can help refine what we mean by "human-like" or cognitively plausible behavior and what constitutes ecologically relevant training tasks. Thus, while I do try to provide some "what", "why", and, in the case of the *instrumental-action* task, "where"-level explanations, the main contribution of this case study is not necessarily a form of understanding, but lies instead in bringing out points for further inquiry.

To sum up, models that are not directly based on pre-existing theory can have their benefits. Process models, e.g., in the form of NNs, can allow for a more open-ended exploration than as-if theories. They can provide alternative or complementary explanations to more normative frameworks, prompt questions about what kind of behavior we consider desirable or cognitively plausible, and help us iteratively refine the design of modeling setups that are suited to the goals of cognitive science rather than engineering aims. The case in the next section carries forward several ideas from this study. In particular, it follows the approach of designing cognitively relevant tasks, training a large NN on them, exploring what the model learns, and comparing its behaviors and representations to those of humans. However, it tries to do so with more explanatory depth than the present case study.

7 Modeling the emergence of early number abilities with Vision-Language Transformers

He had bought a large map representing the sea,
Without the least vestige of land:
And the crew were much pleased when they found it to be
A map they could all understand.

"What's the good of Mercator's North Poles and Equators,
Tropics, Zones, and Meridian Lines?"
So the Bellman would cry: and the crew would reply
"They are merely conventional signs!

"Other maps are such shapes, with their islands and capes!
But we've got our brave Captain to thank:
(So the crew would protest) "that he's bought us the best–
A perfect and absolute blank!"

From "The Hunting of the Snark" by Lewis Carroll

7.1 Study

The fifth case study relates to the core system of numbers (see Figure 7.1). Early number skills represent critical milestones in children's cognitive development and are shaped over years of interacting with quantities and numerals in various contexts. Several connectionist computational models have attempted to emulate how certain number concepts may be learned, represented, and processed in the brain. However, these models mainly used highly simplified inputs and focused on limited tasks. We expand on previous work in two directions: First, we train a model end-to-end on video demonstrations in a synthetic environment with multimodal visual and language inputs. Second, we use a more holistic dataset of 35 tasks, covering enumeration, set comparisons, symbolic digits, and seriation.

The order in which the model acquires tasks reflects input length and variability, and the resulting trajectories mostly fit with findings from educational psychology. The trained model also displays symbolic and nonsymbolic size and distance effects. Using techniques from interpretability research, we investigate how our attention-based model integrates cross-modal representations and binds them into context-specific associative networks to solve different tasks. We compare models trained with and without symbolic inputs and find that the purely nonsymbolic model employs more processing-intensive strategies to determine set size.

A version of this case study was published in the journal Cognitive Science [HD24a].

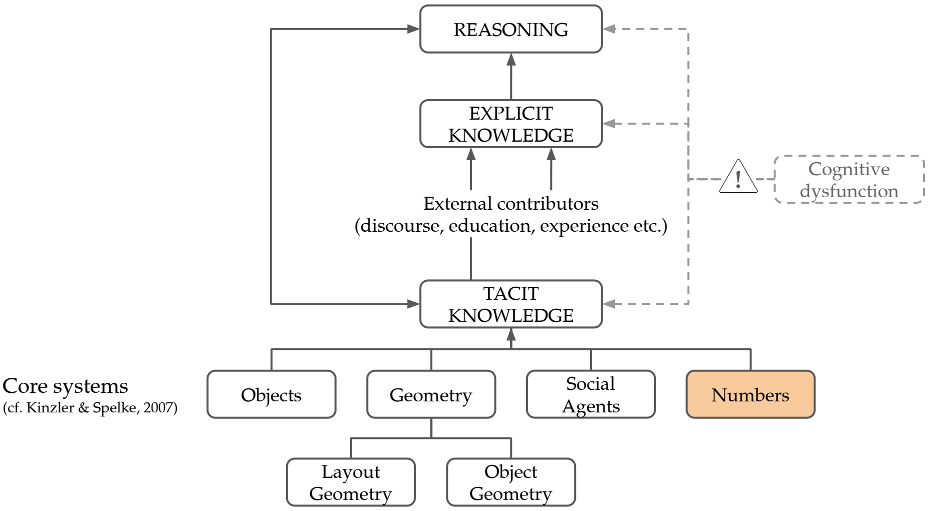


Figure 7.1: Situating the fifth case study in the broader study of cognition. Relevant parts of the framework marked in orange.

7.1.1 Introduction

For many adults, tasks such as counting objects or sorting a set of digits appear simple. For children, however, early number abilities take years to learn. Mastering these skills involves developing a network of concepts that encompasses language, visuospatial abilities, and executive functions [Zha16]. This knowledge later forms the basis for more complex capabilities, e.g., arithmetic. Given the integral role of numbers in our daily lives, questions about how we learn, represent, and process them have occupied cognitive scientists, neuroscientists, and psychologists for decades, forming the multidisciplinary field of numerical cognition.

In this field, connectionist computational models have long played an important part. Often referred to as artificial neural networks, they take inspiration from the way information is stored and processed in the brain via neurons and synapses. As such, they represent concrete implementations of ideas on how at least small subsystems in the brain acquire and process concepts, which can be evaluated against behavioral and neural data. Connectionist models are thus invaluable tools in elucidating critical aspects of learning processes. Their outputs and behavior are inherently shaped by their architecture, training algorithms, and hyperparameters. Additionally, and perhaps more insidiously, these characteristics are influenced by the designers' choice of input modalities as well as the complexity and variety of tasks addressed.

As we show in a brief literature review in section 7.1.2, numerical cognition researchers have been able to reproduce observations from human experimental data using a wide range of approaches. However, many previous computational models operate only on binary images or vectors. When multiple modalities are involved, these are usually processed via specialized modules that are sometimes trained separately. Furthermore, computational modeling studies have mainly focused on a single task type, such as comparing quantities or counting. This setup contrasts with the way humans acquire number knowledge. Children learn through interaction with complex multimodal environments where they encounter number and magnitude concepts in various contexts and concurrently with learning a number of other skills [FRB82].

Faithfully reconstructing a child's brain and experiences is, of course, outside our current abilities. Still, using overly abstracted inputs may artificially impose a stricter separation of input pre-processing and task solving than would naturally occur. Furthermore, considering only isolated skills neglects the interactions between concepts that characterize natural learning and information processing. The main purposes of this work are to introduce a greater but still controlled realism into the modeling of early number abilities and to analyze the points of similarity and difference with empirical research and other models in the literature.

Our approach entails training a model on 35 tasks related to enumeration, set relations, symbolic digits, and seriation. Our goal is not to optimize model accuracy or training times on these tasks – in fact, we are precisely interested in cases where the model struggles or learns more slowly. The tasks draw inspiration from a suite of tests designed to assess young children's early number abilities, which includes hypothesized and empirically validated learning trajectories to serve as comparisons. The model learns end-to-end from video demonstrations in a synthetic environment with visual and language inputs.

Specifically, we examine the following questions: (1) In which order does the model acquire tasks, and how does this compare with findings from educational psychology? (2) On a behavioral level, how do model outputs and error patterns compare with human data and previous computational modeling studies? (3) On a mechanistic level, to what extent does input modality or task specialization emerge in the model? (4) On a behavioral and mechanistic level, what is the effect of removing tasks involving symbolic numbers from the training data?

7.1.2 Related Work

We begin with an overview of previous connectionist models in numerical cognition. Most early connectionist models in numerical cognition focused on numerosity detection and comparison. One of the first such studies was that of Dehaene et al. [DC93]. Their modular architecture processed simple non-verbal visual and auditory inputs using hand-crafted connections and accounted for several psychophysical effects observed in humans. S. A. Peterson et al. conducted a computational study on enumeration and proposed two models, one based on the ACT-R theory [And83] and one a feedforward architecture, which provided good qualitative fits to results obtained in empirical studies [PS00].

Ahmad et al. introduced a multi-network modular system, also focused on determining input numerosity [ACB02]. The architecture used various independently trained neural network types, including recurrent connections and self-organizing maps, and showed some adherence with experimental data from children. Verguts et al. [VF04] and Verguts et al. [VFS05] studied the mental representation of numbers using connectionist models inspired by neuro-scientific findings. They proposed a number representation system using place coding, linear scaling, and constant variability on the mental number line, reproducing error patterns similar to humans on number comparison tasks.

Several computational studies have also focused on spatial aspects of numerical cognition. Mareschal et al. designed a modular cascade-correlation generative network for sorting arrays of numbers [MS99]. Similar to children, the model showed soft stage transitions and variation in performance within stages. Gevers et al. extended the work of Verguts et al. to study the interaction between number and space representations in parity judgment and number comparison tasks [Gev+06]. Their model exhibited the SNARC effect [DBG93], a phenomenon where people tend to respond faster to small numbers located to their left and to large numbers located to their right. Q. Chen et al. further expanded the model, adding hand-crafted biologically inspired layers to represent space explicitly and associate numbers with it [CV10]. The resulting model simulated various experimental data and effects related to spatial attention and dysfunction.

Many initial computational models had relatively few parameters and sometimes involved hand-crafted connections. Recently, researchers have increasingly embraced the paradigm of DL, inspired by the complex, layered organization and functioning of the human cerebral cortex. Stoianov et al. investigated the emergence of visual number sense using a NN trained on binary images [SZ12]. They observed that some neural units acted as "emergent numerosity detectors", resembling the response profiles of monkey parietal neurons.

Since then, several studies have found number-selective neurons even in randomly initialized, entirely untrained NNs [Kim+21; NN21], suggesting that signals that co-vary with numerosity can emerge spontaneously from the statistical properties of bottom-up projections in multi-layered architectures. When explicitly trained on number tasks, a range of NN models have been shown to estimate numerosity at a level comparable to humans. Architectures proposed so far include deep feedforward networks and Differentiable Recurrent Attention Models [Che+18], stacked autoencoders [TZM20], RNNs [She+21], Deep Belief Networks, and Hierarchical CNNs [CSS21].

The computational models discussed so far have been systems trained to classify or reconstruct static inputs usually limited to one modality, such as vision. Several studies have taken a more embodied approach to number learning, exploring the implications of training agents that carry out actions in an environment. Most of these investigations have been in the area of developmental cognitive robotics, where the main focus has been on the benefits of gestures, such as pointing or finger counting, for learning number representations faster, more accurately, and more in line with psychological phenomena observed in humans [RCB11; RCB12; Di +14; De +14; DVC15; Di 17; Di 18; DM19]. Furthermore, Dulberg et al. trained an Emergent Symbol Binding Network on a subset counting task using a two-step training curriculum [DWC21]. Although the model was not physically embodied, it was trained by interacting with an environment via RL.

Most closely related to our work is that of Sabathiel et al. [SMS20b]. Their model consisted of a LSTM and a convolutional LSTM module and was trained on four tasks: counting objects, counting events, reciting numbers, and counting out a subset. The model learned these tasks in a supervised manner in an environment consisting of a 4×4 grid with two binary features at each location, denoting the presence of an object and the agent's hand, respectively. The network developed a strategy of "mentally tagging" objects during counting [SMS20b] and abstract number representations employed across tasks [SMS20a].

We follow a similar approach in that we train a NN on multiple number-related tasks from demonstrations and investigate the model's learned representations. However, we significantly expand the number of tasks and use more complex visual inputs. Motivated by the recent successes of attention-based models in processing sequential, multimodal data, we also use a different architecture, namely, a Transformer. Indeed, we already made use of this architecture in chapters 3 and 6. We provide some background on Transformers in the following section because this is the first time that our research questions require us to introspect *within* the model.

7.1.3 Background: Transformers

Transformers are a type of DL architecture first proposed by Vaswani et al. [Vas+17]. While they originated in NLP, Transformers have since spread to other domains; they are now applied to many forms of data, including images, videos, audio signals, and protein structures [PG23]. They also form the backbone of the now-ubiquitous Large Language Models (LLMs)s. The main change Transformers introduced to the field was a shift from sequential to parallel processing of time series data. Before Transformers, most NLP models used RNNs. In an RNN, inputs, such as tokenized words or characters, are added one after the other. The model then learns which inputs and intermediate computation results to retain for how long in order to succeed on a given task. To do this, it must update its hidden states after each time step, as they form the inputs for subsequent calculations.

In contrast to RNNs, Transformers receive an entire context window, such as a sentence or paragraph, at a time. They maintain access to all the information in this window without having to learn to "remember" it. Because a transformer essentially treats all time steps independently, it can process them in parallel, leading to considerably faster computation than the recurrent approach. The main units that carry out the input processing are a transformer's attention heads. As the following sections presuppose an understanding of the attention mechanism, we seek to provide some intuition on the topic with a toy example.

In Figure 7.2, we illustrate the workings of a single attention head with the following toy task: The model receives a visual input consisting of a circle, rectangle, or triangle, which may be red, blue, or green. It is asked about this input's shape or color. Let us assume that our inputs are a green triangle and the question "What color is this?". We first translate language and vision inputs into binary vectors E_{Lang} Ⓐ and E_{Vis} Ⓑ. Note that this is a simplification for illustrative purposes, not how we encode visual inputs in our actual model (see section 7.1.4). The binary vectors serve as inputs to the attention head.

One attention head consists of five single-layer neural networks: W_V, W_K, W_Q, W_O, and W_{Pred}. The networks' weights are initially random and learned through training on question-answer pairs via backpropagation. In our illustration, model weights have already been optimized. Each network serves a different function. W_K Ⓒ receives language input and produces activation vectors K Ⓓ. W_Q Ⓔ receives visual input and produces an activation vector Q Ⓕ. Inspired by information retrieval terminology, K and Q are referred to as "keys" and "queries"[Vas+17]. Query vectors represent what the model is looking for, whereas keys act as signals to match against the queries.

Because W_K and W_Q have the same number of output neurons $d = 64$, K and Q have the same dimensionality and can be combined via their inner product. This combination allows the model to relate information from both modalities. We divide the key-query product by a scaling factor $\sqrt{d_k}$ and apply a softmax function to keep values between 0 and 1. The result, $A(Q, K)$ Ⓖ, is often referred to as an "attention heatmap" [RCW15]. It shows the strength of the match between the query and each key. $A(Q, K)$ is specific to the context, i.e., combining the same question with another visual input would result in a different heatmap.

$A(Q, K)$ is combined with the output V Ⓘ (values) of the value network W_V Ⓗ. Analogous to how values in databases are the actual data associated with a key, a "value" in the attention mechanism is a transformed representation of the input (in this case E_{Lang}) that contains the actual content to be focused on. Multiplying $A(Q, K)$ with V yields the attention output $A(Q, K, V)$ Ⓙ, which represents the weighted sum of values, where the weights are determined by the attention heatmap.

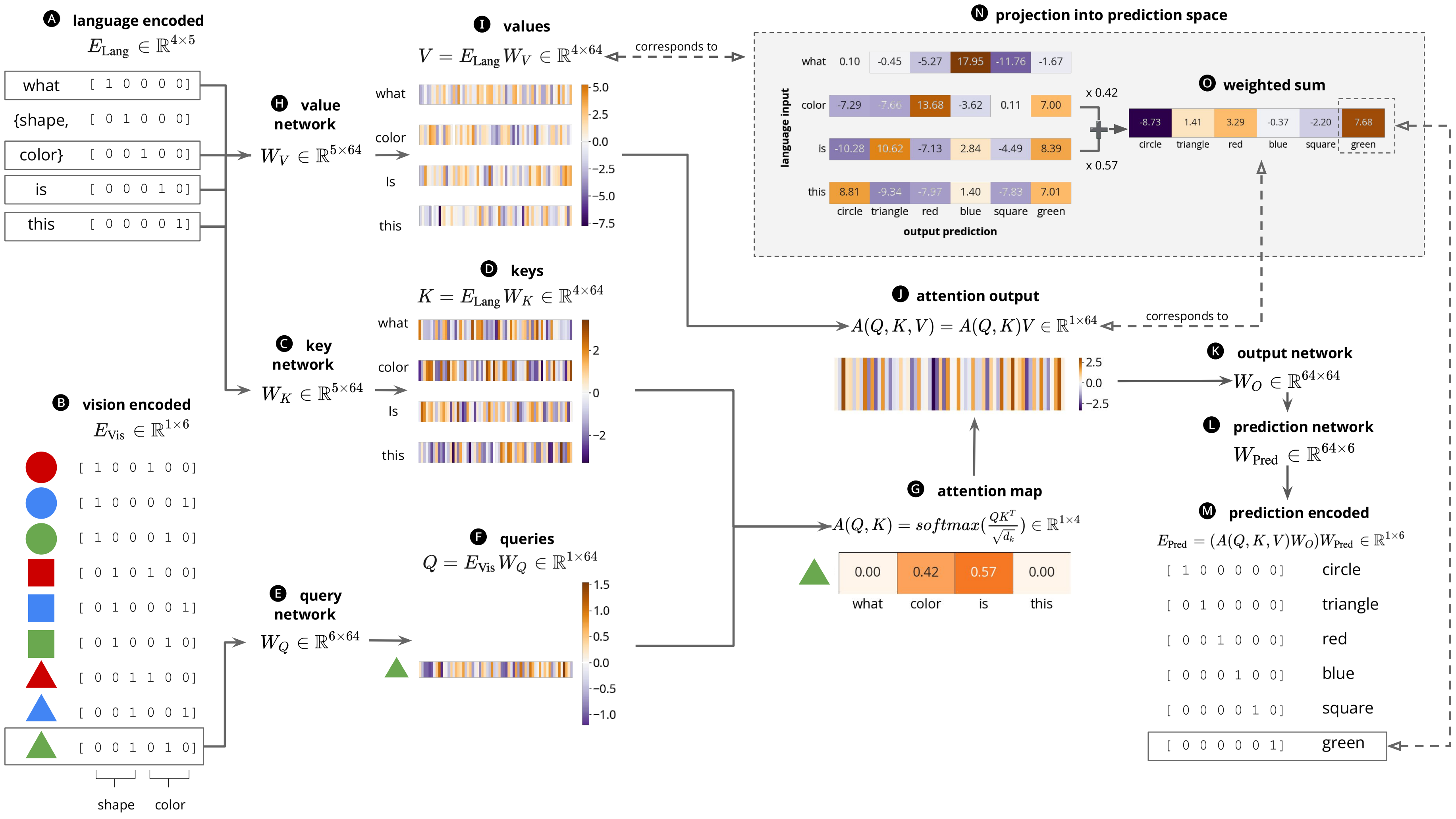

A language encoded
$E_{\text{Lang}} \in \mathbb{R}^{4\times5}$
what [1 0 0 0 0]
{shape, [0 1 0 0 0]
color} [0 0 1 0 0]
is [0 0 0 1 0]
this [0 0 0 0 1]
B vision encoded
$E_{\text{Vis}} \in \mathbb{R}^{1\times6}$
[1 0 0 1 0 0]
[1 0 0 0 0 1]
[1 0 0 0 1 0]
[0 1 0 1 0 0]
[0 1 0 0 0 1]
[0 1 0 0 1 0]
[0 0 1 1 0 0]
[0 0 1 0 0 1]
[0 0 1 0 1 0]
shape
color
H value network
$W_V \in \mathbb{R}^{5\times64}$
C key network
$W_K \in \mathbb{R}^{5\times64}$
E query network
$W_Q \in \mathbb{R}^{6\times64}$
I values
$V = E_{\text{Lang}} W_V \in \mathbb{R}^{4\times64}$
what
color
Is
this
5.0
2.5
0.0
−2.5
−5.0
−7.5
D keys
$K = E_{\text{Lang}} W_K \in \mathbb{R}^{4\times64}$
what
color
Is
this
2
0
−2
F queries
$Q = E_{\text{Vis}} W_Q \in \mathbb{R}^{1\times64}$
1.5
1.0
0.5
0.0
−0.5
−1.0
corresponds to
N projection into prediction space
language input
what 0.10 -0.45 -5.27 17.95 -11.76 -1.67
color -7.29 -7.66 13.68 -3.62 0.11 7.00
is -10.28 10.62 -7.13 2.84 -4.49 8.39
this 8.81 -9.34 -7.97 1.40 -7.83 7.01
circle triangle red blue square green
output prediction
x 0.42
x 0.57
O weighted sum
-8.73 1.41 3.29 -0.37 -2.20 7.68
circle triangle red blue square green
J attention output
$A(Q,K,V) = A(Q,K)V \in \mathbb{R}^{1\times64}$
corresponds to
2.5
0.0
−2.5
G attention map
$A(Q,K) = softmax(\frac{QK^T}{\sqrt{d_k}}) \in \mathbb{R}^{1\times4}$
0.00 0.42 0.57 0.00
what color is this
K output network
$W_O \in \mathbb{R}^{64\times64}$
L prediction network
$W_{\text{Pred}} \in \mathbb{R}^{64\times6}$
M prediction encoded
$E_{\text{Pred}} = (A(Q,K,V)W_O)W_{\text{Pred}} \in \mathbb{R}^{1\times6}$
[1 0 0 0 0 0] circle
[0 1 0 0 0 0] triangle
[0 0 1 0 0 0] red
[0 0 0 1 0 0] blue
[0 0 0 0 1 0] square
[0 0 0 0 0 1] green

Figure 7.2: Toy example illustrating the workings of a single attention head.

We pass $A(Q,K,V)$ through the output network W_O Ⓚ and feed the result to the prediction network W_{Pred} Ⓛ. This gives us the correct answer to the question: "green" Ⓜ. While the activation vectors in Figure 7.2 are not human-interpretable, we can translate them into intermediate predictions by directly inputting them to W_O and W_{Pred}. Doing this for V shows that each word in the question triggers different answers Ⓝ. E.g., the "color" vector activates the output "red". This pairing is arbitrary – with a different random weight initialization, "red" might, e.g., be maximally activated by "this". If we linearly combine the activations according to our attention heatmap $[0.0 \quad 0.42 \quad 0.57 \quad 0.0]$, we obtain a vector Ⓞ that translates to the correct output "green". $A(Q,K)$ can thus be seen as a "selector" of the most likely answer among the options encoded in V, based on the linguistic and visual context.

7.1.4 Methods

Tasks Our dataset is based on a curriculum proposed by Resnick et al. [RWK73]. Inspired by Gagné's framework of "learning hierarchies" [Gag68], the authors operationalized early number concepts as a suite of tasks, ordered by what they hypothesized to be an optimal match for children's natural sequence of acquisition. In two empirical studies, they turned many of these tasks into diagnostic tests, which they administered to pre-kindergartners, kindergartners, and students in their second week of elementary school [WRB71; Wan73]. Thus, the children had been exposed to little or no formal maths education at the point of testing. The authors applied multiple scalogram analysis [Lin63] to the test scores to identify dependencies in the relationships among children's abilities. They then compared the empirical patterns of acquisition they found against their hypothesized learning hierarchies. Resnick et al.'s task suite constitutes an excellent basis for our dataset, as it encompasses a wide range of skills related to the concept of number, including hypothesized and, in part, psychometrically validated results from human studies. The suite covers enumeration, set comparison, symbolic numerals, and sorting.

Table 7.1 gives an overview of the tasks we used, ordered by difficulty as hypothesized by Resnick et al. Table 7.2 shows the developmental trajectories in children's learning found by M. C. Wang et al. and M. C. Wang for those tasks that were psychometrically validated. The grouping into task families in Table 7.1 is not a perfect partition, and some tasks may integrate skills from other task types. We distinguish between three numerical concepts that may be involved in a task: quantity, rank, or label [Nie05]. Quantity refers to cardinality, i.e., number of elements in a set. Rank refers to the serial order of an element. In label tasks, numbers are used categorically to identify an object. As can be seen, the dataset encompasses all three usages of number and some tasks that do not explicitly involve numbers but are believed to support the acquisition of number concepts. We go through the tasks in more detail in the following, starting with those related to enumeration.

The first three tasks introduce two important counting principles. A1 asks the agent to recite the count list, starting and stopping at a specified number. Knowing the number sequence, the so-called stable order principle [GG86], is a crucial numerical concept and arguably the first mathematical skill a child acquires [SMS20b]. Children typically learn this principle over several years, between ages 2 and 6 [Mus+14]. In A2, the agent must point at each object in a set exactly once. A3 combines A1 and A2. It requires the agent to say the correct count word as it touches each object – the so-called one-to-one principle [GG86]. In the original curriculum, the child can remove counted objects to decrease the strain on working memory. In our computational implementation, objects disappear after being grabbed and released.

The last four tasks involve the enumeration of fixed sets. A4 and A5 are analogous to A3, except objects do not disappear after being tagged. The set is linearly arranged in A4, reducing the difficulty of tracking which objects have been counted [PL68; SES74]. In A6, the agent must touch a stated number of objects without uttering any number words, then stop. A6 is a version of the give-N task, which has been used in previous studies of children [SC08; Wyn92] and neural networks [SMS20b; DWC21]. In A7, the agent must point at a set of a given size, selecting from two to five options. Unlike the other tasks in this unit, which are primarily concerned with the rank of an element in the count sequence, A6 and A7 require determining the cardinality of a set without counting aloud.

The second unit involves comparing quantities. In B1 and B2, the agent must point at one of two sets containing more or fewer objects, respectively. Resnick et al. considered B2 more challenging than B1, arguing that B2 requires finding a set with extra objects, then choosing its counterpart. It thus involves negative information, which can be difficult for young children. In B3 and B4, the agent receives a digit and an object set and must point at whichever represents the higher (B3) or lower (B4) number. In B5 and B6, inputs consist of five digits and one set. The agent must point at all digits denoting numbers larger (B5) or smaller (B6) than the set. In B7 and B8, the agent must decide which of two rows of objects contains more (B7) or fewer (B8) objects. This task is reminiscent of the Piagetian number conservation test, where two sets are linearly arranged such that equivalence is easy to determine via 1-to-1 comparison. The arrays are then spaced differently to test whether a child still recognizes the sets' equivalence [PGH52]. B9 and B10 are analogous to B1 and B2 but involve three sets.

The third unit relates to symbolic numerals. Children have been shown to start recognizing and manipulating Arabic digits at around 4 or 5 years of age [GMS07; KKL13; Mus+14; Li+18]. In the first three digit tasks, numbers serve a purely nominal role. In C1, the agent receives one to five pairs of digits and needs to match them by placing corresponding digits atop each other. In C2, the agent must point at one of five numerals denoting a stated number. In C3, the agent is asked to state the name of a given digit. C4 is analogous to A7, except the subset size specification is now given by a digit rather than a number word, connecting the numeral to set cardinality for the first time. The following three tasks are ordinal tasks concerned with relations between numbers. C5 and C6 require the agent to point at the larger and smaller of two digits, respectively. In C7, the agent must sort two to four digits in ascending order by dragging them into the correct linear configuration. C8 is similar to A7, except for the set size being denoted by a digit.

The last set of tasks is related to sorting, one of the skills thought to mark a child's entrance into the stage of concrete operations [Res73]. Although most tasks in this unit involve magnitudes rather than numerosity, it has been suggested that seriation is an essential ability for understanding the properties of number [Pia61]. Sorting is generally considered a difficult skill to acquire, learned around 7-8 years of age [Jes78; MK19]. D1, D2, D5, and D6 require the agent to point at the largest, smallest, darkest, or lightest object in a set, respectively. Resnick et al. considered these tasks prerequisites for D3, D4, D7, and D8, where the agent must sort two to six objects according to size by placing them in the correct order. In D3 and D7, objects differ only in the attribute according to which they are to be sorted. In D4 and D8, they vary in more attributes, e.g., shape, size, and luminance. Adding irrelevant cues to objects should make seriation more challenging [TK97]. D9 requires the agent to seriate two to four whole sets by their size and thus involves both cardinality and rank. In D10, objects are arranged in one or two rows. The agent must verbally specify the ordinal position of a pointed-to object.

Table 7.1: Overview of the task types in our dataset.

		Visual input		Representation of relevant number/magnitude					Numerical concept		
ID	**Task description**	**Number of sets**	**Number of digits**	**Word (symbolic)**	**Digit (symbolic)**	**Set size (analog)**	**Object size (analog)**	**Object color (analog)**	**Quantity (cardinal)**	**Rank (ordinal)**	**Label (nominal)**
A1	recite count list	0	0	x						x	
A2	touch all fixed, unordered Os in turn	1	0								
A3	enumerate movable Os	1	0	x						x	
A4	enumerate fixed, ordered Os	1	0	x						x	
A5	enumerate fixed, unordered Os	1	0	x						x	
A6	touch exactly {number word} Os	1	0	x					x		
A7	select set of size {number word}	2-5	0	x		x			x		
B1	select larger of two sets	2	0			x			x		
B2	select smaller of two sets	2	0			x			x		
B3	select larger of a digit and a set	1	1		x	x			x		
B4	select smaller of a digit and a set	1	1		x	x			x		
B5	select all digits larger than size {set}	1	5		x	x			x		
B6	select all digits smaller than size {set}	1	5		x	x			x		
B7	select larger of two rows	1	0			x			x		
B8	select smaller of two rows	1	0			x			x		
B9	select largest of three sets	3	0			x			x		
B10	select smallest of three sets	3	0			x			x		
C1	match digits	0	2-10		x						x
C2	select stated digit	0	5	x	x						x
C3	name given digit	0	1	x	x						x
C4	select set of size {digit}	2-5	1		x	x			x		
C5	select larger of two digits	0	2		x					x	
C6	select smaller of two digits	0	2		x					x	
C7	sort digits	0	2-4		x					x	
C8	touch exactly {digit} Os	1	1		x				x		
D1	select largest O in set	1	0				x				
D2	select smallest O in set	1	0				x				
D3	sort Os by size (Os differ in 1 attribute)	1	0				x	x			
D4	sort Os by size (Os differ in > 1 attributes)	1	0				x	x			
D5	select darkest O in set	1	0					x			
D6	select lightest O in set	1	0					x			
D7	sort Os by color (Os differ in 1 attribute)	1	0				x	x			
D8	sort Os by color (Os differ in > 1 attributes)	1	0				x	x			
D9	sort sets by size	2-4	0			x			x	x	
D10	name position of an O in ordered set	1	0	x						x	

Table 7.2: Comparison of hypothesized and observed developmental trajectories in children's learning, based on M. C. Wang et al. and M. C. Wang. To be read from left to right. Only psychometrically validated tasks with direct counterparts in the current study are shown.

Hypothesized trajectory	A1	A3	A4	A5	A6	A7	B1	B2	B9	B10	C1	C2	C3	C4	C5	C6	C7	D9
Empirical trajectory for numbers zero to five	C1	A1	A3	A5	A4	B2	B1	~~B9~~*	~~B10~~*	C2	A7	A6	C3	C4	C5/C6†		C7	D9
Empirical trajectory for numbers six to ten	C1	A1	B1	B2	~~B9~~*	~~B10~~*	A4	A3	A5	A6	C2	C3	A7	C4	C7	C5/C6†		D9

∗ Excluded because too few subjects mastered the task

† Not distinguished in psychometric analysis

Data generation We translate the tasks of Resnick et al. into an environment of 259 × 259 pixels with 4×4 black panels, each of size 64×64. The panels are separated by white lines of width 1 pixel and can contain one to ten gray-scale objects or a digit from one to ten, depending on the task. Objects can be rectangles, triangles, circles, or ellipses. They are randomly assigned sizes, luminances, and positions. Sizes vary between 8 and 32 pixels in height or width. Luminances vary between 0.1 and 1.0 to ensure sufficient contrast with the background. Objects are initially non-overlapping, but occlusion can occur as the agent moves them around. We represent the agent with the icon of a yellow hand spawned in the upper left corner of a random panel at environment initialization. The hand can be in one of three states: open, pointing, or grabbing. At each time step, there are a total of 24 output options.

The agent can move up, down, left, or right by either small, 8-pixel, or large, 64-pixel steps. It can interact with its environment by grabbing or releasing an object, grabbing or releasing a whole set, or pointing. It can also output number words from one to ten and the word "stop". Unlike previous work, where task IDs were encoded via binary vectors, we prompt the agent with language inputs such as "sort the numbers" or "which row has fewer objects". For each task, we collect 10,000 training examples, 1,000 test examples, and 500 validation examples using a solver which produces demonstration sequences deterministically.

The solver navigates to its target panel in 64-pixel steps, moving first to the correct row, then to the correct column. If necessary, it moves on to its target object within the set, following the same logic. In enumeration tasks, it targets the next untagged object that is closest horizontally, then vertically. If two objects have the same distance, it prioritizes objects to the right, resulting in a row-wise tagging order. This is representative of linear spatial strategies employed by older children [Sha78; WFC87] and adults [PL68] in enumeration tasks. For tasks C1, B5, and B6, the solver targets the next eligible panel with the smallest Manhattan distance to the agent and prioritizes panels below, above, to the right, and the left, in that order. When sorting objects (D3, D4, D7, D8), it goes from darkest or smallest to lightest or largest and places them next to each other at the top of the panel. It orders them from left to right, which is the preferred seriation order in many industrialized groups [Pit+21]. Sorting whole panels (C7, D9) works similarly, but blocking panels may first have to be removed from the grid's top row.

We programmatically checked for and removed any exact duplicates in the training, test, and validation sets. For some tasks, such as A1 and C3, duplicates were unavoidable due to the limited number of task configurations. In these cases, we held out certain combinations that we only allowed to occur in the train, test, or validation split, respectively. We upsampled these combinations such that the overall number of examples remained the same across tasks. Depending on the task, we ensured a uniform distribution of set sizes, prompts, or the number of non-empty panels. This runs counter to the suggestion of S. T. Piantadosi that the developmental trajectory of number knowledge in children is influenced by the Zipfian distribution of numbers they encounter in everyday experience [Pia16]. However, Testolin et al. found that human-like psychophysical effects also occurred for NNs trained with flat number frequencies [TZM20].

We constructed two additional datasets. The first consists of test tasks B1, B2, and D9, with the difference that one set has 11-15 objects rather than one to ten. We use this to test the model's extrapolation on comparison tasks to larger set sizes. The second excludes all tasks involving digits or number words. The construction of this dataset was motivated by proposals in the literature that language plays a key role in learning numeracy skills [TV14; HSP18; PR16] and is a prerequisite for forming certain concepts [GG04; Car11]. Support for this idea comes from studies of cultures without words for larger, exact quantities, such as the Pirahã, the Mundurukú, the Tsimane, and Nicaraguan Homesigners. In adults from these cultures, the ability to represent exact numbers has been found to be limited to the range for which verbal labels are available [PGP22]. We are therefore interested in the effect that training a model only on nonsymbolic tasks has on its performance and inner representations.

Model Having described the tasks we aimed to solve, we now present the architecture we designed to do so. Figure 7.3 shows a visualization of the model. The example in 7.1.3 illustrated the workings of a single attention head – our full model has 512: four attention blocks, each containing eight so-called attention layers with 16 heads. Each head can be thought of as a specialized unit that learns to focus on specific aspects of the input data during training. Using multiple heads in an attention layer allows the model to focus, in parallel, on different aspects of the input within one processing step (where a processing step is all the computations performed in one attention layer). The outputs of all heads in an attention layer are concatenated and then transformed linearly. This aggregation synthesizes the information from all heads.

The result is passed through a feed-forward block, which consists of a small two-layer neural network. Inputs to the attention heads and the feed-forward block first undergo normalization. Normalization and feed-forward block were omitted from the example in Figure 7.3 for simplicity but are commonly used components of attention layers in deeper models as they have been found to stabilize training [Lin+22]. We also employ so-called "residual connections," where the attention layer's output is added to its original input before being passed to the next layer [He+16]. This approach allows later heads to operate on both original inputs and results from previous heads. Multiple attention layers in an attention block enable the model to learn increasingly abstract representations of the input data.

Similar to our toy example, the model receives language and visual inputs. The language input consists of a question or instruction. The visual input is a series of video frames showing the demonstration sequence produced by the deterministic solver.

To pre-process the language input, we encode each word into a binary vector, analogous to the example in Figure 7.2. To pre-process the video frames, we extract regions of interest (ROIs), which may contain individual objects, digits, or an entire panel of objects (Figure 7.3 Ⓐ). Such an ROI-based Transformer approach has previously been applied to tasks like visual navigation [DYZ21b]. We find our ROIs by identifying contours through morphological transformations and thresholding. Specifically, we extract ROIs by eroding each video frame with a 2×2 kernel, dilating it with a 1×1 kernel, and applying a binary threshold of value 15. 15 is the darkest RGB value objects can take in our task environment. We then apply the

Douglas-Peucker algorithm [DP73] to obtain object contours and their bounding boxes. We found that this yields ROIs of sufficient quality for our task environment; for more naturalistic inputs, CNN-based object detectors could be used. We resize all ROIs to RGB patches of size $28 \times 28 \times 3$, then flatten them into 2,352-dimensional vectors. We limit the maximum number of ROIs per frame to 85 due to computational constraints.

Having converted our linguistic and visual inputs to vector form, we feed them into separate embedding layers (Figure 7.3 Ⓑ and Ⓒ) with 60 and 48 output neurons, respectively. These are single-layer neural networks, which produce an activation vector, or "embedding", for each input. So far, those vectors contain no positional information. Therefore, we concatenate the visual embedding of each ROI with its central x and y coordinates and original width and height. For each word embedding, we append a sinusoidal 16-dimensional encoding [Vas+17] representing its relative position in the sentence. The result is a set of 64-dimensional visual and linguistic embeddings. They are passed into the first attention block alongside two special inputs: the class token `CLS` and the memory token `MEM`. The `CLS` contains the model's prediction, i.e., which action to take. The `MEM` vector compresses relevant information in each time step to be used later in the model. These are initially random, "blank" vectors, which each attention layer can modify by adding its output to them.

The first two attention blocks integrate language and visual information for individual frames. As in the example in Figure 7.2, query networks in the first block's first attention layer receive visual input, and key and value networks receive language input (Figure 7.3 Ⓓ). Merging multiple modalities in this way is referred to as cross-attention. A self-attention block follows (Figure 7.3 Ⓔ). Self-attention means that inputs do not come from different sources (e.g., vision and language). Instead, query, key, and value networks all receive the same inputs – in this case, the outputs from the first block. Up to this point, we process all frames in parallel but separately. I.e., each frame is treated independently from previous frames. However, many tasks set out in section 7.1.4 require knowledge of past time steps.

We address this need for a memory mechanism with the last two attention blocks. In the third block (Figure 7.3 Ⓕ), we give the model access to inputs from past frames. The query networks of this block's first attention layer receive the `CLS` tokens output by the second block. The key and value networks receive the `MEM` tokens, concatenated with temporal position encodings (analogous to the word embeddings). We do this because there can be up to 85 ROIs in a frame and up to 100 frames in a video. Due to the quadratic complexity of the matrix multiplications involved in the naive attention mechanism, attending over every object of every previous frame would be computationally prohibitive. By forcing the model to compress relevant information into a single `MEM` vector per time step, we only need to attend over up to 99 instead of 99×85 vectors.

In the last attention block, we give the model access to its past outputs. This information is, e.g., important for tasks that involve counting. Similar to the language input, past actions are converted to binary vectors and processed by an embedding layer (Figure 7.3 Ⓖ) to yield 48-dimensional embeddings, which we concatenate with 16-dimensional temporal position encodings. These action embeddings serve as input to the key and value networks in the fourth block's first attention layer (Figure 7.3 Ⓗ). Query networks receive the `CLS` tokens output by the third attention block (one for each time step). Finally, the `CLS` tokens are processed by an output layer 7.3 Ⓘ), yielding a sequence of action predictions.

Training We trained four models in total. The first three were trained on the dataset containing non-symbolic and symbolic tasks. We used multiple models to determine whether

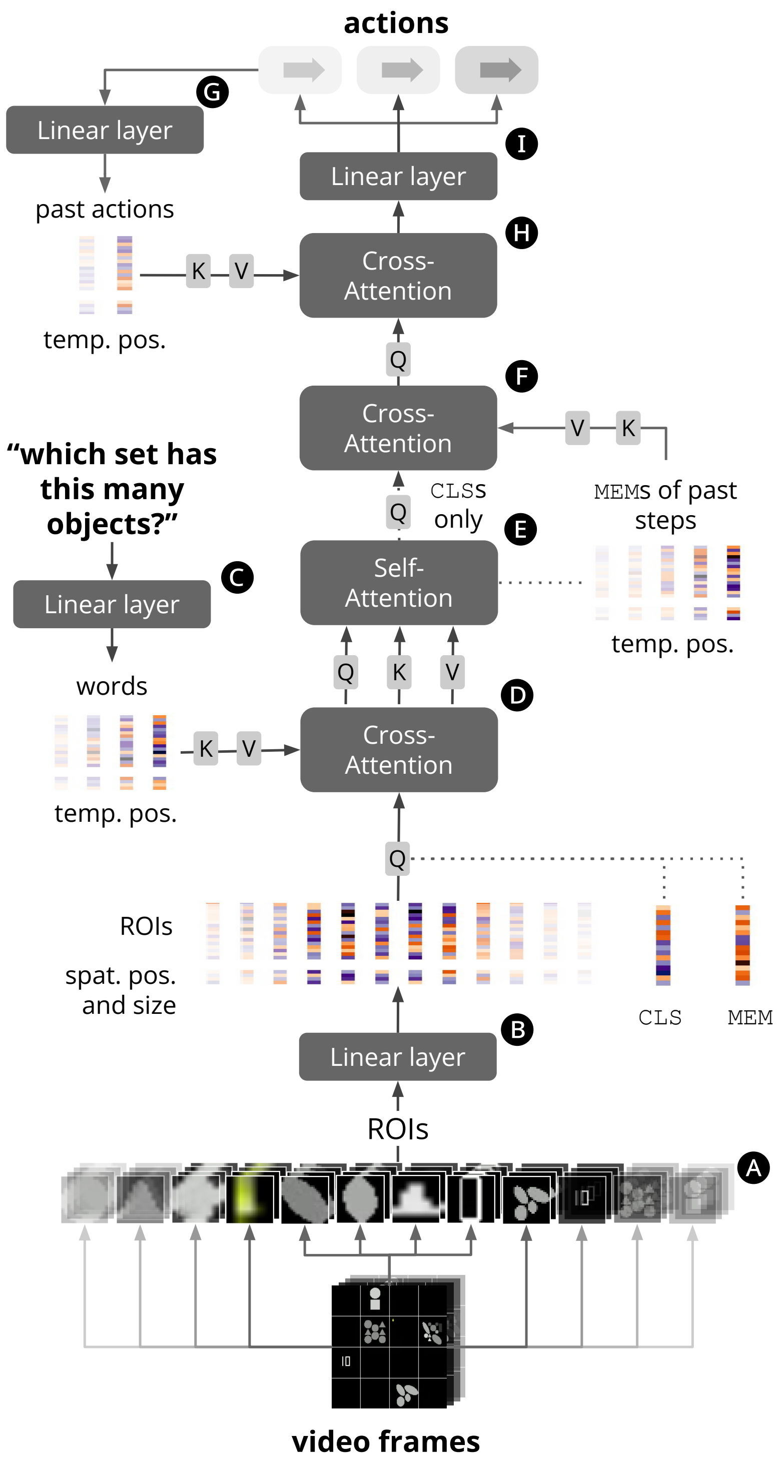


Figure 7.3: Schematic of our attention-based model. Inputs consist of Regions of Interest extracted from each frame in the demonstration videos and a language prompt. They are processed via four attention blocks, the first two of which attend over a single time step. The third and fourth blocks take into account past inputs, compressed into the special MEM token, and the model's past actions, respectively.

Table 7.3: Model accuracy on the test set. Tasks are considered solved correctly if the model's predictions are identical to the deterministic solver's action sequence.

Counting and enumeration			**Set comparison**			**Numerals**			**Seriation and ordinal position**		
ID	**S**	**N-S**	**ID**	**S**	**N-S**	**ID**	**S**	**N-S**	**ID**	**S**	**N-S**
A1	0.99 ± 0.01	-	B1	0.99 ± 0.00	0.99	C1	0.91 ± 0.00	-	D1	0.93 ± 0.00	0.91
A2	0.97 ± 0.01	0.86	*B1+*	*1.00 ± 0.00*	*0.99*	C2	1.00 ± 0.00	-	D2	0.96 ± 0.00	0.95
A3	0.97 ± 0.01	-	B2	0.99 ± 0.00	0.98	C3	1.00 ± 0.00	-	D3	0.86 ± 0.01	0.79
A4	1.00 ± 0.01	-	*B2+*	*1.00 ± 0.00*	*0.99*	C4	0.80 ± 0.02	-	D4	0.82 ± 0.01	0.74
A5	0.97 ± 0.01	-	B3	1.00 ± 0.00	-	C5	1.00 ± 0.00	-	D5	1.00 ± 0.01	0.99
A6	0.95 ± 0.00	-	B4	1.00 ± 0.00	-	C6	1.00 ± 0.00	-	D6	0.98 ± 0.00	0.98
A7	0.82 ± 0.02	-	B5	0.91 ± 0.01	-	C7	0.99 ± 0.01	-	D7	0.99 ± 0.00	0.97
			B6	0.88 ± 0.01	-	C8	0.96 ± 0.01	-	D8	0.98 ± 0.00	0.96
			B7	1.00 ± 0.01	0.99				D9	0.83 ± 0.03	0.64
			B8	1.00 ± 0.00	1.00				*D9+*	*0.84 ± 0.03*	*0.69*
			B9	0.96 ± 0.01	0.91				D10	1.00 ± 0.00	-
			B10	0.95 ± 0.02	0.87						

S models trained on symbolic and non-symbolic tasks
N-S trained on non-symbolic tasks only
+ datasets requiring extrapolation to larger sets of size 11-15

they would display similar final accuracies and training trajectories. The models shared the same architecture and training setup, but their random weight initialization differed. Due to the random shuffling of the dataset, they also received training samples in a slightly different order. The fourth model was trained on the dataset containing only non-symbolic tasks in order to investigate whether it would display differences in performance or internal representations. All models were implemented in PyTorch [Pas+19]. They were trained to predict the deterministic solver's next output at each time step using cross-entropy loss, which is a measure of the difference between model predictions $\hat{y}$ and correct answers y:

$$L(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_i) \tag{7.1}$$

We used the Rectified Adam optimizer [Liu+20] and gradually adjusted the learning rate using a schedule with cosine annealing and warm restarts [LH17]. The scheduler exponentially decayed the learning rate from an initial value of 0.005 to 0.0002 over four passes through the dataset (epochs), after which it was kept constant. This annealing scheme served to speed up initial training. To prevent the model from memorizing the training data too much (overfitting), we used dropout. Dropout is a technique where randomly selected neurons are temporarily disabled to prevent overreliance on individual units. We used a dropout probability of 0.1. We also applied early stopping, meaning we performed validations after every half epoch and stopped training if the model had not improved over three checks. Performance usually stagnated after around 28 epochs. We trained the models in batches of 512 samples at a time. Each epoch, including validation, took ca. seven hours on a 16-core AMD EPYC 7282 server with six GeForce RTX 2080 GPUs.

7.1.5 Results

Overall performance As shown in Table 7.3, the model performs well on most tasks, with an overall average accuracy of 93%. Variation across models trained on the same tasks is minimal, indicating that performance is not sensitive to weight initialization or batch ordering. When tested on comparison and seriation tasks with sets of larger size than seen in training, performance is slightly higher, presumably because of the increased contrast between set sizes. There are, however, tasks on which it consistently reaches lower accuracies, namely, A7, B5, B6, C4, D3, D4, and D9. A7, B5, B6, C4, and D9 all require the integration of several subskills: determining the cardinality of, in the case of A7, C4, and D9, up to five sets and comparing them against either a number or multiple other sets, as well as keeping track of already tagged or obscured panels. Transformers have no recurrent connections; thus, their number of attention layers determines the number of "reasoning" steps they can perform. While the model reaches high accuracy on prerequisites such as, e.g., comparing two sets (B1 and B2), the abovementioned tasks that require multi-step combinations of such subtasks appear to strain its capacity.

The lower performance when sorting objects by size (D3, D4) seems to be due to an issue with size discrimination, as the model successfully sorts objects by luminance (D7, D8). In our environment, an object's size equals its surface area, and differences may be as minor as a few pixels, whereas we enforced larger spacings for color. The model, therefore, needs to retain very granular information about each shape. This may be why the model's accuracy when choosing the smallest or largest object (D1, D2) is 2-7% below its accuracy for choosing the lightest or darkest object (D5, D6). Errors compound when the model has to compare up to six objects during seriation.

The model trained only on non-symbolic tasks (denoted as N-S in Table 7.3) does about as well as the model trained on the full dataset on most tasks related to object attributes, namely D1, D2, D5, D6, D7, and D8. The lower performance on tasks D3 and D4 can again be ascribed to an issue of size discrimination – while the accuracy on tasks D1 and D2 is only 1-2% below the S model, the difference compounds in the case of seriation. The N-S model also achieves similar accuracies as the S model on the two-set comparison tasks B1, B2, B7, and B8, including in the case of extrapolation to sets of larger size.

The fact that its performance is not affected by a lack of symbolic training is in line with findings that it is feasible to compare the cardinality of sets without having mastered symbolic counting; see studies on cultures with a smaller number lexicon [Pic+04b], nonverbal infants [Xu03], animals [BT98; HCH00; NFM02; Dad+09], and neural networks without counting knowledge [DC93].

However, the N-S model achieves lower performance than the S model for pointing out all objects in turn (A2), comparing three sets (B9, B10), and seriation by set size (D9). In the case of A2, this drop might be because, without the inclusion of the enumeration tasks A3-A6, the proportion of tasks that require going through a set one by one is lower, thus putting less emphasis on this skill. In the case of B9, B10, and D9, part of the issue may be that, without symbolic tasks, the model has less exposure to tasks that involve multiple non-empty panels. We also hypothesize that the S model can parse and use numerosity information more efficiently. We investigate this idea further in section 7.1.5.

Training trajectories In addition to the final accuracies reached by the model, we are interested in the order in which the model's performance progresses on the different tasks and whether this aligns with findings from educational psychology. As mentioned in section 7.1.4, our dataset was randomly shuffled. While this contrasts with the sequential way children encounter tasks, neural networks trained on multiple tasks simultaneously have been found to consistently learn easier samples first [Gra+17; WDN21]. This allows us to compare the "implicit curriculum" that emerges for our models with the order of acquisition empirically found by Resnick within and across task families. We show the development of model performance for each task in the course of training in Figure 7.4.

The order of acquisition for the enumeration tasks follows the order found by M. C. Wang et al. for numbers from six to ten (see Table 7.2: the count list is learned first (A1), followed by counting ordered objects (A4), movable objects (A3), unordered sets (A5), subsets (A6), and finally choosing a set of specified size (A7). M. C. Wang et al. did not validate the task of touching each object in turn (A2), but this skill was hypothesized to emerge before tasks A4 and A3. However, our model acquires A2 simultaneously with A5 – likely reflecting the tasks' similar demands on memory, which plays less of a role in A3 and A4.

For set relation tasks, a direct comparison with human data is only possible in some cases, as B3-B8 were not empirically validated. Regarding the tasks that were tested with children (B1, B2, B9, B10), accuracies progress as expected: "More" tasks (B1, B7, B9) are learned before "less" tasks (B2, B8, B10) [Res73; RWK73], and two-set comparisons (B1, B2, B7, B8) are learned before three-set comparisons (B9, B10). In fact, three-set comparisons were excluded from analysis by M. C. Wang et al. because too few subjects mastered them. However, unlike children who first acquire nonsymbolic comparisons, the model begins by learning to select between a digit and a set. We discuss this in more detail towards the end of the section. B5 and B6 are learned last, reflecting the higher demands on the model: it needs to compare a set and multiple digits and keep track of tagged digits, making the task more challenging than just navigating to a single panel and pointing.

For tasks involving numerals, training trajectories only partially align with those found by M. C. Wang et al. [WRB71] and M. C. Wang [Wan73]. Digit identification (C3) does precede digit comparison (C5, C6), which precedes seriation (C7). However, matching digits (C1), which was mastered by human subjects before any other numeral task, is acquired last by the model. The reason may be that, in our setup, C1 is the only task of its kind and involves longer and more complex navigational sequences. Learning to state (C2) and select digits (C3) is also switched compared to children, likely because outputting a number word simply means activating a single node for our model. In contrast, speech production in humans involves more complex articulatory coordination.

Seriation and ordinal position tasks were also not empirically validated by the authors of the original curriculum. However, Jeske investigated prerequisite skills in children tasked with ordering plastic strips of different lengths and found that selection of the longest strip preceded correct seriation [Jes78]. In line with these findings and the hypothesized training trajectory, selecting the largest, darkest, lightest, or smallest object (D1, D2, D5, D6) is achieved first, followed by object seriation (D3, D4, D7, D8), and finally, set seriation by cardinality (D9). Naming an object's ordinal position (D10) is learned earlier than was hypothesized by Resnick. However, other studies have found ordinal concepts to precede cardinal concepts [Bra73] and seriation [Sie71] in children.

We now turn to the training trajectories across task families. The first tasks the model learns are mostly symbolic (A1, C5, C6, C3, B3, B4, C2), followed by non-symbolic two-set comparison (B1, B2, B7, B8), then enumeration and ordinal position tasks (A4, D10, A3, A2, A5, A6, C8).

Concurrently, the model learns to select single objects by a specified attribute (D1, D5, D2, D6). The tasks that develop the latest require comparing or manipulating more than two sets of objects (B9, B10, D3, D7, C1, B5, D4, D8, B6, D9, A7, C4). Training trajectories show gradual development, characteristic of neural networks, and consistent with findings from various aspects of mathematical cognition [MS99; McC+16]. Furthermore, students' development of early numerical competencies is not always linear, and their skill acquisition timelines may differ [PF12]. Similarly, a model's performance on a task will sometimes drop momentarily (see, e.g., A6), leading to a dip in average performance and an increased standard deviation.

In general, the tasks acquired faster by the model are ones with less variability across examples, shorter sequence length, fewer memory requirements, and more exposure – either because there are limited task configurations that were upsampled or because there are very similar tasks that can serve as a scaffold. These are features of most tasks involving number words and digits, which is likely why they are acquired earlier than purely nonsymbolic ones. This contradicts the order observed in children, who typically develop nonsymbolic numerical representations before symbolic ones [WRB71; MA16; Li+18].

Give-N task Having looked at training trajectories across the dataset, we now focus on a task that has received considerable attention in the numerical cognition literature: The give-N task. A prominent proposal for the developmental trajectory on this kind of task is a series of six performance levels: pre-numeral-knower, one-knower, two-knower, three-knower, four-knower, and cardinal-principle (CP) knower [Wyn92; CS06; SC08]. Pre-numeral-knowers will give random amounts in response to a give-N instruction. One-, two-, three, and four-knowers can give out one, two, three, and four objects, respectively, but fail at all other numbers. CP-knowers can solve any give-N task.

According to the knower-level theory, children learn the meanings of numbers one through three or four one after the other. However, once they uncover the cardinal principle, tasks with higher numbers are mastered simultaneously. Several studies support this view, although some have questioned whether a true semantic inductive leap underlies the transition to CP-knower [DEB12]. Others have found that early stages may be noisier than previously assumed [WCB19].

Figure 7.5 shows the training trajectory of our model on task A6 separately for each subset size. The order of acquisition goes from smallest to largest numbers. Performance on subset size one increases first and remains high. The training trajectory for subset size two shows the same concave shape but with an accuracy gap of 10-25%, which is only closed towards the end of training. Subset sizes three and four are learned relatively simultaneously, with an almost linear development slope. Training trajectories for tasks with subsets of size five and up form a group of convex-shaped curves.

Although the graph shows no instantaneous transitions, there is a point around epoch 18 during which the performance on subsets larger than two begins to rise more steeply. This behavior is somewhat in line with the knower-level stages observed in children. However, it may not necessarily reflect any realization of a fundamental underlying principle. The training trajectories are likely also shaped by sequence length and the fact that the "visuo-motor" routines needed to complete tasks with smaller subsets are implicitly contained in those with larger subsets, leading to more training exposure.

The CP trajectory has previously been modeled computationally. Instead of using a connectionist approach, where knowledge is encoded in a set of weights, S. T. Piantadosi et al. proposed a model based on Bayesian program induction [PTG12]. The model learned to

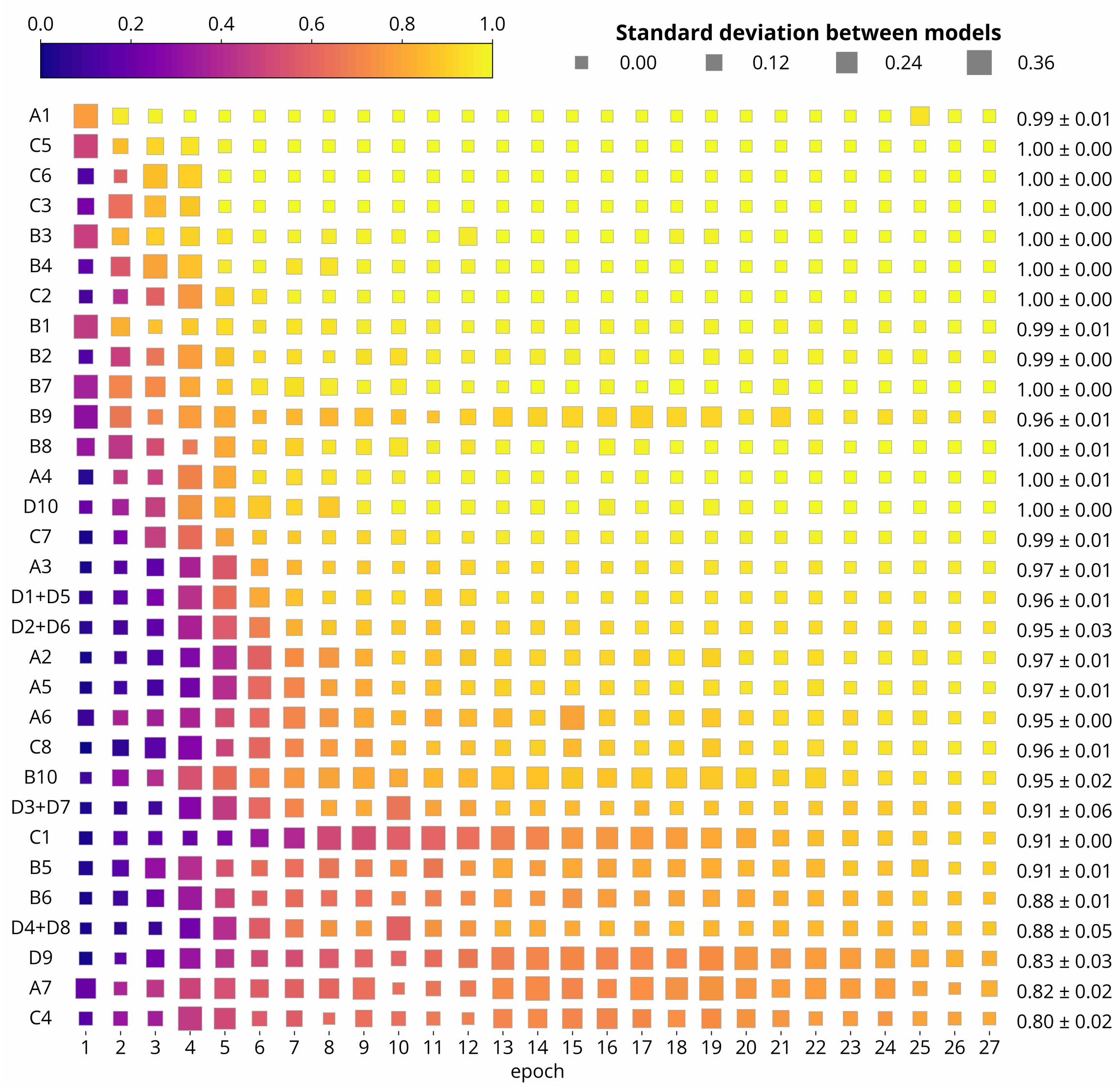


Figure 7.4: Accuracy development across tasks in the course of training. Color encodes performance, while size encodes the standard deviation between the three (architecturally identical) models. Task IDs listed on the left and final model accuracy listed on the right.

combine pre-defined operations, a so-called language of thought, to count occurrences in a set. Its training trajectory mimicked the proposed CP leap. The model by Sabathiel et al. also successfully learned a give-N task, although its learning curves did not follow the CP trajectory [SMS20b]. Dulberg et al. trained an RL agent consisting of specialized pre-trained modules to select *N* items from a binary vector using a curriculum approach [DWC21]. Like in our model, the agent showed a gradual progression, characteristic of neural networks, but did exhibit an inflection during training.

Recognizing exact numerosity Having inspected the model's training trajectories, we now turn to analyzing the trained model on a "behavioral" level, i.e., investigating its output predictions and error patterns. Two core numerical systems are often distinguished in the literature on numerical cognition [FDS04b]: The Object Tracking System and the Approximate Numerical System. The former is said to sustain the fast and precise enumeration of sets

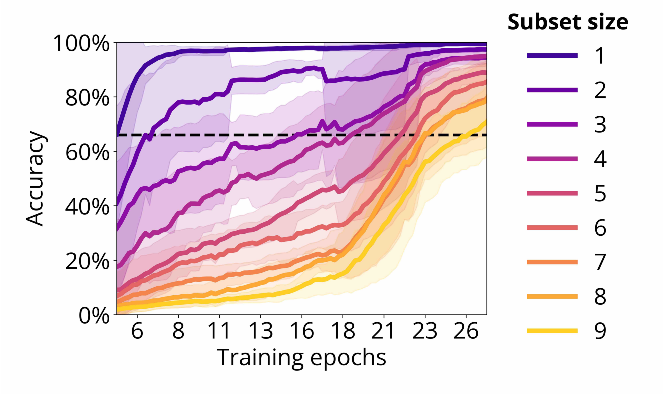


Figure 7.5: Accuracy development for the give-N task, grouped by subset size (smoothed). Shaded regions indicate standard deviation. Dashed line represents the threshold at which a learner is typically considered an N-knower.

with up to five objects without counting, an ability referred to as subitizing. The latter is hypothesized to underlie intuitive estimation and approximation of larger sets. This dichotomy has received support from many investigations of humans and non-human animals [MS82; Rev+08; HS09; BTA10; Agr+12]. However, it has been challenged by some who suggest that a single system is responsible for both subitizing and counting [Pia+02]. Whatever the underlying mechanisms, it has been widely shown that processing smaller numerosities is more precise than processing larger ones.

To see whether this is also the case in our model, we let it interact with 1,000 instances of a task environment requiring it to select a set of a given size (A7). We generate an equal number of tasks for each prompt and plot the target set size against the size of the set chosen by the model in Figure 7.6. The model shows decreasing accuracy and broader response variability with increasing target numerosity, in line with human experimental data. However, performance increases again for larger numerosities. Creatore et al. trained a Deep Belief Network on an enumeration task and observed a similar effect [CSS21]. They noted that this was an artefact of the limited range of numerosities used, which is also the likely explanation in our case.

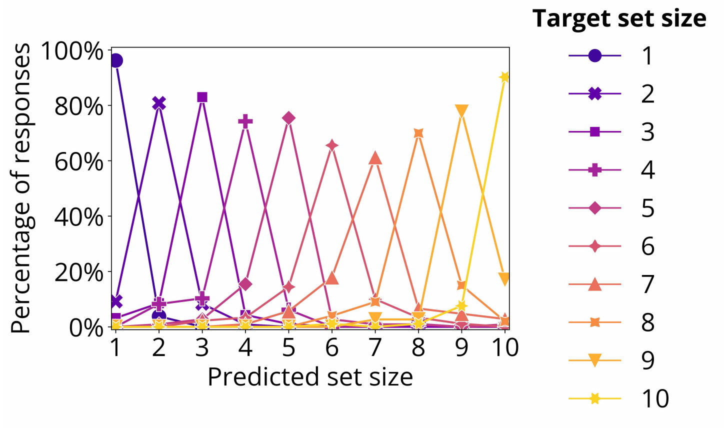
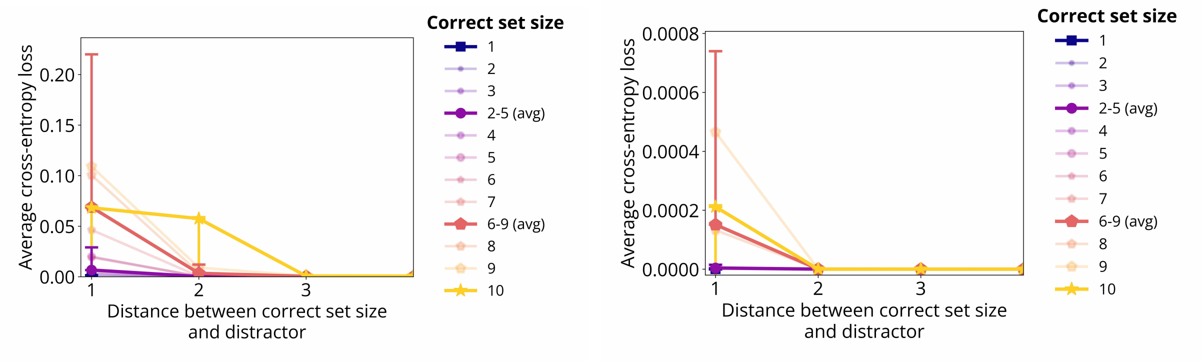


Figure 7.6: Target set size plotted against the size of the set chosen by the model on 1,000 instances of the A7 task environment (choosing a set of stated size).



(a) Non-symbolic set comparisons (B1, B2)

(b) Symbolic digit comparisons (C5, C6)

Figure 7.7: Model loss on symbolic and nonsymbolic comparisons, averaged over time steps within a task. Error bars indicate standard deviation. Similar graphs aggregated by average for better visibility (bold, opaque).

Size and distance effects In infants, adult humans, and a variety of animal species, numerosity comparisons are characterized by size and distance effects: comparisons are faster and more accurate when there is a larger difference between two numbers (distance effect) and when numbers are smaller (size effect) [DDC98]. I.e., comparing 1 vs. 9 is less error-prone than 1 vs. 2, and 1 vs. 3 is easier than 7 vs. 9. A prominent explanation for this phenomenon is that numbers are stored on a "mental number line", where close-by numbers overlap, and their noise is proportional to their value [VF04]. In humans, size and distance effects hold for symbolic and nonsymbolic stimuli [LA15], although they are minute for judgments on number symbols [BG74].

We analyze whether our model displays symbolic and nonsymbolic size and distance effects by evaluating its performance on two-set (B1, B2) and two-digit comparison tasks (C5, C6). Since the model performs very well on these tasks, accuracy is not a meaningful metric to compare. Instead, we use the model's cross-entropy loss on the test data, averaged over time steps within a task. We plot this against the distance between the correct number and its distractor, shown in Figure 7.7. In both the symbolic (Figure 7.7b) and nonsymbolic (Figure 7.7a) cases, target size one has the lowest error and almost no variation, followed by target sizes two to five. Errors and variations increase for target sizes six to nine, particularly in nonsymbolic tasks. Similar to task A7 (section 7.1.5), performance increases for target size ten – again, likely an artefact of the limited range of numbers used. In line with human behavioral studies, the error range for nonsymbolic comparisons is higher than for symbolic comparisons.

Applying the logit lens As mentioned in section 7.1.4, the CLS token contains the model's prediction. Each attention head can contribute to CLS, gradually refining the prediction until it is translated to an action by the model's output layer. However, it is possible to directly read out the prediction's state in any intermediate attention layer. This approach has been dubbed the "logit lens" and shown to provide relatively coherent internal prediction trajectories for LLMs such as GPT-2 [nos20]. Although our model, unlike GPT-2, is not purely text-based, it shares the same architecture. It thus lends itself to applying the logit lens.

We evaluate our model on each test task, decode the nascent prediction in CLS at every attention layer, and log its accuracy. The result is shown in Figure 7.8. We also include the logit lens for the model trained without symbolic tasks, denoted as N-S. How early or late a task reaches high accuracy can be seen as a measure of difficulty – analogous to reaction time in humans: Some tasks require more processing steps, i.e., attention layers, to arrive at a solution. Alternatively, the model may resort to higher attention layers because information about past inputs is only provided after the second attention block (see Figure 7.3).

Outputs of attention layers in the first attention block indicate that they prime the model for the type of answer called for by a prompt. E.g., when asked for an ordinal position (D10) or a digit's name (C3), the initial prediction is a default number such as 5 or 2. For tasks requiring recognition of a final state, the default output is "stop," while for those calling for selecting a panel or object, the default is to point. Any correct predictions in these first attention layers are by chance, e.g., when the agent starts off positioned correctly, then points. The second attention block shows a decrease in correct default answers, suggesting that inhibitory mechanisms set in at this stage.

The order of prediction trajectories mostly fits with the sequence of acquisition found in section 7.1.5. The fastest tasks to reach high accuracies are comparisons (B1, B2, B3, B4, B9, B10, C5, C6) and pure digit tasks (C2, C3). In contrast, tasks involving comparing or manipulating multiple sets, objects, digits, or knowledge of past time steps require more processing steps. This is generally congruent with event-related potential (ERP) studies showing that comparison is associated with modulations of an early component while spatial mappings are associated with later ERP components [TH18]. Less congruently, digit-set comparisons (B3, B4) are among the first tasks to reach high accuracy, whereas studies show high switching costs when humans are asked to compare symbolic and nonsymbolic numbers [LAB12; Fin+21].

The N-S model requires more processing steps than the S model, even when the final accuracy on a task is similar, indicating differences in internal processing. Notably, the accuracy progression in the N-S model is gradual for comparisons of two or three sets (B1, B2, B9, B10). This contrasts with the S model's prediction trajectories on these tasks, which show sudden performance increases between attention layers. However, on the ordinal position (D10) and row comparison (B7 and B8) tasks, the S model's accuracies increase more steadily. This linear progression is particularly striking for D10, which also has the benefit of involving only a symbolic output, making intermediate predictions more human-interpretable. We, therefore, use this task to investigate the processing underlying such gradual prediction trajectories in the following section.

Determining ordinal position It has yet to be understood how humans and animals process non-verbal serial order information. However, behavioral and neural data suggest an imprecise representation of discrete numerical rank, similar to an analog magnitude mechanism proposed for cardinality [Nie05]. Studies in humans and macaques have identified brain areas similarly activated by numerical quantity and rank order information, suggesting a shared system for these processes [Mar+00; NFM02; NMT03; NMT04]. Determining an object's position in a sequence also seems to involve a mixture of cardinal and ordinal number usage in our model.

Figure 7.9 shows two D10 example tasks and how the model's prediction changes after each attention layer of the third attention block. Predictions take the form of probabilistic distributions centered on one or more outputs. These distributions gradually move along the number line. Note that this happens "silently": the model is not trained to output numbers at each time step, only to produce the final answer. The strategy it develops to do so is evocative of an internal counting procedure. However, it does not necessarily go through the count list individually. In Figure 7.9b, it starts directly at the end of the first row with "5", from where

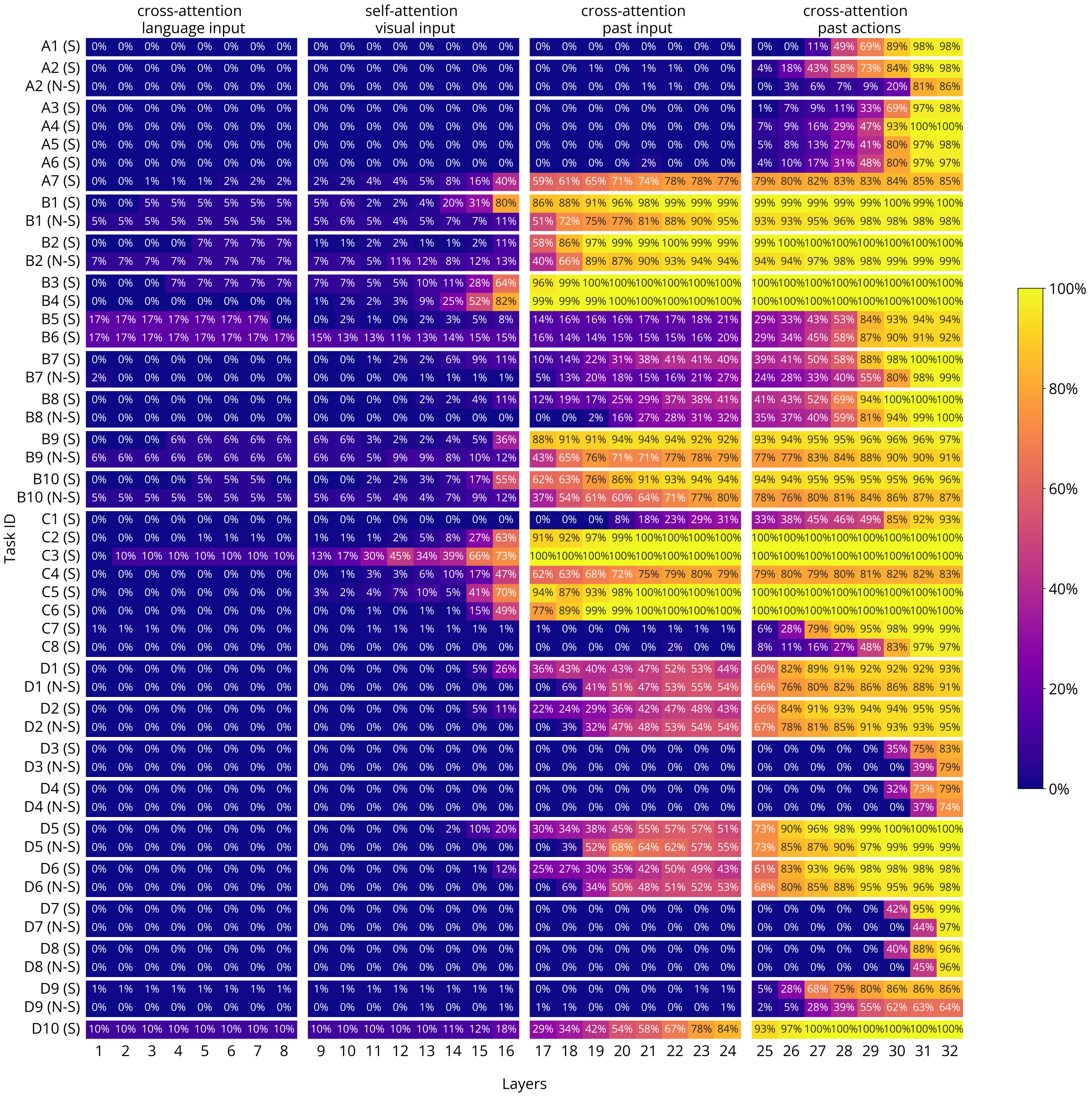


Figure 7.8: Performance of the model on each of the test tasks when decoding predictions at each attention layer, specified on the x-axis. "S" in the y-axis labels denotes the model trained on all tasks. "N-S" denotes the model trained only on non-symbolic tasks. Accuracy encoded via color.

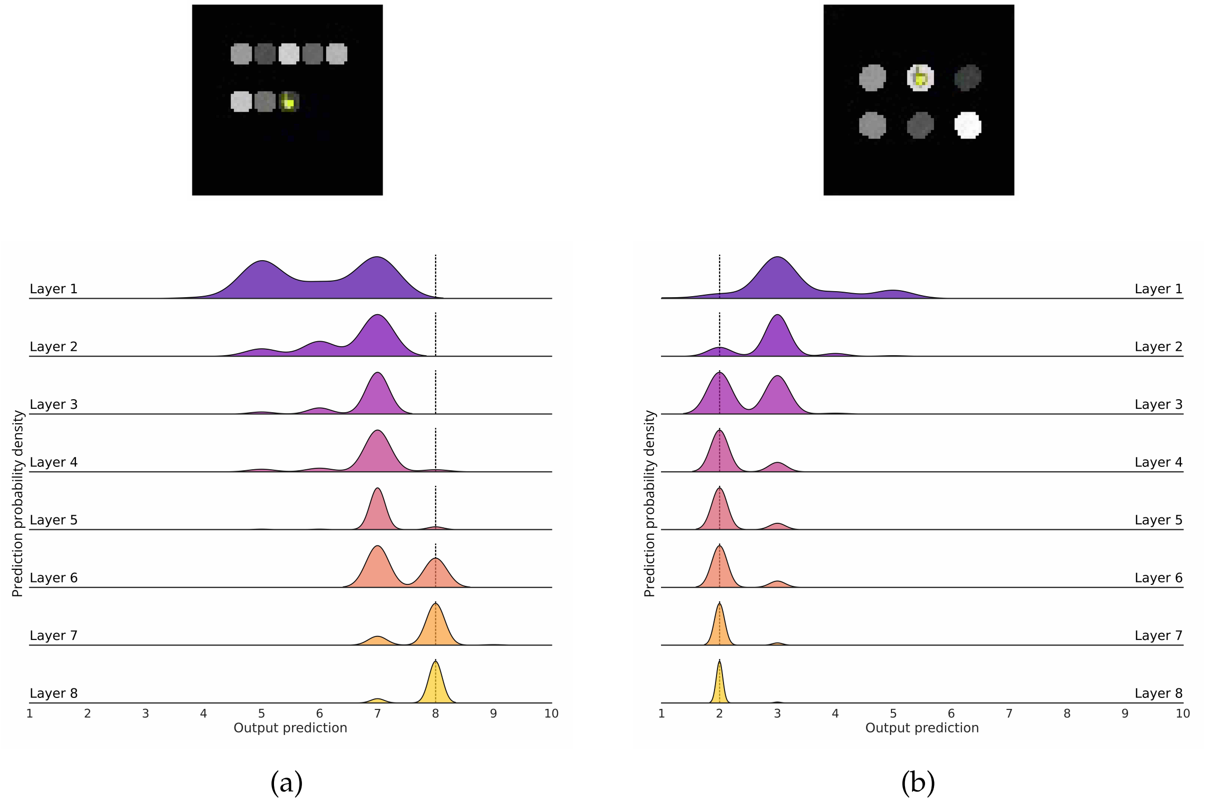


Figure 7.9: Examples of the progression of an individual prediction on task D10 throughout the attention layers of the model's third attention block. Visual task input shown at the top. The agent has to name the ordinal position of the object to which the yellow hand is pointing. The x-axis shows the ten number word outputs. The y-axis shows the probability distribution over these outputs, as predicted by the model, in each attention layer. A dashed line marks the correct output.

it moves up towards "8" (the correct answer), essentially skipping over "6". This behavior is similar to adaptive grouping strategies people employ when enumerating larger groups of objects or solving number line estimation tasks [NFG87; Cam03; SM14; Sch+18]. The model may also start at the end of a row, following the number line in reverse order (see Figure 7.9a).

These examples indicate that the gradual internal progression found for D10 in Figure 7.8 stems from the model internally tagging one object (or group of objects) per processing step until it has identified the requested ordinal position. To provide additional support for this assumption, we plot the accuracy on the D10 test set throughout the third and fourth attention blocks as a function of the target object's distance from the nearest row start or end. The result is shown in Figure 7.10. The model internally reaches its conclusion faster on tasks with target objects closer to a row's edge and with target objects in the first row – consistent with the hypothesis that tasks requiring less "internal counting" involve fewer processing steps.

Integrating multiple modalities In the previous sections, we looked at the model's outputs and error patterns in the context of human behavioral data. We now turn our attention to its internal representations. Many neuroimaging and behavioral studies have investigated where and how the human brain processes numerical inputs. One prominent proposal, the triple-code model [Deh92], argues for three codes with which we mentally represent numbers: symbolic digits, verbal number words, and nonsymbolic quantity representations. The codes are thought

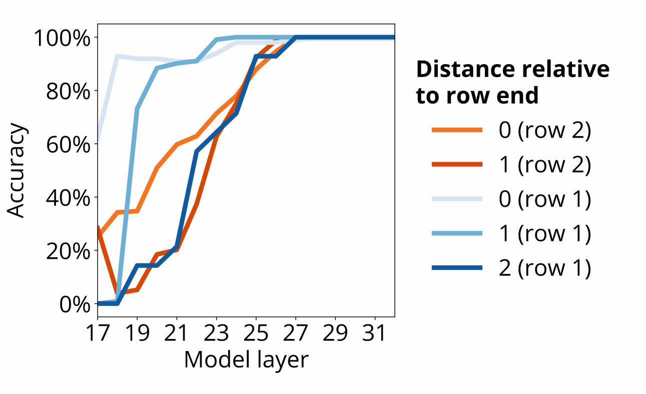


Figure 7.10: Accuracy on the D10 test set throughout the third and fourth attention blocks as a function of the target object's distance from the nearest row start or end. Tasks with target objects in the first and second row plotted separately.

to depend on distinct neural substrates, with visual inputs such as Arabic numerals most likely depending on ventral occipitotemporal structures, verbal representations depending on left frontal and temporal language areas, and analog magnitudes depending on the parietal cortex [Hub+05]. However, functional MRI (fMRI) studies have shown that numerical tasks, even those involving only one representational format, activate a distributed network of areas, including the frontal and parietal lobes [Hub+05].

In this section, we seek to analyze how our model processes and integrates information from different modalities and whether a similar picture of specialized and integrative areas emerges. We begin by creating isolated probes of input stimuli from different modalities. We then feed these isolated inputs to the key, query, and value networks of every attention head in the model and measure how strongly they react to each probe.

Our visual probes consist of 1,051 representative patches, including digits, different luminances, the agent's hand in its three states, shapes of varying resolutions, and panels with object sets of sizes one to ten. We apply the visual embedding layer (Figure 7.3 Ⓑ) to each probe but do not add size or position information. Instead, we create separate size probes, spaced evenly from 4×4 to 64×64, and position probes, spanning 65 locations across the input grid. The language probes consist of 107 vectors representing every word in the vocabulary, encoded by the language embedding layer (Figure 7.3 Ⓒ) and concatenated with each position at which a word may appear in the task prompts. Probes for previous actions consist of all 24 possible outputs encoded by the action embedding layer (Figure 7.3 Ⓖ) and 100 isolated temporal position embeddings.

Finally, we create probes that measure sensitivity to the state of the `CLS` token. We translate all possible actions E_{Pred} back into internal model representations by applying the output layer W_{Pred} (Figure 7.3 Ⓘ) "in reverse". Specifically, we subtract W_{Pred}'s bias term b_{Pred} from E_{Pred} and apply the pseudo-inverse $W^{\dagger}_{\text{Pred}}$:

$$W^{\dagger}_{\text{Pred}}(E_{\text{Pred}} - b_{\text{Pred}})^T$$

For each network in each attention head, we record the ten probes that evoke the largest response, quantified as the sum of the network activations' absolute values. In Figure 7.11, we show which modality these inputs belong to and the strength of the response they elicited. The first two attention blocks integrate language and visual information. Query networks of

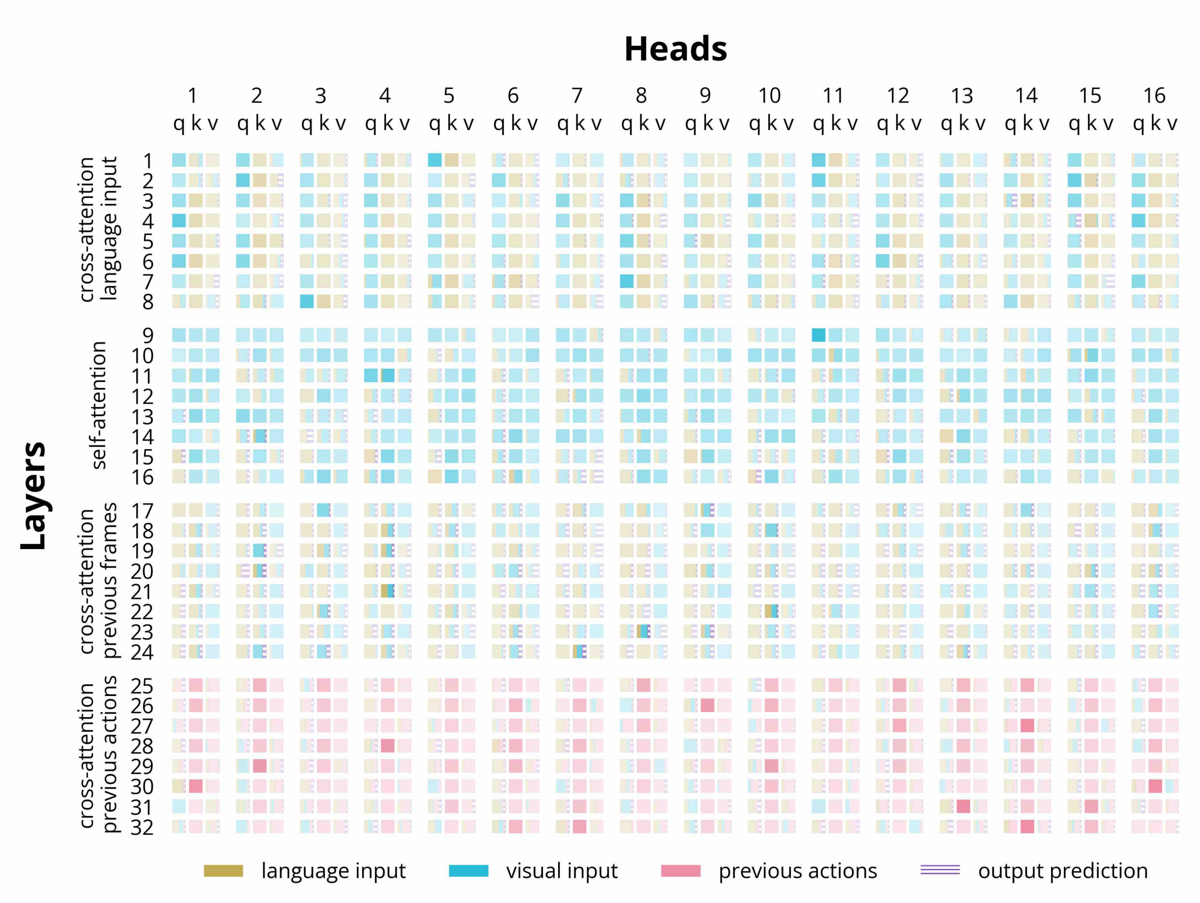


Figure 7.11: Visualization of the sensitivity of every query, key, and value networks in the model to isolated probes from different input modalities. Opacity indicates the strength of activation exhibited by a network in response to the input probes.

heads in the very first attention layer receive visual input. Key and value networks receive language input. As might be expected, query networks in the first block mainly respond to image patches, and key networks mainly respond to words. The value networks react to a mix of language, visual, and output predictions (`CLS`). The partial sensitivity to nascent predictions fits our observations from section 7.1.5 that the model forms “default” outputs at this level. Heads in the second attention block primarily integrate visual information, although some exhibit sensitivity to words.

In the last two blocks, information from past time steps enters the picture. The third block is interesting because the key and value networks in its first attention layer receive `MEM` vectors, i.e., the time step representations produced by the model 7.3 Ⓕ). Figure 7.11 shows that these compressed representations seem to contain a mix of visual, linguistic, and output prediction information. We also see more sensitivity to output predictions, which matches our finding from section 7.1.5 that many tasks are already solved at this stage. In the last block, query networks process a mix of linguistic, visual, and output prediction input, while key and value networks are predominantly sensitive to previous actions.

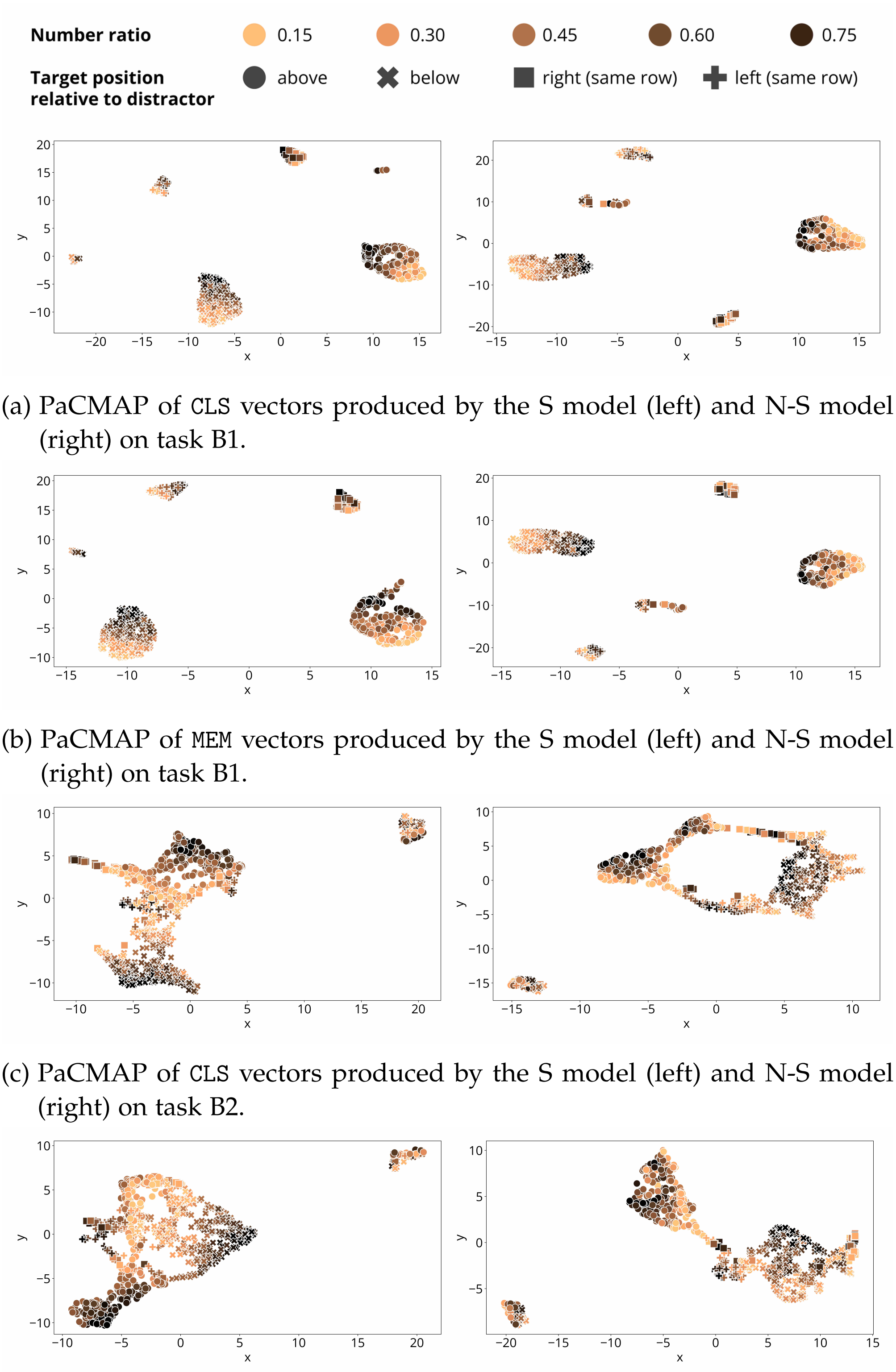
Overall, Figure 7.11 paints a picture of a distributed network of specialized processing units integrating multimodal information. There are very few unimodal heads – primarily in the second attention block. Most heads consist of a unimodal query network interacting with key and value networks sensitive to different modalities. In a few heads, particularly in higher attention layers, single key, query, or value networks respond to inputs from multiple modalities.

The effect of symbolic training Having observed in section 7.1.5 that set comparison tasks require more processing steps in the N-S model than in the S model, we here investigate this finding further. We run both models on the test data for tasks B1 and B2 and collect the inputs to the third attention block, as our analysis in section 7.1.5 showed this to be the point where two-set comparison predictions begin to form. We collect only the time step where the agent is positioned at the correct set but has not yet selected it to facilitate cross-task comparison. We visualize the collected `CLS` and `MEM` vectors using Pairwise Controlled Manifold Approximation (PaCMAP) [Wan+21]. PaCMAP is a dimensionality reduction technique designed to preserve the data's local and global structure. Figure 7.12 gives insight into the differences between the internal representations of the S and N-S models and the role of `MEM` vectors.

We begin with the `CLS` vectors, which encode the model's predictions. For task B1, these form distinct clusters according to the position of the target set relative to its distractor (Figure 7.12a). Within the clusters, tasks with similar number ratios, calculated as the smaller set size divided by the larger set size, are grouped closer together. However, for the N-S model, this stratification is slightly less pronounced. There is also a collection of "miscellaneous" predictions that are not yet well clustered, indicating that further processing steps are needed. `CLS` vectors for task B2 (Figure 7.12c) are less neatly grouped than for B1, which fits with the observation from section 7.1.5 that "less" comparisons are solved in higher attention layers than "more" comparisons. The PaCMAP for the N-S case is almost circular, reflecting that many `CLS` tokens have few neighbors of high similarity. The arrangement indicates that, at this stage, the vectors still contain perceptual details that have already been abstracted away in the S model.

We now turn to the `MEM` vectors (Figures 7.12b and 7.12d), which contain compressed information the model deemed relevant enough to "remember" about a time step. The `MEM` PaCMAPs closely resemble the `CLS` PaCMAPs in their differences between tasks B1 and B2 and S and N-S models, as well as their stratification according to number ratio and target position. This suggests that `MEM` and `CLS` contain similar information. To test this hypothesis, we evaluate the models on tasks B1 and B2 as before but replace the `MEM` vectors with `CLS` vectors after the second attention block. We see no decrease in performance, confirming that the two are interchangeable, at least for set comparison. For other tasks, such as A5, doing this does cause a significant accuracy drop from around 98% to 13%, showing that `MEM` vectors carry crucial additional or complementary information in some cases.

We can conclude that set relations are implicitly quite well defined by attention layer 16, although slightly less so for the N-S model. To quantify this gap further, we train two linear regression models to predict the size of a task's larger and smaller set based on the models' B1 `CLS` and `MEM` vectors. We do this for each attention layer in the second attention block. For the N-S model, the coefficient of determination goes from an average of 83% in the first to 93% in the eighth attention layer. For the S model, it goes from 83% to 97%, suggesting that it produces slightly more precise representations of set cardinality earlier, on which its higher levels can operate. In the N-S model, which appears to require more processing steps, i.e., attention layers, to determine set size, fewer attention layers are available for higher-level operations once cardinality information has been determined. This leads to a lower performance on tasks like set seriation (D9).



(a) PaCMAP of `CLS` vectors produced by the S model (left) and N-S model (right) on task B1.

(b) PaCMAP of `MEM` vectors produced by the S model (left) and N-S model (right) on task B1.

(c) PaCMAP of `CLS` vectors produced by the S model (left) and N-S model (right) on task B2.

(d) PaCMAP of `MEM` vectors produced by the S model (left) and N-S model (right) on task B2.

Figure 7.12: Pairwise Controlled Manifold Approximation (PaCMAP) applied to the `CLS` and `MEM` representations produced by the models trained with both symbolic and nonsymbolic tasks (S) and on nonsymbolic tasks only (NS), collected after the second attention block during the processing of two-set comparison tasks (B1, B2). Proximity of points indicates similarity.

Visualizing an information flow The analyses presented so far have mostly looked at static model weights for one or more entire task families. We now want to provide a glimpse into the dynamics that unfold while processing a single task. We use an information flow graph in the style of Katz et al., who recently proposed this kind of visualization for LLMs [KB23]. We adapt their tool to our multimodal case to show a snapshot of the information flow in the model's 11th attention layer during one time step of the first B4 task in the test set (Figure 7.13). We choose task B4 as it is relatively simple but involves the comparison of a set size and a digit – in this case, a set of size ten and the numeral four (Figure 7.13 Ⓐ). This makes it an interesting case for investigating the two number formats' representations. We choose the 11th attention layer because it is the first point in which the correct action enters the model's top five most likely predictions, indicating that the attention layer's heads play a role in solving the task. Nodes represent groups of activated neurons. Edges represent interactions, with width encoding interaction strength.

The attention layer receives the model's current state as input. This state is a high-dimensional vector that is not human-interpretable. However, we can translate it to an action prediction by directly applying the model's final output layer (Figure 7.3 Ⓘ). We show the five most likely outputs as separate bars (Figure 7.13 Ⓑ). Length indicates certainty. The correct answer is to point because the agent is in the right panel. This action is not yet among the top outputs. The prediction undergoes normalization (Figure 7.13 Ⓒ), which has been found to act as a "semantic filter" in LLMs by dampening the effect of common inputs and boosting the signal of rare tokens [KB23]. In our case, normalization does little except increase the likelihood of the "stop" action.

What follows are the outputs of the key, query, and value networks in the attention layer's 16 attention heads (Figure 7.13 Ⓓ-Ⓕ), each represented by a node. As we saw in Figure 7.11, the networks may encode linguistic or visual information. To "decode" their outputs, we compare their activations with those they exhibited in response to the probes in section 7.1.5 and use the closest match as node labels. Labels are colored according to modality. We also translate each network output to an action prediction, as we did for the attention layer input (Figure 7.13 Ⓑ). The color of each node represents whether this translation yields the correct action (pointing) as the most likely (green) or second-most likely (yellow) output. This color-coding indicates whether a network contributes to the correct prediction.

The results of the interactions between keys, queries, and values pass through the heads' output layers (Figure 7.13 Ⓖ). The individual heads' outputs are aggregated into an updated prediction (Figure 7.13 Ⓗ). This updated prediction is added to the attention layer's original input and processed by further normalization and a feedforward block, which we do not depict for simplicity. The attention layer shown in Figure 7.13 is a relatively early one, and the updated prediction it produces is still almost uniform. However, the correct action, pointing, has now entered the model's top five predictions due to the contributions from the attention layer's heads.

If we consider the mechanism formed by keys, queries, values, and outputs as an associative process, we see that the model retrieves relevant information, including representations learned from other tasks. E.g., there are activations for the visual digit ten, a pointing hand, and the number words "ten" or "four" – none of which are in the task's immediate input. Most heads output the action "point", which modifies the model's top five predictions to include pointing. However, the output predictions "three", "four", or "five" also appear across the attention head. This activation of surrounding number outputs can be explained by looking at the weights in the model's final output layer.

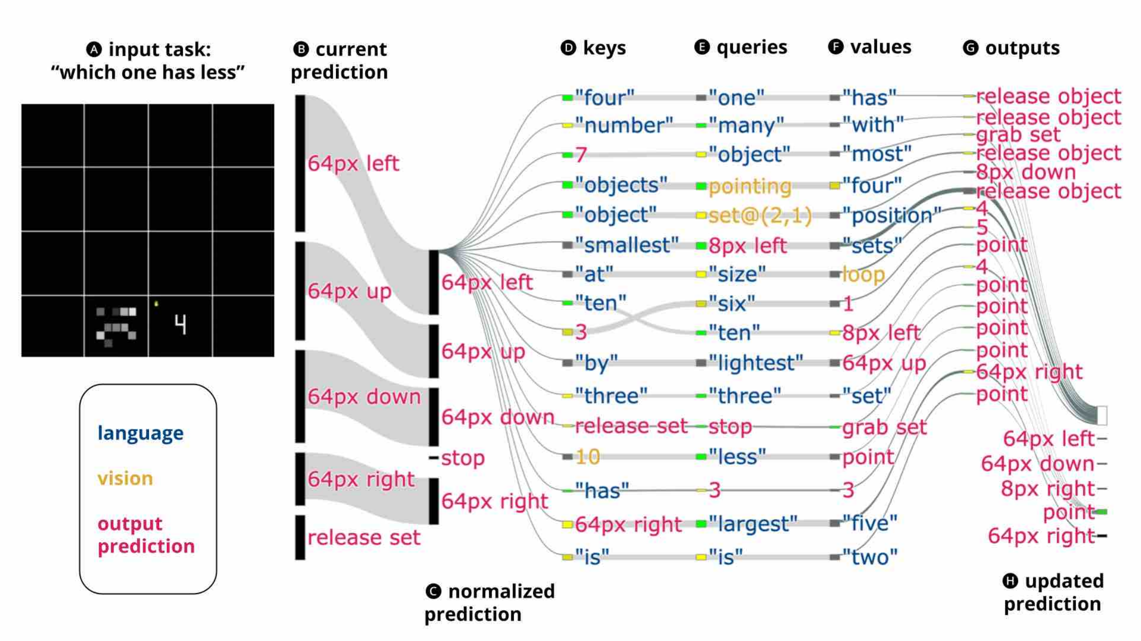


Figure 7.13: A graph of the information flow in the style of Katz et al. [KB23] of the third attention layer of the first attention block while processing one time step of a B4 task. Nodes represent groups of activated neurons. Edges represent interactions, with width indicating interaction strength. Nodes are labeled with the most likely prediction when processed by the model's final output layer, except for keys, queries, and values, where we use the probe from section 7.1.5 eliciting the most similar activation. Node color represents whether activations, when interpreted as predictions, have the correct action (pointing) as either the most likely (green) or second-most likely (yellow) output. The graph should be read from left to right and omits the attention layer's feedforward block for simplicity.

Figure 7.14 shows the cosine similarity of the incoming weights for each possible output. Similarity for weights of neighboring numbers is higher than for numbers further away, leading to a co-activation of close-by numbers in line with Verguts et al.'s proposal of a noisy mental number line [VF04]. The fact that various visual, spatial, and output prediction nodes appear in the graph also fits well with Abrahamse et al.'s proposal that performing a task co-activates perceptual, motor, and goal representations in the brain, binding them into a context-specific network which allows for cognitive control [Abr+16].

Comparing task processing sequences Our dataset spans a range of number concepts and task families, which enables us to compare them from various perspectives. So far, we have looked at the order of acquisition during training in section 7.1.5 and within-model prediction trajectories in section 7.1.5. Finally, we want to compare the model's internal activations while processing different tasks. We run the model on all tasks in the test set and collect each attention layer's 64-dimensional attention head outputs at every time step. We average the recorded activations over the time steps of a single task and sum over the 1,000 tasks in a task family. We take the pairwise cosine similarity for the aggregated activation vectors of each task family as a measure of similarity between their activation trajectories. In Figure 7.16, we present the results in a hierarchically-clustered heatmap.

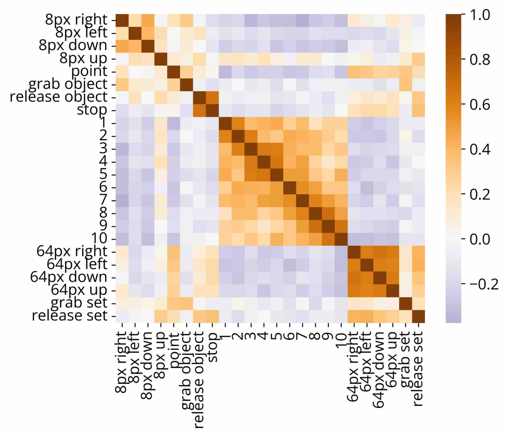


Figure 7.14: Cosine similarity of each output's incoming weights in the model's final output layer.

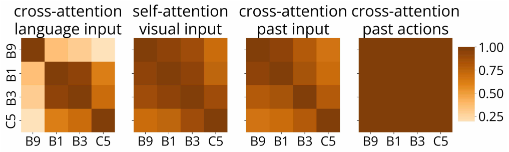


Figure 7.15: Cosine similarity between aggregated activation trajectories for the four tasks in the dataset involving the "more" relation, in each of the four attention blocks.

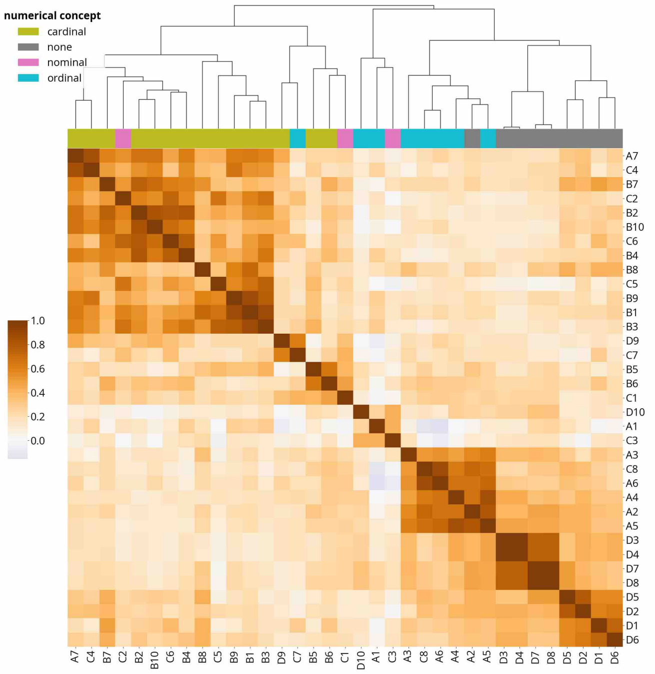


Figure 7.16: Hierarchically-clustered heatmap of the cosine similarity between aggregated activation trajectories for each task in the dataset.

Two over-arching clusters form – one cluster of mainly cardinal tasks that involve set comparisons or exact cardinality (upper left) and one of within-panel seriation and ordinal tasks (lower right). Within the second cluster, there is a subcluster of set enumeration tasks (A3, C8, A6, A4, A2, A5), object seriation tasks (D7, D8, D3, D4), and object selection tasks (D5, D2, D1, D6). Notably, although sorting objects differing in one (D3, D7) and more than one (D4, D8) attribute showed different training trajectories, the activation trajectories in the trained model are almost identical. The map also includes a small cluster of tasks with purely verbal outputs (D10, A1, C3) and one cluster of tasks requiring the manipulation or selection of multiple panels (D9, C7, B5, B6, C1).

Tasks that involve different number modalities but are otherwise identical show high similarity. Examples include A7 and C4, C8 and A6, or D9 and C7. This suggests that the network has learned knowledge and procedures employed similarly in tasks involving different number formats. We investigate where the model processing diverges when solving versions of the same task with different number representations in Figure 7.15. The plot compares four tasks involving the "more" relation (B9, B1, B3, and C5) broken down by attention block. Activation trajectories diverge most in the lower attention layers, then form clusters according to input representation formats: B9 and B1 involve only object sets, and C5 involves only digits. B3, which involves objects and digits, shows equal similarity to both. Activations in the fourth block are almost identical, most likely because its attention layers do not contribute much to these tasks and are essentially skipped during processing (see section 7.1.5).

Several neuroimaging studies have done comparable analyses to investigate activations in the brain during tasks involving different number representations and magnitudes. Results indicate that neural overlap depends on task demands [LA15] and that, besides areas thought to represent numbers, numerical tasks activate more non-specific brain areas related to, e.g., general visuospatial skills [Hub+05]. These findings generally fit with the fact that clusters in activation trajectories in Figures 7.15 and 7.16 in part reflect number representation format and in part similarities in other visual inputs and action sequences.

7.2 Discussion

In summary, the model's training trajectories within and across tasks mostly fit with empirical findings from children where such findings are available. In line with human behavioral data, the model shows decreasing accuracy and broader response variability with increasing target numerosity, as well as nonsymbolic and symbolic size and distance effects. Qualitative analysis of the model suggests an intricately entwined network of specialized and more general processing units rather than a strictly hierarchical and segregated set of modules. Using isolated probes, I show where in the model information is integrated via multimodal attention heads. I explore the interplay between attention heads in action by visualizing an exemplary information flow. The visualization illustrates how attention heads retrieve cross-modal information related to, but not necessarily present in, the model's immediate input. I compare aggregated activations across tasks and find that overlap in activation trajectories reflects similarities in inputs and task demands.

Inspired by discussions in the literature on the role of language in numerical cognition, I train a model only on non-symbolic tasks. The model performs well on two-set comparisons and tasks related to object attributes, in line with findings that some proto-quantitative skills can develop without language. However, it performs less well on tasks involving more than two sets. I compare the internal processing and embeddings of the models trained with and without symbolic tasks. I conclude that the model trained without symbolic tasks requires more processing steps to determine set sizes, leaving fewer capacities for more advanced operations involving multiple sets. This offers a concrete, computationally implemented demonstration of how differences in exposure to symbolic number tasks can give rise to differences in internal representations and processing strategies.

In the context of this thesis, this study represents the culmination of several ideas we have already encountered in the previous chapters. It especially resembles the BIB case in that I train a large NN on a set of cognitively relevant tasks, analyze what it learns, and compare the model's behavior and internal representations to those of humans where possible. This relates to three topics I briefly want to outline in the following sections: Aristotelian vs. Galilean psychology, epistemic opacity, and exploratory modeling in science.

7.2.1 Aristotelian and Galilean psychology

The distinction between Aristotelian and Galilean psychology refers to two different approaches to psychological research and theory, as characterized by Kurt [Kur31]. These approaches diverge in their methodologies, perspectives on psychological phenomena, and the emphasis on understanding or categorizing behaviors and mental processes. Note that, despite its name, Galilean psychology is not directly related to the notion of Galilean idealization discussed in section 4.2.2.

Aristotelian psychology is defined by its reliance on categories derived from everyday observations, such as "normal" vs. "pathological." This approach is top-down, meaning it starts with broad taxonomies into which it fits encountered phenomena. Explanation takes the form of subsuming observations under broader categories or laws without necessarily investigating the underlying mechanisms. The focus is on consistency and replicability. Variability is often dismissed as noise or as something outside the scope of scientific inquiry. Researchers using this strategy prioritize group averages, which are thought to capture the essence of the forces governing behaviors and mental processes [HC17].

Much of cognitive neuroscience can be said to follow this paradigm, as observations are often considered "explained" when they have been attributed to the activation of a particular part of the brain [Hom20]. According to Hommel, the Aristotelian approach may be unavoidable in the early days of a field but is ultimately "limited to re-describing the available findings in a modeling language [Hom20, p. 1297]".

In contrast, Galilean psychology seeks a more nuanced, mechanistic understanding of psychological phenomena. It moves away from binary categorization and instead explains behaviors and mental processes in terms of gradations and common principles. This approach does not rely on everyday categories but begins with basic, well-understood mechanisms and tries to use them to explain a broad spectrum of observations. The emphasis is on identifying the components and processes that underlie behavior and cognition. Variability, both inter- and intra-individual, is seen as important data that a good mechanistic theory should account for rather than as noise to be ignored [HC17]. Hommel has argued that achieving progress in the field of cognitive science requires moving towards a more Galilean psychology [Hom20].

But do NNs lend themselves to developing the kinds of mechanistic theories that this approach calls for? We have seen in the glyph study how a NN can provide how-possibly explanations of a target phenomenon without presupposing a deep understanding of the model itself. However, when the target phenomenon in question concerns how the mind performs certain computations, it becomes necessary to "look under the hood" of the model – a challenge in the case of NNs, as they can exhibit very complex dynamics.

7.2.2 Epistemic opacity and explainable AI

Humphreys defines a system as epistemically opaque relative to a cognitive agent X at a time t if the agent does not know all the Epistemically Relevant Elementss (EREs) of the system at that time [Hum09]. EREs are the components, processes, or steps within a computational system that must be understood to fully grasp how the system functions and produces outcomes. These elements can include algorithms, data transformations, and the underlying logic that drives the system's behavior. Crucially, epistemic opacity is not an inherent property of a system but is relative to an agent's knowledge and capabilities [Zed21]. A system might be opaque to one person but transparent to someone with different expertise or resources.

Humphreys also discusses the concept of "essential" epistemic opacity, which occurs when it is practically impossible for an agent to know all the EREs of a system – not because of a lack of effort, intelligence, or technology, but due to the agent's very nature [Hum09]. I.e., there may be cases where our fundamental cognitive limitations prevent us or any agent with similar constraints from ever completely understanding a system [Alv]. According to some accounts, NNs are instances of such inherently and unavoidably opaque systems; the sheer volume and complexity of the operations involved make it impossible for us to grasp every detail at once, and trying to explain them is futile [Bur16; Rud19; Alv].

However, not everyone shares this view. Indeed, there is a whole field within ML, Explainable Artificial Intelligence (XAI), which aims to demystify the workings of NNs. This drive toward developing interpretability methods is fueled by a mix of practical, ethical, and epistemological reasons.

One of the primary motivations for XAI is to build user trust and encourage the adoption of ML models, particularly in high-stakes areas such as healthcare and finance [Vel20; Kri20]. In these domains, the reliability of AI systems is crucial, as their decisions can significantly impact individuals' lives. Moreover, the ability to explain AI decisions is not just a technical necessity but represents a moral obligation to ensure fairness and accountability in AI-driven outcomes [Dig20]. This is especially pertinent as legal frameworks in various jurisdictions increasingly mandate transparency in AI-based decision-making processes. From the perspective of developers and researchers, better insights into models' inner workings may enable the identification and correction of errors or biases, thereby improving performance and fairness [ZZ18; WF21; Pra+22]. Additionally, shedding light on which features significantly influence predictions could provide a deeper domain-specific understanding of the processes and phenomena being modeled [RL18; WF21].

Yet, while most scholars agree that it would be desirable for ML models to be interpretable, there is very little consensus about what interpretability means and how to achieve it [DK17; Lip18; Wat21]. To name a few prominent proposals, Doshi-Velez and Kim define interpretability as the ability to explain an ML system or to present it in understandable terms to a human [DK17]. Similarly, Lipton discusses notions of transparency, suggesting that a model might be considered transparent if a person can contemplate the entire model at once or if each part of the model admits an intuitive explanation [Lip18]. However, such definitions often run into a recursive problem by relying on terms like "understandable" or "intuitive," which themselves require further clarification [Sul22]. Others, such as Ratti and López-Rubio, attempt to define interpretability using more pragmatic criteria, like the ability to perform precise and successful material manipulations based on the information provided by the model [RL18]. Despite these and numerous other attempts to define and quantify XAI terms, the field still lacks clear, agreed-upon definitions for interpretability and its cognates [Kri20].

Krishnan puts forward that XAI researchers struggle to pin down these concepts because "there is not, in fact, anything that it is to *be* an interpretation or an explanation of an algorithm. There are only [. . .] lists of facts that are more or less useful for particular purposes" [Kri20, p. 491]. To connect this back to Humphreys' idea of epistemic opacity, interpretability, explainability, or transparency are not inherent system properties. Instead, they are specific to the model, its EREs, and the epistemic agent. Following this line of thought, it is certainly possible to try to make specific models for specific situations easier to grasp for a particular user group. However, searching for blanket definitions or universal metrics might not be a coherent or achievable goal.

In this way, NNs have something in common with brains: They are highly complex systems that we can probe in ad hoc ways to answer specific questions. However, we do not have an all-encompassing grasp of their behavior and underlying mechanisms. In situations like these, where comprehensive theoretical frameworks are still missing, scientists often employ exploratory modeling techniques.

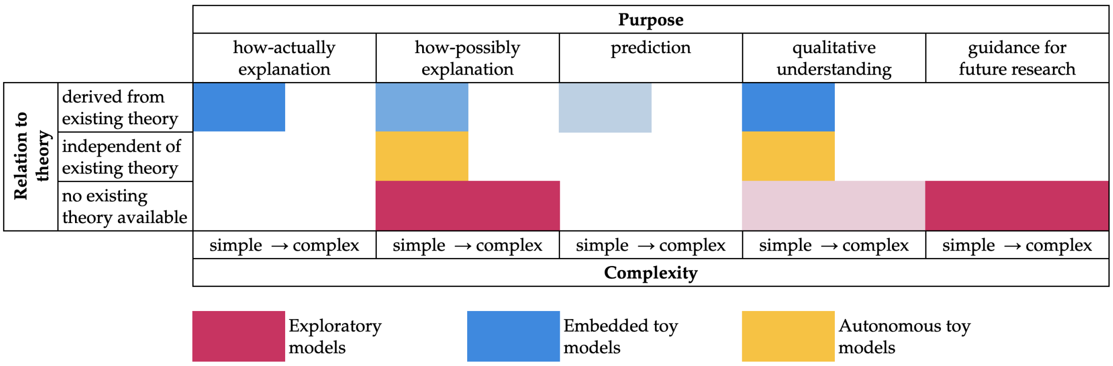


Figure 7.17: Comparison of toy models and exploratory models.

7.2.3 Exploratory modeling

The "standard view" of science traditionally emphasizes the role of experiments as a way to test hypotheses and theories. According to this view, experiments are primarily designed to confirm or falsify specific predictions. This perspective strongly emphasizes the deductive-nomological notion of explanation, where experiments are seen as a means to validate the causal mechanisms posited by a theory. However, this "standard view" does not fully capture the diversity of experimental practices in science. Besides hypothesis testing, experimentation can also have important exploratory uses, including generating hypotheses in the first place. Exploratory experimentation and modeling are particularly relevant in contexts where a well-formed body of theoretical knowledge is not yet available, or the subject matter does not lend itself to a straightforward theoretical description [Gel16b]. This is arguably the case in many areas of the cognitive sciences [McC09; JK17; Zer+19].

In this way, exploratory models have much in common with toy models, especially autonomous ones. Both types of models focus on the modal dimension of modeling and can be used to generate how-possibly explanations. However, toy models and exploratory models are not entirely synonymous. Toy models can be independent of or embedded in theory, while exploratory models are used when no established theory exists. Toy models are highly simplified and idealized to the point of stylization. Exploratory models, on the other hand, can vary in complexity and detail.

Finally, the primary aim of toy models is to provide qualitative understanding, while exploratory models aim to guide future research. They are not necessarily intended to provide definitive answers but to open up new lines of inquiry and help formulate more precise questions. In that way, exploratory models can lay the groundwork for more focused investigations in the form of toy models, which can, in turn, help refine or expand the scope of initial explorations [Gel19]. Figure 7.17 summarizes the similarities and differences between toy and exploratory models.

Gelfert distinguishes between four functions of exploratory models [Gel16b]. The first use is as a starting point for future inquiry. These models allow scientists to identify key variables and their possible interactions and outline the scope of the research. By doing so, they can highlight areas that require further investigation and lead to new hypotheses that can be tested empirically. The second use is as a proof-of-principle demonstration. In this context, a model is constructed to show that some mechanism or process could, in principle, feasibly lead to the observed phenomenon. The third use is to generate potential explanations for observed phenomena. These models allow scientists to identify which accounts are more consistent with the available evidence and help narrow down the range of possible explanations. The fourth use is to assess the suitability of the target phenomenon for investigation. These models can help re-evaluate the boundaries and characteristics of a phenomenon, leading to a more precise and operational definition.

Despite the ability of exploratory modeling to offer important insights, it comes with some caveats that require careful consideration. Models, by necessity, simplify the complex phenomena they aim to represent. This simplification is a double-edged sword; it makes models tractable and accessible but can lead to misrepresenting or overlooking critical aspects [McC09]. It is, therefore, crucial not to conflate a model with the reality it seeks to explain – to distinguish the map from the territory, as it were. Understanding a model can be an important step in understanding a target system, but they are not the same thing [Gel16b; Sul22].

Furthermore, it is essential not to conflate how-possibly with how-actually explanations. The inferences drawn from exploratory models are often under-constrained, and multiple models may account for the same data [Fri15]. Therefore, a model's ability to fit existing data does not necessarily validate it as a true representation of the underlying processes [McC09]. Conversely, a model's failure to fit data could be due to seemingly minor implementational or conceptual choices [YD16]. It does not automatically invalidate the core ideas of a model. Misinterpreting model success or failure can lead to premature conclusions about the validity of the hypotheses being explored [McC09].

Finally, as previously discussed, there is rarely such a thing as a "perfect and absolute blank" in modeling – our observations are significantly shaped by the theories and concepts we already hold. Much like idealization, this, too, has benefits and drawbacks. As discussed in the third case study, the background knowledge and assumptions built into our models are what link them to scientific theory and the world, even if they are not accurate representations. They can help guide exploration in what would otherwise be an infinitely large space of possibilities [JK17].

However, the fact that all modeling is "theory-laden" means that every step, including data collection, pre-processing, data interpretation, labeling, and algorithm design, is infused with some degree of - often implicit - bias [Kit14; Leo14; Bur16; Zer+19; Des+22]. While there is no way to avoid this altogether, it is important to acknowledge how existing theories and expectations shape the series of decisions involved in modeling because these affect the model's behavior and the knowledge we can derive from it.

In DL models, so many parameters are optimized algorithmically, outside our direct control as designers, that this can impart a false sense of objectivity or "blankness". In reality, however, the learning process is influenced by a series of decisions based on implicit knowledge. For example, I chose the architecture and size of my models with the expectation that this would elucidate aspects I was interested in studying. As I became more familiar with the models, I refined my choices to navigate the various trade-offs involved between computational resources and predictive performance. For some decisions, such as the proportions of data used for training, validation, and testing, I relied on best practices in the field of ML. Other hyper-parameters I chose based on my experience with similar projects.

In that regard, I am no different than other cognition researchers working with complex models. However, I attempted to generalize the problems across input modalities as well as task types in such a way that it reduced my ability to make choices that were highly targeted to or optimized for narrow tasks and datasets. This was an effort to increase the applicability and generalizability of the models, even if it came at the cost of optimizing for specific benchmarks.

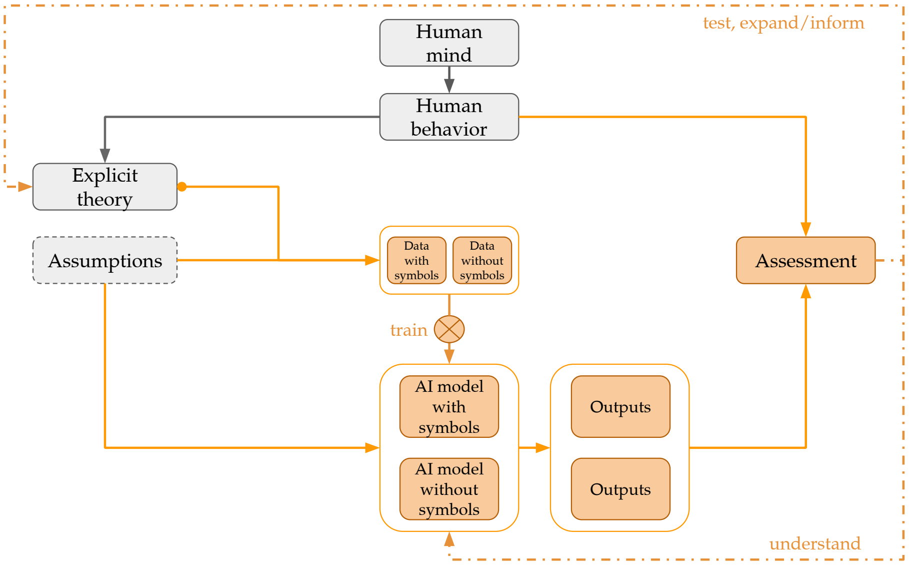


Figure 7.18: Overview of relations between human cognition, theory, assumptions, data, model, and outputs in the fifth case study. Relevant components performed or generated by us shown in orange. Components provided by third parties shown in gray. Circle at the beginning of an arrow indicates the starting point of investigation.

7.2.4 Relation to the guiding questions

The modeling process involved in this study resembles several of the previously presented cases (see Figure 7.18). It is similar to the glyph study in that my starting point lies in the cognitive science literature, which informs the design of my dataset. I train large NNs on this dataset and analyze their behaviors and internal representations. This top-down approach contrasts with many of the previous models in numerical cognition, which are often simpler and involve components with manually defined semantics. In this regard, the case study is similar to the BIB case, where the NN also presents an alternative to a model based on a different paradigm (in the case of BIB, HBToM). However, I analyze the NN's inner workings in more depth. Thus, the goal is not just to prompt new questions and inquiries, as with the BIB study. Instead, I aim to better understand the model – analogous to the gSCAN case – and potentially to inform the theory that inspired the study – analogous to the glyph case.

This feedback loop between cognitive science literature and model assessment leads us back to the first criticism against NNs, namely, their lack of connection to theory. As discussed in section 5.2.2, most models are linked to our knowledge of the world through the questions they are designed to investigate. In this study, the literature on numerical cognition heavily informed the dataset design, model analysis, and interpretation of results. However, as has perhaps become apparent in the study, this is not a field where many definitive, widely accepted theories exist. There are still more open questions than answers, and competing proposals and models abound. Thus, numerical cognition is an area where exploratory modeling can be helpful. In section 7.2.3, I outlined the four functions of exploratory models according to Gelfert. Two of these functions are especially relevant to this work: exploratory models as proof-of-principle demonstration and as a starting point for future inquiry.

Regarding the first function, this study reinforces and amplifies previous findings that early number skills can emerge from the general learning mechanisms of NNs. E.g., the model's "implicit curriculum" forms without imposing an order of task presentation or explicitly modeling maturational changes, which have been hypothesized to underlie transitions in children's learning [MK19]. It produces size and distance effects without an innate, spatially organized "mental number line", a prevalent explanation for this phenomenon in humans [ZPU02; Har+13]. In general, it develops a network of specialized and more general processing units. This functional organization emerges from objective-based training without enforcing topological constraints on model connections.

Of course, these findings do not preclude the presence of certain neural structures supporting number skills in the brain. However, they demonstrate that innate circuitry is not the only possible source of explanation. The model can thus help inform and expand our views on numerical cognition, moving towards a more nuanced Galilean approach to psychology.

Given that the model reaches high accuracy on most tasks, including comparisons requiring extrapolation to larger set sizes, it could also serve as a starting point for further *in silico* exploration of hypotheses about the biological mind. Discrepancies between model and human behavior are particularly interesting in this regard because they provide clues about factors at play in human learning that may be missing in the setup [McC09; CK19].

For instance, the proposed model learns symbolic tasks faster than non-symbolic ones. I attributed this to symbolic tasks involving less variability and, often, shorter sequence length. A more human-like acquisition order may arise with a more realistic dataset where digits vary in appearance, and outputting number words requires producing individual phonemes. Alternatively, symbolic tasks may be introduced later in training, or changes to the architecture may be needed. Furthermore, future work could investigate the role of maturational changes in learning by gradually increasing model capacity and comparing internal representations or processing strategies to those emerging from a priori full-scale models. The model could also be ablated to simulate hypotheses about developmental disorders.

Turning to the second criticism, biological implausibility, it is certainly fair to say that the NN used in this study does not accurately represent the brain. It is much more computationally limited, its neuronal dynamics differ, and training and inference are split into separate phases that do not reflect organic learning processes [SK20; LKC20]. However, the modeling setup represents a step forward from previous proposals regarding cognitive plausibility.

Many of the studies outlined in section 7.1.2 used comparatively small models, abstracted inputs, and few specific tasks, as this was conducive to their goal of understanding model representations and processing. In DL, models are trained on naturalistic data and evermore general tasks. However, the focus is generally on performing well on benchmark datasets rather than analyzing the models' inner workings. While there is undoubtedly room and good reason for both approaches, I have tried to find a middle ground: I use a large, relatively general-purpose NN. I train it on the circumscribed domain of early number knowledge, then analyze it in depth. Many of my analyses reveal representations and processing strategies that could only emerge from a sufficiently complex, i.e., more cognitively plausible setup.

Employing NNs in smaller, controlled environments that capture essential properties of natural experience and focusing more on the "how" than the "how well" can benefit both cognitive science and AI. Cognitive scientists can use AI developments to broaden their models' scope, allowing them to analyze phenomena that cannot emerge when studying isolated concepts. For AI researchers, better insights into NNs can yield a more realistic assessment of model capabilities and motivate improvements in architectures or input data. For example, analyses of the proposed model pointed to the limitations of its purely feedforward nature, underscoring the importance of recent efforts to introduce recurrent weight sharing and adaptive halting mechanisms to Transformer-based architectures [Mes+22; CT22].

As seen throughout the study, design decisions at every level significantly impact what a model can be taught. Even two models with identical architectures and similar task performance may develop diverging internal processing mechanisms if trained on different inputs. Fields such as cognitive science and developmental psychology have long studied the experiences that shape what and how we learn. This expertise can inform the design of ecologically relevant training inputs that induce more human-like representations and processing in NNs, ultimately making them more cognitively plausible.

We now come to the last criticism against NNs, namely, their supposed inability to provide explanations. Scaling up to more cognitively plausible tasks inherently introduces a higher level of complexity, both in the inputs and in the increasingly powerful models needed to process them. As discussed in section 7.2.2, some consider these models instances of essentially epistemically opaque systems – i.e., systems that we will never fully grasp due to our cognitive limitations as humans. While it may be true that an all-encompassing understanding of these model's inner workings is out of reach, the preceding discussion illustrates that NNs are not the entirely impenetrable black boxes they are often made out to be. Specifically, I hope I have shown that the ubiquitous Transformer models can be made more interpretable through the application of appropriate analysis techniques.

Because NNs allow for almost unfettered access, they can be subjected to various analyses. Researchers can causally manipulate individual elements, continuously collect data at different levels of detail, and selectively lesion or stimulate models in a way that is currently impossible with biological brains. When guided by specific research questions, it is thus possible to identify EREs of a modeling setup and to provide explanations at different levels of the Marr hierarchy. These can range from comparing the effect of exchanging a higher-level building block, such as the training dataset, to producing lower-level mechanistic accounts, such as the one presented in the section 7.1.5.

To sum up, in many areas of research, comprehensive theoretical frameworks are not yet available. In such cases, NNs can serve as exploratory models. Much like the captain's blank map in the chapter's epigraph, open-ended inquiries involving models that contain fewer rigid assumptions can have their benefits. Unlike purely bottom-up, descriptive models, they have the potential to produce surprising results that diverge from the expected behavior. Such results can provide proof-of-principle demonstrations or a starting point for future inquiry, including the design of more targeted empirical studies with human participants.

While such models may fall short of how-actually explanations, they can help us expand our views on a topic and perhaps support a shift towards a more nuanced Galilean approach to psychology. Given the ability of NNs to process complex inputs, they can help cognitive scientists expand the scope of phenomena they are able to model. Conversely, AI developers can take inspiration from the cognitive sciences to design ecologically relevant training environments and potentially move towards more cognitvely plausible models. Finally, NNs are amenable to a much wider range of post-hoc analysis techniques than biological brains.

In section 7.2.3, I have briefly presented the four uses of exploratory models proposed by Axel Gelfert: starting points for future inquiry, proof-of-principle demonstrations, generating potential explanations, and assessing the suitability of a target phenomenon. This case study and the three studies preceding it have mostly served the first three uses of exploration. In the following chapter, I will present a model that illustrates the fourth exploratory function.

8 Modeling decision-making processes in medical ethics with Fuzzy Cognitive Maps

> *It seems to me that the test of "Do we or do we not understand a particular subject in physics?" is, "Can we make a mechanical model of it?*
>
> William Thomson

8.1 Study

The sixth case study relates to explicit knowledge and reasoning (see Figure 8.1). As this is quite a broad topic that could fill several theses in its own right, the case study focuses on a specific example: the reasoning involved in decision-making in medical ethics. Although machine intelligence applications are becoming increasingly prevalent in healthcare, medical ethics remains largely unexplored from a technical perspective. We propose an approach based on Fuzzy Cognitive Maps (FCMs), which builds on Beauchamp and Childress' prima-facie principles. The FCM's weights are optimized using an EA to provide recommendations regarding the initiation, continuation, or withdrawal of medical treatment. The final model approximates the answers provided by our team of medical ethicists reasonably well and offers a high degree of interpretability.

The model was developed as part of an interdisciplinary collaboration between the Chair of Data Processing and the Institute of History and Ethics in Medicine at the Technical University of Munich. We first presented our approach in an article in the American Journal of Bioethics in 2022 [Mei+22a]. The paper elicited several responses from the medical ethics community [SHK22; Cha22a; DES22; CD22; BP22; Rah+22; Cha22b; GB22b; KG22; BFG22; Sab22; DFR22; PB22], to which we replied in a separate article [Mei+22b]. We also produced a companion paper geared towards a more technical audience, which was published in the proceedings of the 2022 International Conference on Fuzzy Systems [Hei+22]. The following text is based mainly on this technical companion paper but also contains fragments of the first two publications.

8.1.1 Introduction

Some ethical questions that arise in clinical settings have obvious answers. Others are more complicated. Should doctors continue treating a child who still has a small chance of long-term survival against her will? Should one carry out a procedure that has adverse medical effects when the patient insists on being treated this way? Should medical personnel put a person on a ventilator following a suicide attempt when she had signed a do-not-resuscitate order many years ago? These are just some examples of medical ethical dilemmas in which our moral intuitions appear to give conflicting advice.

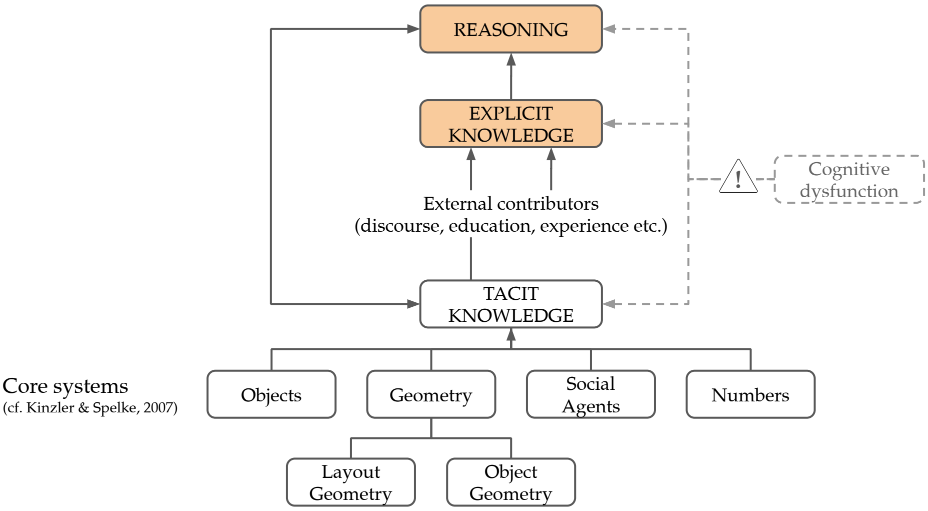


Figure 8.1: Situating the sixth case study in the broader study of cognition. Relevant parts of the framework marked in orange.

The need for ethical counseling triggered the appointment of clinical ethics committees. In the early 1980s, only 1% of US hospitals employed ethics committees, and on average, they reviewed not more than a single case per year [You+83]. Nowadays, ethics committees are in operation in most major hospitals [16]. Not only is the number of clinical ethics committees increasing, but so is the number of cases that are brought before these institutions [SMP12]. Among other reasons, this is due to advancements in healthcare technology that enable doctors to carry out procedures that raise entirely novel ethical questions [Per92; FM01].

In recent years, ML technology has become more and more prevalent in medicine. ML-based systems are now being used to support physicians in areas like disease diagnostics, medical image analysis, care coordination, or precision medicine [Hab+21]. Yet, despite the growing number of clinical ethics cases brought before committees, the realm of ML approaches to *moral* decisions remains largely unexplored. To our knowledge, the only attempts at such a system were GenEth [AA18] and its predecessor version, MedEthEx [AAA06]. Both are ethical dilemma analyzers proposed by Anderson et al., which use Inductive Logic Programming (ILP) to codify decision principles for medical ethics dilemmas.

Although these studies delivered interesting first results, they applied to a restricted set of cases, namely, ones in which mentally competent patients refuse to undergo treatments that medical professionals judged as being beneficial for them [AAA06]. An updated version also included a scenario in which patients are reminded to take their medication [AA18]. In addition to the studies' limited scope, ILP is known to struggle with noisy inputs and non-symbolic domains where data is inexact or uncertain [EG18]. In clinical reality, however, medical staff face a wide variety of situations, and there is seldom an "objectively correct" solution. A system designed to provide ethical guidance in medicine can, therefore, be expected to receive contradictory inputs from different experts across cases. Since translating such conflicting instances into purely symbolic representations using classic ILP would be difficult, we here propose a different approach based on FCMs.

We argue that fuzzy technology is especially well-equipped to handle such an ambiguous domain and that FCMs are an intuitive tool for modeling medical-ethics decision processes. We present an FCM that covers a variety of cases regarding the initiation, continuation, and withdrawal of medical treatment. We further show how such a model can be optimized to approximate expert recommendations with relatively little data using an EA. We begin by providing some background on medical ethics, FCMs, and EAs. We then introduce our FCM model, the data set used to optimize it, the pre-processing steps we applied, and the details of our implementation. We report on the quantitative evaluation of our experiments as well as on qualitative results. Finally, we discuss the strengths and limitations of our approach and point out ways in which our work could be used and expanded upon.

8.1.2 Background

Medical ethics So far, applications of machine ethics have mainly been focused on situations like trolley dilemmas in autonomous driving (see, e.g., Awad et al. (2018). A major challenge for such undertakings is to a) choose an appropriate ethical theory and b) translate often vague ethical concepts into computer-friendly representations. Clinical ethics is in a uniquely advantageous position in this regard. This is because there exists a set of four prima-facie principles, introduced in 1979 by Beauchamp and Childress, which today is the dominant methodology for doing bioethics and is regarded by many as the de facto standard in the Western world [Vea20]. Beauchamp and Childress' approach facilitates our choice of ethical framework, both because of the consensus surrounding it and because it lends itself well to algorithmic implementations. The four principles are [BC13]:

Beneficence: medical interventions should promote patients' well-being. The treatment option that will offer the greatest benefit to the patient should be selected. A benefit would be an increase in life expectancy, an improvement in the quality of life, or both.

Non-maleficence: medical interventions should not harm the patient. This is often in conflict with the principle of beneficence since there are very few treatment options that are completely free of risks and side effects.

Patient autonomy: medical interventions should only be carried out in accordance with the patient's preferences.

Justice: the distribution of costs and benefits within the healthcare system should be adequate and fair; in contrast to the other three principles, justice does not only pertain to the individual patient but governs the allocation of medical resources throughout the whole community.

Fuzzy Cognitive Maps FCMs were first proposed by Kosko in 1986 [Kos86] and have since been adopted across a variety of domains, such as social and political sciences, robotics, medicine, or environmental studies [PS13]. They are a graph-based way of modeling a set of concepts $\mathcal{C} = \{C_1, C_2, C_3, \dots, C_M\}$, represented as nodes, and the cause-and-effect relationships between them, represented as weighted directed edges $W : \mathcal{C} \times \mathcal{C} \rightarrow [-1, 1]$. Each node takes on a fuzzy value, which represents how active a concept is at a given point in time. Each edge is assigned a fuzzy weight, where positive values represent a causal increase, and negative values represent a causal decrease. The absolute weight value determines the magnitude of the causal effect one concept has on another.

To simulate a chain of causal reasoning for the modeled domain, the values of $\mathcal{C}$ for each time step are calculated as the weighted sum of their input concepts, passed through a non-linear function. Equation 8.1 displays the most widely used activation rule for FCMs, where $A_i^{(t)}$ is the activation value of concept C_i at the time step t and w_{ji} denotes the value of the edge between C_i and C_j. A is applied iteratively until a stop condition is met and produces a state vector with updated activation values for all concepts at each time step. f denotes a transfer function that ensures the activation values of all concepts stay within the desired range, usually a sigmoid or hyperbolic tangent function [Fel+19].

$$A_i^{(t+1)} = f\left(\sum_{\substack{j=1\\i\neq j}}^{M} w_{ji}A_j^{(t)}\right) \tag{8.1}$$

This update mechanism is similar to the one used in neural networks, and indeed FCMs can be seen as “interpretable recurrent neural networks that include fuzzy logic elements during the knowledge engineering phase”[Fel+19, p. 1710]. However, in contrast to most traditional neural networks, the strength and polarity of the weights between nodes are usually not learned via backpropagation. Instead, it can either be specified manually by domain experts or learned using Hebbian, error-driven, or hybrid approaches [Fel+19].

Evolutionary algorithms An EA is a probabilistic search method inspired by evolution. It starts with a pool of solutions, the so-called “population". At each generation, the fitness of each solution, also called a chromosome, is evaluated. Each chromosome is made up of a collection of solutions for individual parameters to be optimized (genes). Depending on their fitness value, chromosomes may be chosen for the next generation. Some solutions are randomly selected to produce new solutions through crossover and mutation operators. Since their introduction in the 1970s, EAs have been applied to optimization problems in a variety of fields, such as biology, finance, and engineering [Kum+10].

8.1.3 Methods

Fuzzy Cognitive Map Design Due to their intuitive visualizations as causal graphs, FCMs lend themselves well to communicating ideas in interdisciplinary studies such as ours. We made use of this by first consulting our team of medical ethicists regarding the parameters they considered relevant when assessing cases. Our joint goal was to find a set of concepts sufficient to cover a range of situations but not so large as to make the model overly specific and complex. We then proposed several visual representations of how these concepts might be connected and iteratively refined our FCM based on the ethicists’ feedback. The resulting graph is shown in Figure 8.2.

The largest part of the FCM relates to the concept of autonomy. When patients possess full decisional capacity, doctors should normally follow their treatment preferences. In the case of comatose or otherwise incapacitated patients, one consults the advance directive – if such a document was drafted and if it applies to the situation at hand. Doctors can also obtain consent to a treatment from the patient’s surrogate decision-maker. If the patient’s preferences cannot be established, doctors proceed according to what is deemed to be in the patient’s best interest [Bun19], essentially relying only on BENEFICENCE and NON-MALEFICENCE. These different sources of obtaining consent may also conflict – for instance, when an advance directive is only partly applicable and clashes with the surrogate’s instructions.

We model this decision process using three types of nodes: input, intermediate, and output. Input nodes represent all the facts of a medical case, e.g., whether the patient is considered an adult (the age of majority varies between countries). For input nodes, the values are

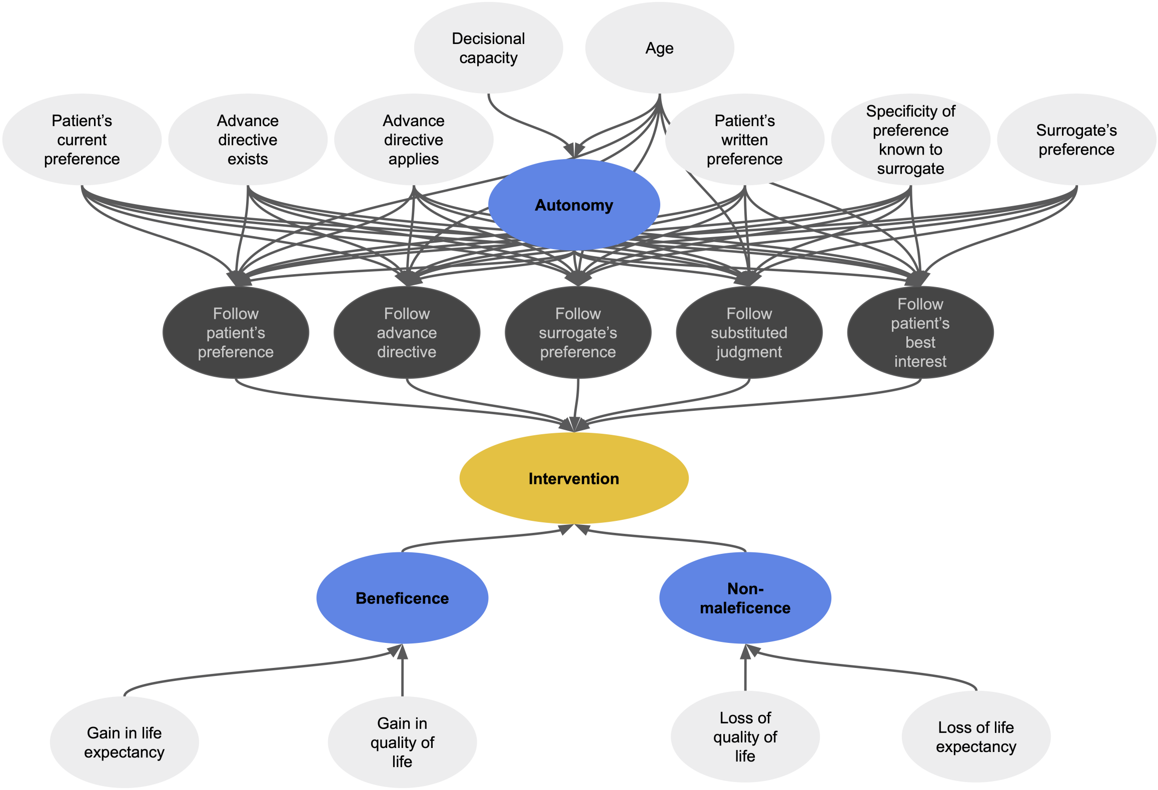

Decisional capacity
Age
Patient's current preference
Advance directive exists
Advance directive applies
Autonomy
Patient's written preference
Specificity of preference known to surrogate
Surrogate's preference
Follow patient's preference
Follow advance directive
Follow surrogate's preference
Follow substituted judgment
Follow patient's best interest
Intervention
Beneficence
Non-maleficence
Gain in life expectancy
Gain in quality of life
Loss of quality of life
Loss of life expectancy

Figure 8.2: Causal graph of the proposed model. Input nodes are shown in light gray, nodes representing Beauchamp and Childress' principles in blue, nodes representing intermediate decisions in dark gray, and the output node in yellow. Gains and losses in quality of life and life expectancy undergo pre-processing as outlined in section 8.1.3. Dark gray nodes are also fully connected amongst each other (not shown for the sake of simplicity).

given by the user, and activation stays constant. The output node represents the system's recommendation regarding the intervention in question. It takes values between 0 and 1, where 0 means strongly opposed to, and 1 means strongly in favor of an intervention. Intermediate nodes represent concepts such as Beauchamp and Childress' principles.

The concepts of BENEFICENCE and NON-MALEFICENCE receive as inputs the gains and losses in quality of life and life expectancy that the intervention in question will likely bring about, as estimated by a doctor. Autonomy is determined by the patient's possession of decisional capacity and whether they are an adult or a minor. AUTONOMY does not influence the output node directly. Instead, together with other inputs relating to, e.g., advance directives and surrogates, it influences a set of intermediate decisions, such as following the patient's wishes or their advance directive. These intermediate decision nodes, in turn, influence the final outcome. More than one intermediate decision at a time may be correct, e.g., if the surrogate's wishes are in agreement with the patient's. In total, our final model consists of 12 input nodes, 8 intermediate nodes and one output node for the recommendation regarding the intervention.

We decided to exclude cases involving the principle of justice. This is for two reasons: first, the handling of medical ethics cases in which justice is a major factor is partly dictated by regulations that vary from country to country. To implement these juridical differences computationally would have been impractical. Secondly, the question of how to define and quantify fairness is still a hotly debated topic in the field of MI. Many, often conflicting, metrics have been proposed, but as of yet, no consensus on the matter exists [VR18]. For now, we therefore focus on cases involving only the principles of AUTONOMY, BENEFICENCE, and NON-MALEFICENCE.

As the transfer function for our node activations we chose a modified sigmoid function

$$f(x) = \frac{1}{1 + e^{-r(x-b)}} \tag{8.2}$$

where r and b are parameters optimized individually for each node. The reasoning behind this is that some aspects of cases in medical ethics, such as AUTONOMY, may be almost binary, whereas others, such as BENEFICENCE, are more continuous. By making the activation function's slope and offset optimizeable parameters, the model can take this into account.

Dataset The data set used in this study consists of 69 cases sourced from the medical ethics literature [AS89; DHP10; FM01; JB16; PP10; Per92; SG08]. It includes case types such as pregnancy and abortion, consent in minors, advance directives and consent in adults, patients' refusal of treatment, requests for provision of futile treatment, withdrawal of treatment, and issues in mental health. For each case, we collected 20 input features, listed in Table 8.1. We also collected a ground truth label specifying a medical ethicist's intervention recommendation, expressed as a number from 0.0 (strongly opposed to the intervention in question) to 1.0 (strongly in favor), as well as the proposed intermediate decisions. For each case, our team of ethicists filled out a form with questions pertaining to these 20 input features.

Pre-processing Features 1 – 3, 5 – 7, and 9 – 18 in Table 8.1 are re-scaled to range [0, 1]. Features regarding treatment preferences (13, 16, 18) are re-scaled to range [-1, 1], with unknown or not applicable values located at zero:

$$x_{\text{norm}} = (b - a)\frac{x - \min(x)}{\max(x) - \min(x)} + a \tag{8.3}$$

where a and b are the lower and upper bound of the desired range, respectively.

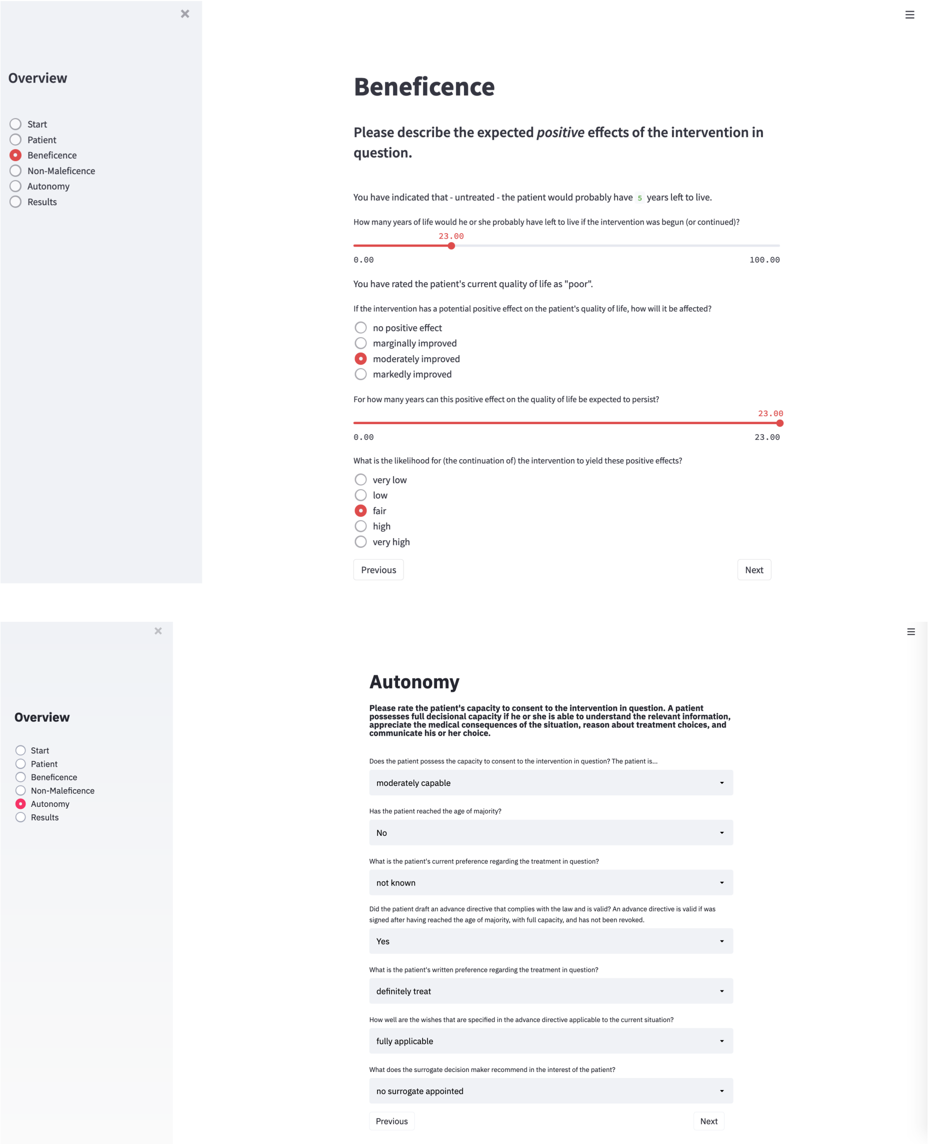
Loosely inspired by Kahneman and Tversky's prospect theory [KT13], the input values regarding the potential risks and benefits of an intervention are pre-processed to take into account the patient's risk preferences regarding potential benefits vs harms of an intervention, as well as their preferences regarding their emphasis on quality of life vs life years. To do this, we map all "objective" inputs concerning probabilities for risks and benefits, as well as quality of life, to a "subjective" value. For probabilities, we use

$$p^* = \frac{dp}{dp + (1 - p)} \tag{8.4}$$

where d is different for potential risks and benefits, and a large value of d corresponds to a high-risk/high-gain attitude. It is determined by the patient's preference (feature 20) and the empirically chosen mapping given in Table 8.2. Figure 8.4 shows the objective and subjective probabilities for a gain or loss of value 1.0, for different patient preferences.

Table 8.1: Overview of dataset features.

Category	Feature		Possible values
Health status	1	Estimated years left to live if intervention was not begun (or not continued)	[0,100]
	2	Patient's current quality of life	{very poor, poor, fair, good, very good, excellent}
Beneficence	3	Estimated years left to live if intervention was begun (or continued)	[0,100]
	4	Estimated positive effect on patient's quality of life	{none, marginal, moderate, significant}
	5	Estimated duration of positive effect on patient's quality of life	[0,100]
	6	Estimated likelihood of positive effect on patient's quality of life	{very low, low, fair, high, very high}
Non-maleficence	7	Estimated years of life lost due to (continuing) the intervention?	[0,100]
	8	Estimated negative effect on patient's quality of life	{none, marginal, moderate, significant}
	9	Estimated duration of negative effect on patient's quality of life	[0,100]
	10	Estimated likelihood of negative effect on patient's quality of life	{very low, low, fair, high, very high}
Autonomy	11	Patient's capacity to consent	{definitely incapable, marginally capable, moderately capable, definitely capable}
	12	Age of majority reached	{Yes, No}
	13	Patient's current preference	{definitely treat, rather treat, rather not treat, definitely not treat}
	14	Valid advance directive	{Yes, No}
	15	Applicability of advance directive	{fully applicable, partly applicable, marginallyapplicable, not applicable}
	16	Patient's written preference	{definitely treat, rather treat, rather nottreat, definitely not treat}
	17	Familiarity of surrogate with patient's wishes	{fully familiar, partly familiar, marginally familiar, not familiar}
	18	Recommendation of surrogate decision maker	{definitely treat, rather treat, rather nottreat, definitely not treat}
	19	Patient's emphasis on gains in life expectancy (+1) or increase in quality of life (-1)	[–1,+1]
	20	Patient's risk preference	{very high risk/very high gain, high risk/high gain, medium risk/medium gain, low risk/low gain, very low risk/very low gain}



Beneficence

Please describe the expected *positive* effects of the intervention in question.

You have indicated that - untreated - the patient would probably have 5 years left to live.

How many years of life would he or she probably have left to live if the intervention was begun (or continued)?

23.00

0.00 100.00

You have rated the patient's current quality of life as "poor".

If the intervention has a potential positive effect on the patient's quality of life, how will it be affected?

○ no positive effect
○ marginally improved
● moderately improved
○ markedly improved

For how many years can this positive effect on the quality of life be expected to persist?

23.00

0.00 23.00

What is the likelihood for (the continuation of) the intervention to yield these positive effects?

○ very low
○ low
● fair
○ high
○ very high

Autonomy

Please rate the patient's capacity to consent to the intervention in question. A patient possesses full decisional capacity if he or she is able to understand the relevant information, appreciate the medical consequences of the situation, reason about treatment choices, and communicate his or her choice.

Does the patient possess the capacity to consent to the intervention in question? The patient is...

moderately capable

Has the patient reached the age of majority?

No

What is the patient's current preference regarding the treatment in question?

not known

Did the patient draft an advance directive that complies with the law and is valid? An advance directive is valid if was signed after having reached the age of majority, with full capacity, and has not been revoked.

Yes

What is the patient's written preference regarding the treatment in question?

definitely treat

How well are the wishes that are specified in the advance directive applicable to the current situation?

fully applicable

What does the surrogate decision maker recommend in the interest of the patient?

no surrogate appointed

Figure 8.3: Example screenshots of our user interface for collecting model inputs.

Table 8.2: Mapping of patient risk profiles to values for d.

Patient preference	Value of d for gains	Value of d for losses
very high-risk/very high-gain	5	0.1
high-risk/high-gain	4	0.5
medium-risk/medium-gain	1	1
low-risk/low-gain	0.5	4
very low-risk/very low-gain	0.1	5

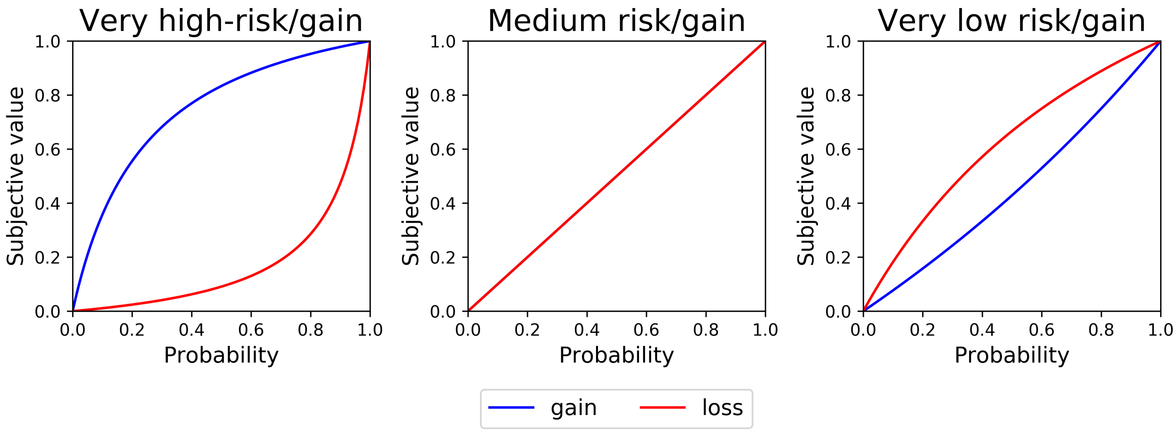


Figure 8.4: Examples of the mapping of “objective" probabilities for potential consequences of an intervention to “subjective" probabilities, according to a person’s risk preferences.

For weighting quality of life, we use

$$q^* = \frac{\alpha q}{\alpha q + (1-q)}, \qquad \text{where} \quad \alpha = \begin{cases} \frac{1}{1-2c}, & 1 \le c \le 0 \\ 1+2c, & 0 < c \le 1 \end{cases} \tag{8.5}$$

and c denotes the parameter which was input by the ethicist as feature 19.

Gains and losses in quality of life and life years are considered separately and calculated as follows:

$$\begin{aligned} q_{\text{dec}} &= q^*_{\text{base}} \cdot (1 - k_{\text{dec}}) \\ q_{\text{inc}} &= q^*_{\text{base}} \cdot (1 + k_{\text{inc}}) \end{aligned} \tag{8.6}$$

$$\begin{aligned} \Delta q^*_{\text{dec}} &= q^*_{\text{base}} - q^*_{\text{dec}} \\ \Delta q^*_{\text{inc}} &= q^*_{\text{inc}} - q^*_{\text{base}} \end{aligned} \tag{8.7}$$

$$Q_{\text{gain}} = p^*_{\text{inc}} \cdot \Delta q^*_{\text{inc}} \cdot \min(t_{\text{inc}}, t_{\text{base}}, t_I) \tag{8.8}$$

$$Q_{\text{loss}} = p^*_{\text{dec}} \cdot \Delta q^*_{\text{dec}} \cdot \min(t_{\text{dec}}, t_{\text{base}}, t_I) \tag{8.9}$$

$$\begin{aligned} T_{\text{gain}} = p^*_{\text{inc}} \cdot (\ & q^*_{\text{base}} \cdot \max(0, t_I - t_{\text{base}}) \\ & + \Delta q^*_{\text{inc}} \cdot \max(0, t_{\text{inc}} - t_{\text{base}}) \\ & - \Delta q^*_{\text{dec}} \cdot \max(0, t_{\text{dec}} - t_{\text{base}})\ ) \end{aligned} \tag{8.10}$$

$$\begin{aligned} T_{\text{loss}} = p^*_{\text{dec}} \cdot (\ & q^*_{\text{base}} \cdot t_{\text{loss}} \\ & + \Delta q^*_{\text{inc}} \cdot \max(0, t_{\text{inc}} - t_{\text{I}}) \\ & - \Delta q^*_{\text{dec}} \cdot \max(0, t_{\text{dec}} - t_{\text{I}})\ ) \end{aligned} \tag{8.11}$$

$$\begin{aligned} \text{where :} & \\ q_{\text{base}} &= \text{quality of life without intervention} \\ t_I, t_{\text{base}} &= \text{years left to live with/without intervention} \\ t_{\text{loss}} &= \text{years lost due to intervention} \\ k_{\text{dec}}, k_{\text{inc}} &= \text{harm/benefit to quality of life} \\ q_{\text{dec}}, q_{\text{inc}} &= \text{worse/better quality of life after intervention} \\ t_{\text{dec}}, t_{\text{inc}} &= \text{duration of harm/benefit to quality of life} \\ p_{\text{dec}}, p_{\text{inc}} &= \text{probability of harm/benefit occurring} \\ Q_{\text{dec}}, Q_{\text{inc}} &= \text{total subjective gain/loss in quality of life} \\ T_{\text{dec}}, T_{\text{inc}} &= \text{total subjective gain/loss in years of life} \end{aligned}$$

Figure 8.5 shows the effect of this kind of pre-processing. Given the choice between a life twice as long but at half of the current quality of life or a shorter life at a higher quality of life, the option will look subjectively different depending on what one values more, even though the objective product of time and quality is the same. Analogously, an operation that comes with the risk of an earlier death but entails an increase in the quality of life will appear more attractive to a patient who focuses on potential benefits and is prepared to take risks to obtain them. It will look less attractive to someone who is more risk-averse.

Evolving the Fuzzy Cognitive Map We optimize the model with an EA to learn three sets of parameters: b and r used to calculate the transfer function for each node and the weight matrix W that determines the sign and strength of all the connections in the graph. Given the FCM's 21 vertices and 54 connections, there are 96 learned parameters (two values per node, one per connection). The initial values for b are drawn from a random normal distribution with a mean of 0.5 and a standard deviation of 0.05. The initial values for r and W are drawn from a random uniform distribution in the interval $[1.5, 5]$ and $[-1, 1]$, respectively. Weights on the diagonal of the adjacency matrix are not optimized and are held constant at 1.0. We let the algorithm run for 1000 generations and 30 solutions per population. Per generation, seven parent solutions are chosen using steady-state selection. We use random mutation and set the percentage of genes to mutate to 0.05. The fitness function utilized to evaluate each generation is the multiplicative inverse of the Root Mean Squared Error (RMSE)

$$\text{Fitness} = \frac{1}{RMSE} \tag{8.12}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{8.13}$$

where N is the number of cases, y is the prediction of the model calculated using the current r, b, and W parameters, and $\hat{y}$ is the ground truth provided by the human annotators.

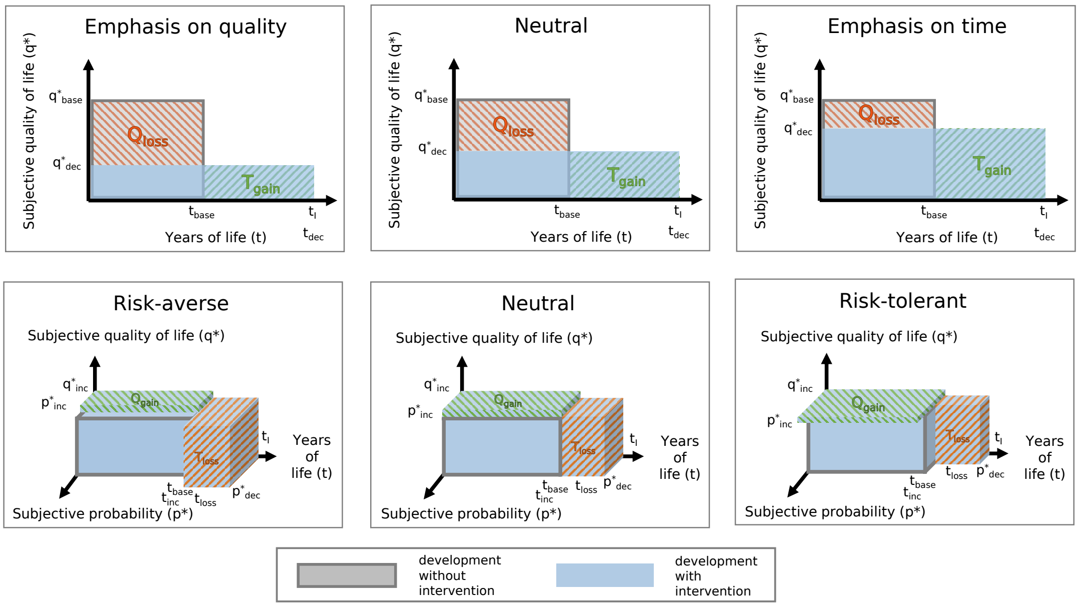


Figure 8.5: Schematic visualization of the re-weighting of quality of life and probabilities for the estimated consequences of an intervention, depending on the patient's preferences. The first case (top row) shows a scenario without uncertainties regarding patient outcomes. The intervention doubles the patient's life expectancy but halves their quality of life. The "subjective value" areas for the scenario without the intervention (gray) and with the intervention performed (blue) are equal to a neutral person (center). For a person preferring quality of life (left), the gray area is larger than the blue. For an individual emphasizing years of life (right), the blue area is larger than the gray. The second case (bottom row) shows a scenario where there is an equal probability that an intervention may improve the patient's quality of life and reduce their life expectancy. This yields equal-depth subjective value cubes for a neutral person. To a risk-averse individual (left), the negative outcome looks more probable and the positive less likely. For a risk-tolerant person (right), the opposite is the case.

8.1.4 Results

Evaluation metrics Due to the small size of our data set, we evaluated the model with stratified *k*-fold cross-validation. The number of folds was set to 3, and the remaining parameters were left at their default values. The cases were stratified by case category. For each of the three folds, ten models were optimized using each fold's training set. Models were run for a maximum of 50 timesteps and stopped earlier if the difference between all node activations in two successive time steps was ≤ 0.01. We then compared the output of a model's intervention recommendation to the labels in a fold's test set. We report the average performance over the 10 models as well as the standard deviation, using Mean Absolute Error (MAE) and the model's binary accuracy as metrics.

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \tag{8.14}$$

$$\text{Acc} = \frac{(TP + TN)}{TP + FP + TN + FN} \tag{8.15}$$

TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives. We here regard cases in which both the prediction and the ground truth > 0.5 as true positives, cases in which both the prediction and the ground truth ≤ 0.5 as true negatives, cases in which the prediction > 0.5 and the ground truth ≤ 0.5 as false positives, and cases in which the prediction ≤ 0.5 and the ground truth > 0.5 as false negatives.

Performance One challenge in trying to optimize and evaluate a decision-making model for cases in medical ethics is the scarcity of available data. Collecting examples is a labor-intensive task, as each case must be annotated and assessed by an expert. We were, therefore, not able to collect a very large data set. Furthermore, in ethics, the right course of action is much more debatable than in nearly any other domain, with even experts occasionally disagreeing about what to recommend in one and the same case. We nonetheless report a quantitative measure of the models' deviation from the consensus between our experts. However, our main focus in this evaluation is on *how* the evolved model emulates the decision-making process, i.e., on qualitative results.

Table 8.3 shows the average MAE and classification accuracy over three folds and 10 different models. As can be seen, the models approximate the training cases very well, but performance experiences a drop in test cases – a classic sign of overfitting. This is not too surprising on a data set as small as this one. One may also notice that, for our input data and FCM configuration, applying evolved activation functions yielded much better results than the commonly used *tanh* activation function. We present some examples of these evolved activation functions in the next section.

Evolved activation functions Figure 8.6 shows node activation functions that have been adapted through the EA. The function for AUTONOMY almost resembles a step function, leading to an activation only if both of its inputs (decisional capacity and age of majority) are high enough. This is in line with regarding decisional capacity in patients as a binary notion, which many clinicians indeed do. The evolved function shown for BENEFICENCE, which receives inputs preprocessed as described in section 8.1.3, is almost linear. The evolved function for INTERVENTION has a more conventional sigmoid shape, close to what is often used in FCMs and output layers of neural networks.

Evolved weights Inspecting the static weights of an FCM tells only part of the story, as dynamic effects emerge from the interplay of input data, activation functions, and weights over multiple time steps. However, we do see some elements that are in accordance with human ethical intuition, such as the status of the patient's decisional capacity and the variable "age of majority" having an equal influence on AUTONOMY, and AUTONOMY having a positive influence on FOLLOW THE PATIENT'S WISH.

Table 8.3: Results of 3-fold cross-validation over 10 models.

Activation function	Avg. MAE		Avg. Acc.	
	Train	Test	Train	Test
tanh	0.47± 0.33	0.47± 0.33	0.50± 0.05	0.50± 0.06
evolved	**0.11± 0.07**	**0.23± 0.12**	**0.92± 0.10**	**0.75± 0.20**

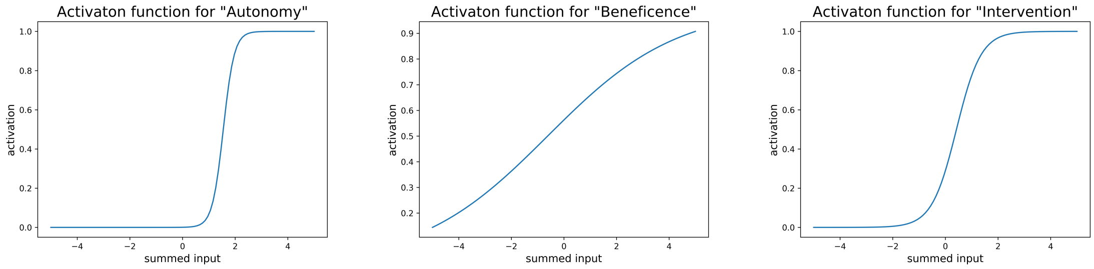


Figure 8.6: Examples of evolved concept activation functions.

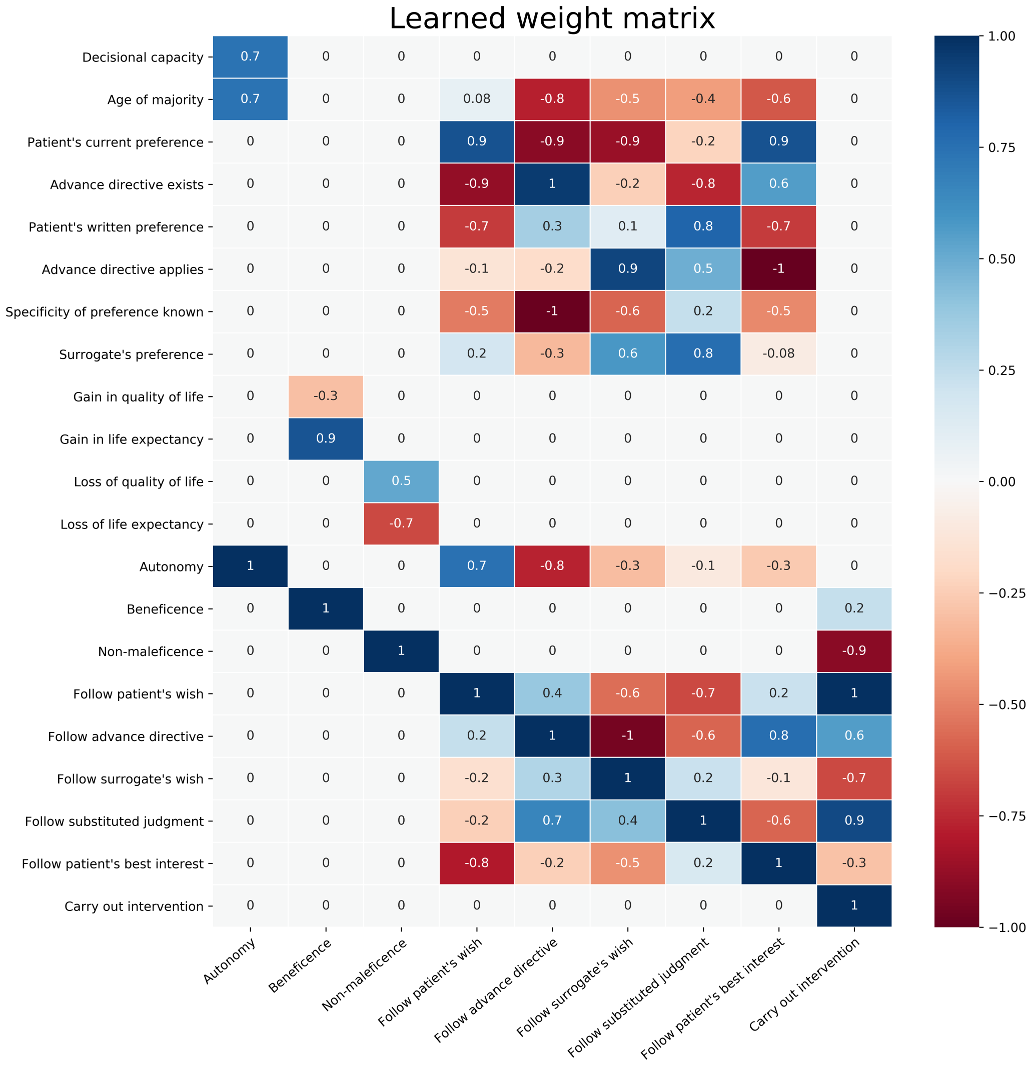


Figure 8.7: Heatmap of a model's evolved weights.

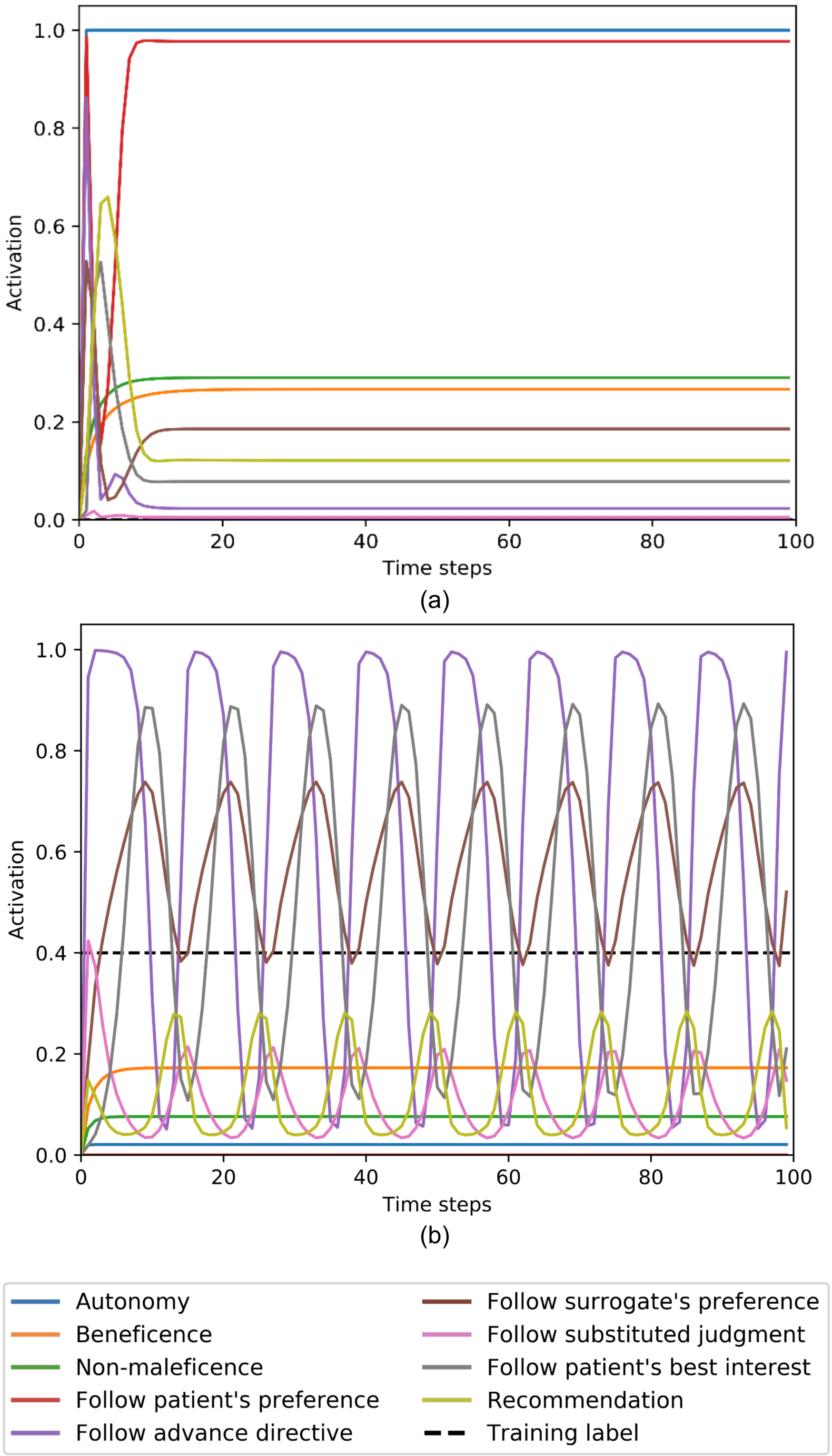


Figure 8.8: Activation values of the model's nodes over the course of simulation for two example cases – one concerning the competent refusal of treatment (top) and the other concerning advance directives (bottom). In the first case, the model quickly converges to a fixed point, whereas it enters a limit cycle in the second case. In both cases, the intervention recommendation (light green) and the ground truth (dotted) are below 0.5, which would constitute a correct classification.

Example Simulations To show the components inspected above in action, we visualize the activations for the concepts in the FCM over time for two example cases. In Figure 8.8a, a Jehovah's Witness refuses a blood transfusion that would save their life. The ethically "correct" answer in this case is not to intervene because the patient has taken this choice fully informed and in possession of decisional capacity. In the simulation, activations for the concepts of Follow the patient's best interest and Intervention initially increase but quickly fall again as the concept of Follow the patient's wish begins to dominate.

In Figure 8.8b, an adult patient with multiple sclerosis has signed a do-not-resuscitate order many years ago. After a suicide attempt, her husband wants doctors to continue treating her. Here, the concepts of Follow the advance directive, Follow the patient's best interest, and Follow the surrogate's wish are in conflict. This is a rare example in which the system entered a limit cycle. Most simulations converged to fixed points. Although limit cycles are often regarded as something to be avoided in FCMs [MA06], here they provide an interesting figurative illustration of the ethical dilemma at play. Despite oscillating, the activation magnitudes show the ethically correct ordering – Follow the advance directive is highest, next comes Follow the patient's best interest, then Follow the surrogate's wish. The activation for Intervention also stays between 0.0 and 0.3 and thus below or around the expert-specified solution of 0.4.

8.2 Discussion

In this case study, we showed that an FCM with evolved activation functions and weights could be used to model the decision-making process behind several categories of medical ethics cases, and we found the model's main advantage to be its high interpretability. Because nodes and edges have human-assigned semantics, FCMs can be intuitively visualized as causal graphs – and weights, activation functions, and simulation runs can be meaningfully inspected. Furthermore, the FCM permits the addition, deletion, and modification of nodes and connections. This way, the model can also be adapted to regional, cultural, and juridical differences or expanded to accommodate more types of medical ethics cases.

Although our model was already able to approximate expert decisions to some extent with our limited data set, more examples would be needed to reach higher accuracies. Collecting additional data from different ethicists would also allow us to better estimate inter-annotator agreement and compare it with the model's deviation from human-provided answers. Furthermore, data pre-processing steps could be adapted to handle more realistic disease progressions, and Beauchamp and Childress' principle of justice could be integrated into our existing FCM.

As it stands, the model would not be ready for clinical use but could, for example, be used for educational purposes such as training medical students and aspiring ethicists. Another possible use case would be to collect data from ethicists of different cultures or institutions. One could then make use of the interpretable properties of the FCM by evolving separate models for the different sets of ethicists and inspecting the similarities and differences between the various evolved decision models. Both of these potential use cases relate to the idea of models as mediators, which I briefly outline below.

8.2.1 Models as mediators

Morrison and Morgan's account of models as mediators sees models as essential tools that bridge the gap between theoretical frameworks and observations. Models are constructed with elements from both theory and empirical evidence, as well as additional components that may not be directly derived from either domain. This construction grants models a form of partial autonomy, which enables them to mediate between theory and the world. Models can operationalize theories so that their empirical implications may be explored. Models can also simulate conditions that are difficult or impossible to replicate in reality, thereby extending the reach of scientific investigation. Because of their status as partially independent tools, they can provide a kind of epistemic access that is not available through either pure theory or direct observation alone. We can distinguish between two ways in which models may be a source of learning: model use and model construction [MM99].

Learning from using models Knowledge and understanding are usually seen as the result of successful explanations [Dié15]. However, explanation is not the only path to learning. Polanyi uses the term "tacit" to describe the kind of knowledge that cannot be fully captured through explicit instructions or formal language but is instead acquired through experience and practice. In the context of scientific modeling, researchers often develop a tacit "feel" for a model's behavior and implications. They may, for instance, develop an intuitive understanding of how certain parameters affect the system's behavior, how to interpret model results, and what a model's strengths and limitations are. This implicit knowledge emerges through hands-on experience with a model and is often challenging to articulate or teach directly [Gel16b].

There are several ways in which models can support tacit learning. As concrete artefacts, models provide a physical or virtual embodiment of theoretical concepts or systems [Knu11]. By manipulating these models, epistemic agents can explore the implications of changes in variables or conditions in a controlled, observable manner [JS06]. This interaction can help transform abstract concepts into tangible experiences that foster a deeper understanding of underlying mechanisms or principles [Knu11]. It provides a direct, experiential form of learning that is often more impactful than passive observation or theoretical study alone [Gel16d].

By providing a physical or visual representation that can be interacted with, models help to externalize thought processes [Knu11]. Making complex ideas more accessible in this way can be helpful on an individual level, but it also facilitates collaborative learning and communication [JS06]. Models can serve as powerful tools for sharing ideas or theories in an intuitive manner [Gel16d]. Allowing multiple users to engage with a model simultaneously can promote discussion, debate, and collective problem-solving [JS06]. In contexts like education and interdisciplinary research, models can thus bridge communication gaps between participants with diverse backgrounds and expertise [Knu05].

Learning from constructing models In addition to the epistemic value of interacting with an existing model, the process of model building itself can be a significant source of tacit knowledge [MM99; RK22]. Building a model compels epistemic agents to engage actively with a subject matter and articulate their current understanding and personal theories about it [MM99]. It requires identifying and delineating problem spaces and encourages the formulation of specific epistemic aims [GP17]. Much like the process of writing, modeling represents a way of externalizing ideas and facilitating cognitive processing and reasoning [Knu11].

Constructing a model can help make tacit knowledge explicit and organize it in a way that allows learners to critically examine their assumptions and understanding [MM99]. Through this examination, they can identify gaps in their knowledge, misconceptions, and areas requiring further exploration [GP17]. The result is usually a continuous, iterative process of model refinement that leaves a "cognitive residue" [SPG91]. I.e., the result of modeling is not just the external artefact but also the builder's mental model and internal conceptual change. Furthermore, when modeling takes place in collaborative settings, it allows learners to pool their knowledge and construct a shared understanding of a system, process, or set of domain principles [JS06].

8.2.2 Relation to the guiding questions

Among the works discussed in this thesis, this sixth and last one likely represents the most atypical use of ML models. At first glance, the building blocks of the modeling process are the same as we have previously seen (see Figure 8.9). As in most of the case studies so far, a domain-specific theory or body of literature (in this case, Beauchamp and Childress' prima-facie principles) forms the starting point of our investigation. This theory informs both the design of our FCM and how we collect the input features and human judgments that make up our dataset. In contrast with the case studies presented so far, these design decisions were made in collaboration with medical ethicists. As discussed in previous chapters, this process was, as is unavoidably the case in all modeling endeavors, also shaped by implicit assumptions. In traditional ML fashion, we then train our model and assess its accuracy on the dataset.

However, the model's performance is of secondary importance in this case. It only serves to show that the proposed approach works well enough to provide a proof of concept. Instead, the study's main contribution lies in operationalizing medical ethics theory into a concrete model. This model can then be used to communicate ideas across disciplinary boundaries and form the basis to discuss both the suitability of the approach and, potentially, the human decision processes it was designed to replicate. Let me briefly elaborate on the second point.

The model is optimized to mimic human labelers' judgments on medical ethics cases. These judgments are informed by the labelers' values and priorities, which affect how they weigh the factors of a case against each other. This weighing is usually a more or less implicit process. However, the tacit assumptions that underlie human decisions may manifest in certain patterns, which the model can be expected to capture when trained on the labelers' decisions. Analyzing what the model has learned could thus help us better understand how human experts reason about medical ethics cases, including any unconscious subjective biases that play into their decisions and may go unnoticed otherwise [Hey+17; Haa+20].

With the points above in mind, let us return to the guiding questions of this thesis, starting with the objection against NNs due to their supposed lack of theoretical basis. The case study illustrates how models can mediate between theory and the world. By operationalizing an abstract theory like the prima-facie principles and translating it into a concrete FCM, we can explore its empirical implications in a way that is not possible by simply looking at words on a page. As a matter of fact, even calling the qualitative and mutually conflicting prima-facie principles a theory may be an over-statement by quantitative standards. In line with the fourth function of exploratory modeling proposed by Gelfert (see section 7.2.3), bringing the empirical implications of such a framework to light can help raise new questions about the suitability of the model's target system.

Indeed, several of the commentaries sparked by our original target article discussed whether principlism (or the way we implemented it) was an appropriate basis for decision-making in medical ethics [GB22b; BFG22; PB22; Rah+22; DFR22]. Thus, the case study serves as an

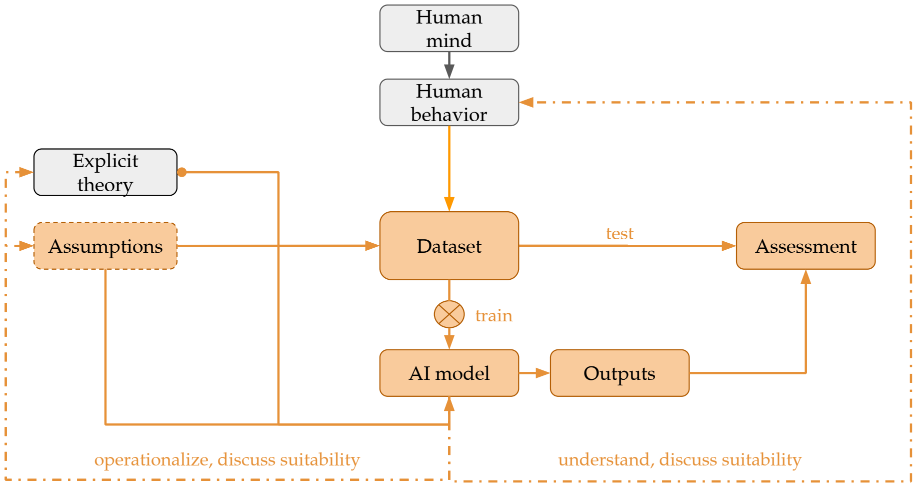


Figure 8.9: Overview of relations between human cognition, theory, assumptions, data, model, and outputs in the sixth case study. Relevant components performed or generated by us shown in orange. Components provided by third parties shown in gray. Circle at the beginning of an arrow indicates the starting point of investigation.

example of how a model can provide a "tool for critical self-reflection by [...] revealing causal assumptions, uncovering limitations of its adopted variables of interest [...] and suggesting the need for alternatives" [AG19, p. 445].

The second criticism against NNs, namely, their lack of biological plausibility, certainly applies here. Out of all the case studies in this thesis, this model is perhaps the furthest from representing the biological brain: It consists of only 21 nodes, all of which have manually defined semantics. In that sense, the FCM is more akin to the bottom-up cognitive models we sought to complement with our top-down NN approach in the last two case studies. However, it is precisely the model's simplicity and human-interpretable components that allow it to fulfill its exploratory and, as we will come to shortly, explanatory epistemic aims. Thus, the study serves as another case in point that maintaining a plurality of models is important in any discipline, as different modeling paradigms can have complementary strengths and weaknesses.

Finally, this case study brings a new perspective to the third criticism against neural networks, i.e., their inability to provide explanations. I have previously discussed the two issues contained in this criticism, namely, whether we can explain neural networks themselves and whether we can explain phenomena *through* neural networks. As in previous case studies, we include some post-hoc analyses of, e.g., model weights or activation functions. However, what is particularly interesting about this study is that it demonstrates that explanation in the traditional sense does not constitute the only path to understanding a model or a real-world phenomenon.

As emphasized in the models-as-mediators account by Morrison and Morgan, the fact that models are manipulable artefacts allows for a kind of interaction that can foster a tacit "feel" for a system. In our specific case, the FCM indicates how Beauchamp and Childress' principles are weighted and even provides graphical visualizations of this process. Since the model can be visualized as a directed causal graph with human-designated semantic meanings, one can understand and interpret these weightings in terms of cause-effect chains. Observing the model's behavior, therefore, enables users to compare its processing of the respective case to their own reasoning, which will be instructive irrespective of whether one agrees or disagrees with the machine-generated solution.

Importantly, this kind of model interaction is possible in real time. Gaining hands-on experience through engaging with reactive systems is undoubtedly more informative for users, such as medical students or aspiring ethicists, than working solely with static training devices like ethics books. The causal structure embedded in the FCM allows for probing counterfactual scenarios and for assessing the quantitative impact of individual parameters: how does the weighting of the principles of beneficence, non-maleficence, and autonomy change if one drops a particular assumption or adds a novel one? What influence does the modification of each variable exert on the final recommendation? Through this kind of self-guided exploration, users can build up an intuitive understanding of the model and its target domain that would be difficult to acquire through verbal explanation alone.

In addition to the model's explanatory potential for users, it also served as a valuable tool in our interdisciplinary collaboration as designers. Producing a concrete implementation of a model of medical ethics decision-making required us to engage deeply with the topic and articulate our collective understanding and perspectives. The process compelled us to identify key ethical dilemmas, evaluate different theories regarding whether and how they could be operationalized, and clarify our epistemic aims with the project. Through testing a range of scenarios and model architectures, we iteratively defined our scope and converged on a shared understanding of what we were modeling. Thus, the act of constructing the FCM presented in this case study not only resulted in an external artefact. It also left a lasting imprint on our mental models and conceptual frameworks that a more conventional form of explanation could not have provided.

To sum up, models, including neural networks, can serve as mediators between theory and the world and provide a kind of epistemic access that is not available through abstract frameworks or observation alone. Models such as the one presented in this study can serve as tools for externalizing thought processes, communicating ideas, and assessing their suitability. Furthermore, hands-on experience with models can lead to a tacit, intuitive understanding of a model and its target system that may be difficult to achieve with conventional, verbal explanations. Importantly, neural networks can fulfill these epistemic functions without needing to be biologically plausible.

9 General discussion

> *Mirrors are wonderful things. They appear to tell the truth, to reflect life back out at us; but set a mirror correctly and it will lie so convincingly you'll believe something has vanished into thin air, that a box filled with doves and flags and spiders is actually empty, that people hidden in the wings or the pit are floating ghosts upon the stage. Angle it right and a mirror becomes a magic casement; it can show you anything you can imagine and maybe a few things you can't. Stories are in one way or another mirrors. We use them to explain to ourselves how the world works or how it doesn't. [...] Fantasy is a mirror, a distorting mirror, and a concealing mirror set at 45 degrees to reality, but a mirror nonetheless, which we use to tell ourselves things we might not otherwise see.*
>
> Neil Gaiman

In this thesis, I set out to examine the debate surrounding the use of NNs in the study of cognition. I began by outlining the key arguments on both sides of this discussion. Namely, proponents highlight both the ability of NNs to account for higher-level cortical processing and the level of experimental access that *in silico* studies offer. On the other hand, critics argue that NNs lack a theoretical basis, are not biologically plausible, and do not provide explanations. I then explored how these criticisms relate to different philosophical perspectives on the nature and role of models in science. This led me to the following questions: How valid are each of the three criticisms against the use of NNs in the study of cognition, and what are the implications for the epistemic utility of NN models?

I explored these questions through six case studies, each focusing on a different aspect of cognition research. Each case study addressed a set of targeted questions and posed its own implementational challenges. Mirroring the structure used throughout this thesis, I first summarize my technical approaches to overcoming those challenges in Table 9.1. For each case study, I then provide an overview in Table 9.2 of how each NN model relates to the three criticisms and what kind of epistemic value it can offer.

As shown in Table 9.1, I employed a diverse range of models, from state-of-the-art DL architectures to more classic or less well-known techniques – often within the same study. These choices were made by taking into account the epistemic goals of each work. Because my thesis operates at the intersection of ML and cognitive science, these goals reflect the epistemic interests of both disciplines. The first two studies align mostly with the performance-driven paradigm dominant in DL. In the fourth and fifth studies, I made use of similar DL techniques but adapted them to the aims of cognitive science. I employed cognitively-inspired diagnostic tasks and performed more critical, in-depth post-hoc analyses than are commonly offered in ML publications. The third and sixth studies represent a less typical way to employ NNs, namely, as tools for simulating and learning from hypothetical scenarios. Taken together, I hope these works demonstrate the wide range of uses to which we can put the ML toolbox and how we can benefit from drawing on a plurality of modeling paradigms.

Informed by the six case studies, I now want to return to my guiding questions and summarize my overall conclusions.

Table 9.1: Overview of the technical challenges, methodology, and results of the six modeling case studies.

Case study 1: Modeling patient-specific activity patterns with Transformers to detect psychotic and non-psychotic relapses		
Task	Identifying psychotic and non-psychotic relapses in patients using biosignals captured by wearable sensors	
Challenges	• Missing and corrupted real-world data • Only unsupervised anomaly detection possible • Inter-patient variability	⟹ Feature engineering and data preprocessing ⟹ Proposed timestamp prediction as pre-training task ⟹ Patient-individual models
Model(s) used	Transformer (ensembles)	
Optimization	Backpropagation	
Model analyses	N/A	
Results	Ranked 1st on Track 2 and 3rd on Tracks 1 of the 2024 ICASSP e-Prevention Grand Challenge	

Case study 2: Modeling the emergence of compositional generalization on the grounded SCAN dataset with selective attention		
Task	Investigating the potential of two human-inspired inductive biases (selective attention and egocentric location encoding) to improve performance and sample efficiency on gSCAN	
Challenges	• Implementation of selective attention • Identification of factors contributing to performance	⟹ Non-differentiable attention module optimized via EA ⟹ Minimal model to allow for extensive analyses
Model(s) used	• ESN • MLPs	
Optimization	Backpropagation and CMA-ES	
Model analyses	• Ablation studies • Neuron pruning • Statistical error analyses • Loss landscape visualization	
Results	• ≈ 60 times fewer parameters than previously proposed models • Accuracies comparable with previously proposed models on most test splits, even when trained on only 2% of the full dataset • Outperforms previously proposed models by 65 to 86% on adverb-to-verb generalization	

Table 9.1: Technical challenges, approaches, and results of the six case studies. (Continued)

Case study 3: Modeling the emergence of letter shapes with drawing-based signaling games		
Task	Exploring factors *in silico* that have been hypothesized to influence the shapes of letters in human writing systems	
Challenges	• Approximating human communicative situations • Lack of pre-existing datasets • Lack of automated metrics for letter shape analysis	⟹ Drawing-based sender-receiver game setup ⟹ Self-composed datasets ⟹Used HOG and proposed a symmetry metric based on auto-correlation
Model(s) used	• CNNs • HuBERT (pre-trained, self-supervised speech model) • Linear models	
Optimization	Backpropagation and CMA-ES	
Model analyses	• Statistical analyses and visualization of evolved glyphs • Ablation studies • Association rule mining on evolved glyphs	
Results	Identified statistics of pre-training data, canvas shape, and architectural model properties as relevant factors	

Case study 4: Modeling the emergence of intuitions about agents' goals, preferences and actions with Video Transformers		
Task	Testing the ability of the Transformer attention mechanism to capture relationships between agents and their goals from observation	
Challenges	• Input dimensionality (length of BIB video clips) • Preserving agent identities	⟹ Selective attention using only top-*k* patches ⟹ Auxiliary reconstruction loss
Model(s) used	Video Transformer	
Optimization	Backpropagation	
Model analyses	• Statistical error analyses • Occlusion analyses • Visualization of layer activations • Decoding experiment using linear probes	
Results	Ranked 1st in the Machine Visual Common Sense Challenge's BIB Track at the 2022 European Conference on Computer Vision	

Table 9.1: Technical challenges, approaches, and results of the six case studies. (Continued)

Case study 5: Modeling the emergence of early number abilities with Vision-Language Transformers		
Task	Comparing the behavior and internal representations of a model trained on a range of early number skills to humans	
Challenges	• Lack of existing datasets • Model opacity	⟹ Self-created dataset ⟹ Design of targeted analyses and adoption of XAI techniques
Model(s) used	Multimodal Transformer	
Optimization	Backpropagation	
Model analyses	• Analysis of training trajectories • Targeted error analyses • Logit lens • Probing studies • Ablation studies • Visualization of embeddings • Visualization of an information flow • Activation pattern comparisons	
Results	• Model behavior largely aligns with reported human data • Emergence of functional organization in the model	

Case study 6: Modeling decision-making processes in medical ethics with Fuzzy Cognitive Maps		
Task	Examining the feasibility of an AI-based decision support system for medical ethics cases	
Challenges	• Choice of ethical framework • Operationalization of abstract ethical principles • Accounting for patient-specific preferences • Lack of existing datasets • Subjectivity of ethical judgments	⟹ Used prima-facie principles ⟹ Close collaboration with medical ethicists ⟹ Individual pre-processing based on prospect theory ⟹ Creation of suitable dataset ⟹ Supplementation with qualitative evaluation
Model(s) used	FCM	
Optimization	EA	
Model analyses	• Inspection of evolved weights and activation functions • Visualization of example simulations	
Results	• Approximates judgments of medical ethicists fairly well • Handles a much wider range of cases than previous proposals	

Table 9.2: Overview of the relation between the guiding questions of this thesis and the six modeling case studies. Degree to which the three criticisms against NNs apply listed on the left. Implications for the epistemic utility of NNs listed on the right.

Case study 1: Modeling patient-specific activity patterns with Transformers to detect psychotic and non-psychotic relapses		
Connection to theory	Very limited: • Only loose connection through abductive inference • Patterns identified through induction as potential inspiration for follow-up real-world experiments	⟹ For *some* use cases, NNs can have pragmatic utility as epistemic enhancers by virtue of their predictive performance without necessarily having to be based in theory, biologically plausible, or provide much in the way of explanation.
Biological plausibility	Very limited: • Not considered	
Ability to offer explanation	Very limited: • Only teleological	
Case study 2: Modeling the emergence of compositional generalization on the grounded SCAN dataset with selective attention		
Connection to theory	Moderate: • Model and dataset at least partially informed by theory and created in pursuit of questions relevant to cognitive science	⟹ NNs can be connected to theory by taking inspiration from the cognitive science literature in designing training environments, architectural constraints, forms of regularization, or augmented loss functions. This not only allows for the investigation of questions relevant to cognitive science but may, in some cases, also improve NN performance.
Ability to offer explanation	Moderate: • On an individual level, offers explanations at different levels of the Marr hierarchy • On a meta-level, comparing different models on a dataset like gSCAN can serve to reveal relevant success factors	⟹ NNs can be explained at various levels, and benchmarking the predictive performance of different models can serve as a stepping stone to explanation. Whether an explanation is successful always depends on the individual needs of the epistemic agent.
Biological plausibility	Limited: • Translation of inspirations from cognitive science takes place at a highly abstract level	⟹ Idealization may decrease biological plausibility but can be a useful epistemic strategy to isolate factors of interest.

Continued on next page

Table 9.2: Relation between guiding questions and the six case studies. (Continued)

Case study 3: Modeling the emergence of letter shapes with drawing-based signaling games		
Connection to theory	Considerable: • Computational implementation of an empirically supported hypothesis	⟹ NNs are usually connected to theory by virtue of the research questions that motivate their construction and the knowledge built into them. This link can take a more explicit form in the case of embedded toy models.
Biological plausibility	Very limited: • Highly simplified and idealized toy model	⟹ NNs can be of epistemic use without faithfully representing the brain – in fact, they need not represent any existing target system at all. Models have a modal dimension, meaning they can be used as tools for exploring hypothetical systems or scenarios.
Ability to offer explanation	Moderate: • Model serves as an embodied "how-possibly" explanation of empirical regularities in letter shapes	⟹ NNs can not only be the target of explanations but can serve as (how-possibly) explanations of real-world phenomena in their own right. For NNs to serve this function, it may not be necessary to understand the model itself in detail – high-level "what" and "why" explanations may be enough.
Case study 4: Modeling the emergence of intuitions about agents' goals, preferences and actions with Video Transformers		
Connection to theory	Limited: • Dataset (but not NN model) informed by cognitive science theory	⟹ Process models in the form of NNs can allow for a more open exploration than as-if theories.
Biological plausibility	Very limited: • Direct comparison with human infants on a behavioral level, but no attempt at increasing plausibility of the model itself	⟹ Analyzing what NNs learn can prompt questions about what kind of behavior we consider desirable or cognitively plausible, and help iteratively refine the design of modeling setups that are suited to those goals.

Continued on next page

Table 9.2: Relation between guiding questions and the six case studies. (Continued)

Ability to offer explanation	Limited: • Provides some “what”, “why”, and “where”-level analyses, but mainly serves to raise questions rather than provide explanations	⟹ NNs can provide alternative or complementary explanations to more normative frameworks, and serve as starting points for future investigations.
Case study 5: Modeling the emergence of early number abilities with Vision-Language Transformers		
Connection to theory	Limited: • Educational psychology informs dataset design, and model analyses are contextualized within the numerical cognition literature	⟹ When comprehensive theoretical frameworks are not yet available, NNs can serve as exploratory models that provide a starting point for future inquiry or proof-of-principle demonstrations.
Biological plausibility	Limited: • No attempt at increasing biological plausibility of the model itself, but increased task and model complexity provides a step towards more cognitively plausible models compared to previous work	⟹ Given the ability of NNs to process complex inputs, they can help cognitive scientists expand the scope of phenomena they are able to model. Conversely, AI developers can take inspiration from the cognitive sciences to design ecologically relevant training environments and potentially move towards more cognitively plausible models.
Ability to offer explanation	Considerable: • Uses a wide range of techniques from interpretability research to investigate model behavior and representations	⟹ NNs are amenable to a much wider range of post-hoc analysis techniques than biological brains.
Case study 6: Modeling decision-making processes in medical ethics with Fuzzy Cognitive Maps		
Connection to theory	Considerable: • Designed to operationalize Beauchamp and Childress’ prima-facie principles	⟹ Models, including neural networks, can serve as mediators between theory and the world and provide a kind of epistemic access that is not available through abstract frameworks or observation alone.

Continued on next page

Table 9.2: Relation between guiding questions and the six case studies. (Continued)

Biological plausibility	Very limited: • Consists of only 21 nodes, all of which have manually defined semantics	⟹ Models can be used for externalizing thought processes and communicating ideas without needing to be biologically plausible.
Ability to offer explanation	Considerable: • Provides a reactive tool through which users can explore medical ethics dilemmas and build up an intuitive understanding of the model and its target domain	⟹ Hands-on experience with (NN) models can lead to a tacit understanding of a model and its target system that may be difficult to achieve with conventional explanations.

9.1 Validity of the criticisms against neural networks in the study of cognition

I begin by discussing the validity of the three criticisms against the use of NNs. I will briefly recapitulate each argument and then provide my responses to it, drawing from the modeling case studies and epistemological discussion I have presented in this thesis.

9.1.1 Lack of basis in theory

The first criticism against the use of NNs in the study of cognition is that these models were not designed to test hypotheses about biological brains. Design decisions in DL are more informed by heuristics and engineering goals than pre-existing knowledge of neural computation. Thus, critics argue there is an insufficient theoretical link between AIs and cognitive science.

Most (NN) models are implicitly linked to what we know of the world. It is true that the field of DL is primarily focused on task performance and computational efficiency rather than testing hypotheses about biological brains. However, the construction of models is always informed by some degree of theory or existing knowledge. This includes performance-focused models such as the ones presented in chapters 3 and 4. Every modeler brings with them a set of values, assumptions, and expertise that shape the kinds of questions they ask and the modeling choices they make. Even models of hypothetical systems, such as the one in chapter 5, are rarely developed in a theoretical vacuum but are connected to our understanding of the world through the considerations that motivated their creation.

The connection between NNs and theory can be strengthened in several ways. The connection to theory can be enhanced by building on the cognitive science literature when designing and training NNs. This can take different forms: explicitly building "embedded toy models" (as in chapter 5), designing ecologically relevant learning tasks and environments (as in chapters 6 and 7), or building cognitively inspired inductive biases into NNs through architectural constraints, regularization techniques, or loss functions (as in chapter 7). Taking inspiration from cognitive science in this way not only allows for the investigation of questions relevant to the study of the mind, but may also improve the performance of NN models.

Less principled (NN) models can have exploratory benefits. At the same time, the lack of a strong underlying theory can actually be beneficial in certain contexts. As noted by Thomas

S. Kuhn, paradigms influence the kinds of questions that can be asked [Kuh97]. In contrast to highly idealized as-if theories, process models like the ones discussed in chapters 6 and 7 can take a more open-ended, descriptive approach. They thus have the potential to uncover unexpected patterns or relationships that may not “fit into” existing frameworks. This move towards a more nuanced Galilean psychology can be particularly valuable in contexts like cognitive science, where widely accepted theories are often not yet available.

In such cases, exploratory modeling with NNs can serve as a starting point for future inquiry, provide proof-of-principle demonstrations, suggest potential explanations for observed phenomena, and help assess the suitability of a target phenomenon for investigation. Thus, rather than being hindered by a lack of theory, researchers can leverage this flexibility to take a more inductive, data-driven approach to uncover patterns and mechanisms that may then inform the development of more robust theoretical frameworks. The exploratory phase can lay important groundwork and open up new lines of inquiry that would not be possible if the research was overly constrained by existing theory from the outset.

9.1.2 Biological implausibility

The second criticism against the use of NNs in the study of cognition is that NNs lack most of the dynamics of biological neural networks. For instance, they do not produce spike-based representations, have different constraints regarding power efficiency and memory, and use the biologically implausible backpropagation algorithm for training. Critics argue that this renders them useless as stand-ins for the brain.

Isomorphic representation is not always necessary and can even decrease epistemic usefulness. The criticism that NNs are not biologically plausible is a valid one. It is true that NNs lack many of the dynamics and constraints of biological “wetware”. Even proponents concede that NNs can, at best, be considered highly abstract models of the brain. However, the biological plausibility of a model is not always a relevant or necessary criterion for its epistemic utility. All models, to some degree, are “false” or inaccurate depictions of reality. The act of modeling inherently involves idealization and simplification. This can take the form of Galilean or minimalist approaches, depending on whether an idealization is meant to be “corrected” or not. The key is to ensure that the model does not diverge too far from reality, rely on pseudoscientific ontology, or have insufficient predictive power. However, when these conditions are fulfilled, idealization is often the very thing that allows us epistemic access to a system. For example, our ability to derive insights from the models in chapters 4 and 5 hinged on omitting factors irrelevant to the core phenomena being modeled.

Cognitive science and AI can inform each other to create more plausible models. Cognitive science and AI can work together to create NN models that, while not necessarily biologically accurate, are at least more cognitively plausible. As shown in chapter 7, adapting the ML pipeline to the epistemic purposes of cognitive science, rather than relying on existing ML benchmarks and performance metrics, can help model cognitive phenomena that only emerge at a sufficiently high level of complexity. Creating more cognitively grounded tasks, such as gSCAN or BIB, and using them to benchmark different NN models can aid in pre-selecting promising candidates for further inquiry. On the other hand, AI researchers can benefit from carefully designing their modeling setups and drawing on insights from cognitive science. Incorporating expertise on the experiences that shape human learning can inform the design of ecologically relevant training inputs, which can induce more human-like representations and processing in NNs in desirable ways.

9.1.3 Inability to offer explanations

The third and last criticism against the use of NNs I have considered in this thesis is that they do not offer explanations. Critics argue that the "black box" nature of NNs inherently limits our ability to understand both the models themselves and their target phenomena.

Epistemic opacity is not an inherent property of (NN) models. The question of whether NNs can offer explanations is a complex one that does not have a simple yes or no answer. As we have seen, there are many different notions of what constitutes an explanation: deductive-nomological, inductive-statistical, causal mechanistic, unificationist, or pragmatic accounts, explanations at the "what", "why", "how", or "where" level of the Marr hierarchy, and how-actually vs. how-possibly explanations. Throughout this thesis, I have aligned myself with the pragmatic view that epistemic transparency is not an inherent property of models but is contingent on what the user is trying to understand. Thus, we cannot make a blanket statement about the explanatory power of NNs because it depends on the user's prior knowledge, the type of explanation sought, and how the model is being used.

If guided by specific questions, NN models afford a wide range of explanations at different levels of analysis. It is true that DL models can appear as opaque systems that are difficult – perhaps even impossible – for humans to fully grasp. NNs may not lend themselves to the same types of explanations as traditional computational models. However, as seen throughout this thesis, this does not mean they cannot be analyzed and leveraged for scientific inquiry. When guided by specific research questions, NNs can be examined at different levels of granularity to offer inductive-statistical, causal mechanistic, and unificationist explanations at the "what", "why", "how", or "where" levels of Marr's hierarchy. In line with arguments by NN proponents, these models provide a higher level of experimental access than human brains. As shown in chapter 5, they also allow for the simulation of scenarios that are impossible to replicate in reality. This makes NNs not only useful for inspiring research on biological cognition but also intrinsically interesting to cognitive scientists. After all, cognitive science is the study of how agents, including artificial ones, perform tasks.

NNs can be used to explain external phenomena without a complete understanding of the model itself. Besides trying to answer questions about NNs themselves, researchers can also leverage these models to gain insights into the target phenomenon of biological cognition. NNs may not be able to provide detailed deductive-nomological "how-actually" explanations of the brain, as they often lack the nuanced representations and dynamics of biological neural networks. However, they can still serve as useful "how-possibly" explanations if they are sufficiently connected to the real-world system of interest. Importantly, this approach does not require a complete understanding of the NN model itself. The EREs that drive a target phenomenon can often be found at higher levels of abstraction. For example, chapter 5 provided a how-possibly explanation of statistical regularities in letter shapes, and chapter 7 demonstrated how a lack of exposure to symbolic numbers could decrease performance on non-symbolic comparisons. In both cases, the most relevant ERE was the dataset.

Explanation is not the only way toward understanding. Finally, (NN) models like the one proposed in chapter 8 can provide an alternative path to understanding that goes beyond traditional forms of explanation. They can serve as "mediators" between theory and data. By interactively exploring the behavior of NNs, users and designers can build up tacit knowledge and intuitions about a model or a problem domain. This hands-on experience with a model's inputs, outputs, and internal representations can yield insights that are difficult to capture through verbal theorizing or mathematical analysis alone. In this view, models are not just tools for testing theories but active participants in the process of understanding. By engaging with the model's behavior, users can develop new questions, refine their conceptual frameworks, and gain a feel for the space of possible explanations.

9.2 Epistemic functions of neural network models

The case studies presented in this thesis illustrate the diverse epistemic functions that NNs can serve in the cognitive sciences (see Figure 9.1). They can be leveraged for prediction, exploration, and explanation - often in complementary ways [CK19].

Prediction NNs designed with a focus on predictive power can be helpful in two key ways. First, they can act as tools to reach practical aims. In these applications, the model's output is the primary epistemic goal rather than providing an explanation of the underlying mechanisms. For example, a NN trained on real-time sensor data could detect or predict when a patient with schizophrenia is likely to experience a psychotic episode, allowing clinicians to intervene proactively. In such cases, the model's predictive performance is the main criterion for success, not its interpretability or alignment with theory.

Second, predictive benchmarking, where NNs are evaluated on standardized datasets like gSCAN or BIB, can serve as a starting point for further inquiry and explanation. These comparative studies can point to promising candidate models that warrant more in-depth analysis and guide model selection for subsequent explanatory work. Additionally, analyzing the factors that contribute to a model's predictive success, such as architectural choices or training regimes, can provide insights into the underlying principles governing the target phenomenon. So, while predictive benchmarking does not itself constitute an explanation, it can serve as a stepping stone by identifying models worthy of further investigation and by suggesting hypotheses about the factors driving predictive performance.

Exploration Additionally, NN models can serve an exploratory function, opening up new avenues for research. One key way they contribute to exploration is through proof-of-principle demonstrations. For example, the second case study showed that selective and joint attention mechanisms could improve compositional generalization by helping to isolate relevant inputs. The fourth case study showed that the Transformer attention mechanism could be helpful in learning semantic categories like social agents and goals from end-to-end training, and in capturing the relationships between these categories. Finally, the fifth case study demonstrated how phenomena like "implicit curricula" of early number skills, size and distance effects, and functional organization could emerge from the general learning mechanisms of NNs. While not definitive, such results can motivate and guide further inquiries.

NN models can also help assess the suitability of a cognitive phenomenon for computational modeling. E.g., the sixth case study on decision-making in medical ethics operationalized Beauchamp and Childress' prima facie principles into a NN. The model provided a tangible proposal of how these principles could be applied, sparking debate among commentators in the medical ethics community. By implementing the abstract principles in computational form, the model facilitated scrutiny of their empirical implications and suitability as a basis for moral reasoning. This demonstrates how models, including NNs, can serve as a way to communicate ideas, form the basis for discussion, and perhaps ultimately help concretize and re-evaluate theoretical constructs.

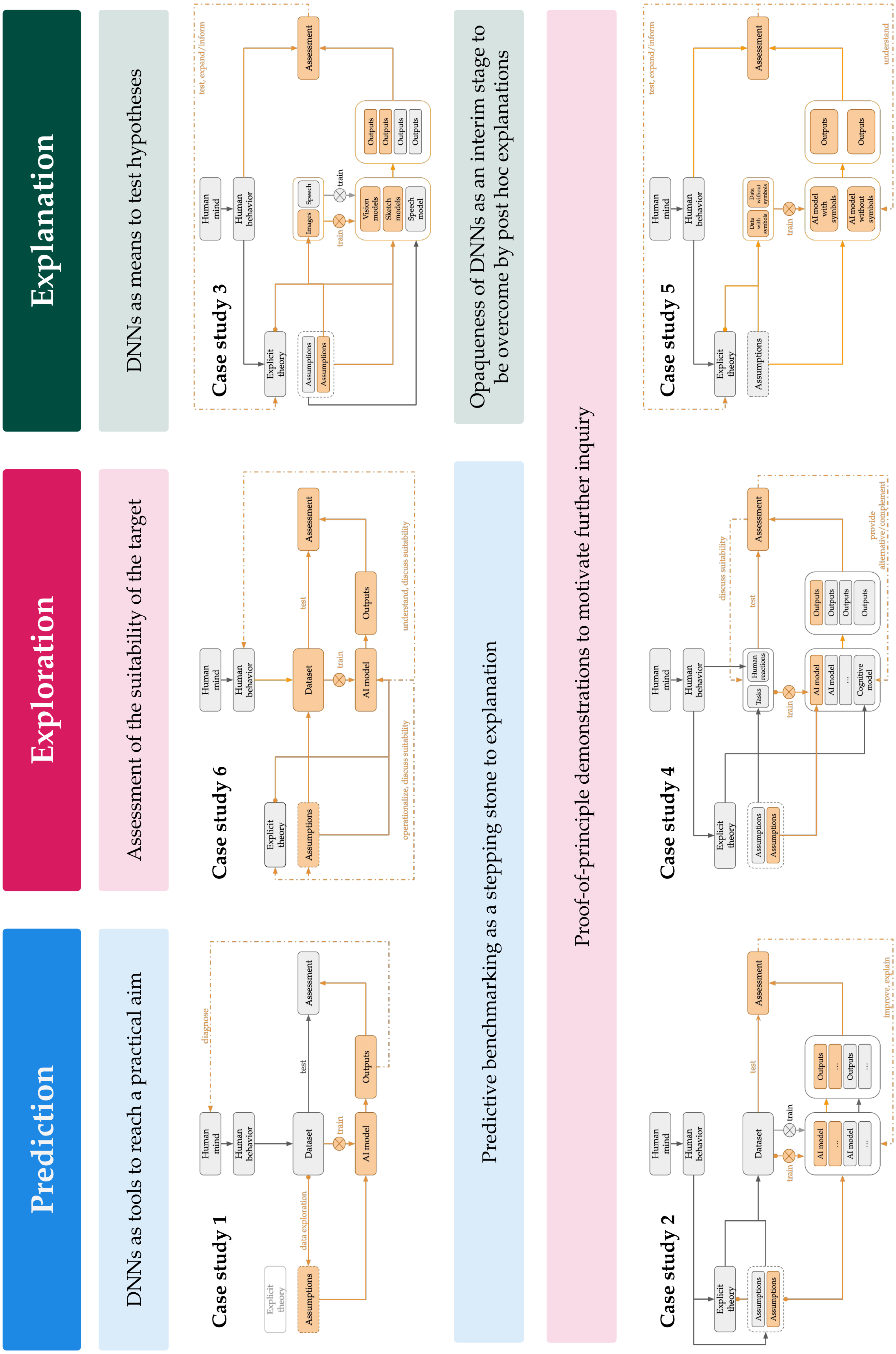


Figure 9.1: Overview of the main epistemic functions served by the models in the six case studies presented in this thesis. Loosely based on Cichy and Kaiser [CK19].

Explanation Finally, NNs can be leveraged to provide explanations in two distinct ways. First, they can serve as computational implementations of hypotheses about cognitive phenomena, enabling those hypotheses to be tested and refined. For example, in the third case study, I evolved glyphs for NNs exposed to different pre-training data and evaluated how well the resulting glyphs' geometric characteristics matched those of human writing systems. Analyzing the glyph shapes resulting under different training regimes provided support for the ecological hypothesis that writing systems reflect the visual statistics to which our perceptual system has adapted. Thus, NNs can serve as tools for operationalizing hypotheses in a way that enables them to be empirically evaluated.

Second, when the goal of a study requires a detailed understanding of a NN itself, there is a wide range of post-hoc explanatory techniques available. While NNs may initially appear as opaque black boxes, researchers can "look under the hood" to some considerable degree. For example, the fifth case study on numerical cognition used methods like the logit lens, specific error analyses, probing the representations encoded in different heads and layers, visualizing model embeddings, and illustrating an information flow to shed light on the model's inner workings. Thus, even though NNs are highly complex, it is possible to design experiments to investigate specific questions about what a model has learned using interpretability tools.

9.3 Conclusion

As Neil Gaiman notes in this chapter's opening quote, mirrors "appear to tell the truth, to reflect life back out at us." In much the same way, models are typically expected to accurately represent and reflect their target phenomena. They are tools that allow us to observe, measure, and understand what is happening around us – and inside our minds. However, Gaiman also points out that mirrors can "lie so convincingly" when set at the right angle. Similarly, models can distort or simplify reality to make it more comprehensible. They are inherently reductive - they cannot capture the full complexity of the world and must necessarily focus on certain aspects while omitting others in order to be useful.

Models, like "magic casements" can also "show you anything you can imagine and maybe a few things you can't." They allow us to explore hypothetical scenarios, test theories, and imagine alternative realities that may not yet exist. Just as mirrors can transport us to fantastical realms, models can open up new vistas of understanding and possibility. Ultimately, models are tools we use to "tell ourselves things we might not otherwise see." They are reflective surfaces that allow us to grapple with the world's complexities, uncover hidden truths, and envision new possibilities. In this and all the ways above, scientific models, including NNs, can be seen as mirrors - distorted, reductive, and yet profoundly illuminating.

List of Figures

List of Tables

Bibliography

[16] "Why Did Hospital Ethics Committees Emerge in the US?" In: *AMA Journal of Ethics* 18.5 (2016), pp. 546–553. ISSN: 2376-6980. DOI: 10.1001/journalofethics.2016.18.5.mhst1-1605.

[AA18] M. Anderson and S. L. Anderson. "GenEth: A General Ethical Dilemma Analyzer." In: *Paladyn, Journal of Behavioral Robotics* 9.1 (2018), pp. 337–357. ISSN: 2081-4836. DOI: 10.1515/pjbr-2018-0024.

[AAA06] M. Anderson, S. Anderson, and C. Armen. "An Approach to Computing Ethics." In: *IEEE Intelligent Systems* 21.4 (2006), pp. 56–63. ISSN: 1541-1672. DOI: 10.1109/MIS.2006.64.

[AAA20] E. Akyürek, A. F. Akyürek, and J. Andreas. "Learning to Recombine and Resample Data For Compositional Generalization." In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. 2020.

[Abr+16] E. Abrahamse, S. Braem, W. Notebaert, and T. Verguts. "Grounding Cognitive Control in Associative Learning." In: *Psychological Bulletin* 142.7 (2016), pp. 693–728. DOI: 10.1037/bul0000047.

[ACB02] K. Ahmad, M. Casey, and T. Bale. "Connectionist Simulation of Quantification Skills." In: *Connection Science* 14.3 (2002), pp. 165–201. DOI: 10.1080/09540090208559326.

[AG19] M. Andrus and T. K. Gilbert. "Towards a Just Theory of Measurement: A Principled Social Measurement Assurance Program for Machine Learning." In: *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019, pp. 445–451. ISBN: 978-1-4503-6324-2. DOI: 10.1145/3306618.3314275.

[Agr+12] C. Agrillo, L. Piffer, A. Bisazza, and B. Butterworth. "Evidence for Two Numerical Systems That Are Similar in Humans and Guppies." In: *PloS one* 7.2 (2012), e31923. DOI: 10.1371/journal.pone.0031923.

[Alv] R. Alvarado. "Explaining Epistemic Opacity." In: *The Science and Art of Simulation II*. Ed. by M. M. Resch, A. Kaminski, and P. Gehring. Springer International Publishing.

[Alv23] R. Alvarado. "AI as an Epistemic Technology." In: *Science and Engineering Ethics* 29.5 (2023). ISSN: 1353-3452. DOI: 10.1007/s11948-023-00451-3.

[And+16a] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. "Learning to Compose Neural Networks for Question Answering." In: *Proceedings of the 15th Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technology (NAACL-HTL)*. 2016, pp. 1545–1554. DOI: 10.18653/v1/n16-1181.

[And+16b] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. "Neural Module Networks." In: *Proceedings of the 29th Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 39–48. DOI: 10.1109/CVPR.2016.12.

[And+19] J. Andreas, M. Baroni, A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, J. Devlin, A. Fyshe, L. Wehbe, et al. "Measuring Compositionality in Representation Learning." In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. 2019, pp. 2227–2237.

[And08] C. Anderson. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." In: *WIRED magazine* 16.7 (2008).

[And20] J. Andreas. "Good-Enough Compositional Data Augmentation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2020, pp. 7556–7566. DOI: `10.18653/v1/2020.acl-main.676`.

[And83] J. R. Anderson. *The Architecture of Cognition*. Lawrence Erlbaum Associates, Inc, 1983. DOI: `10.4324/9781315799438`.

[App72] S. Appelle. "Perception and Discrimination as a Function of Stimulus Orientation: The "Oblique Effect" in Man and Animals." In: *Psychological Bulletin* 78.4 (1972), pp. 266–278. DOI: `10.1037/h0033117`.

[AS89] T. F. Ackerman and C. Strong. *A Casebook of Medical Ethics*. Oxford University Press, 1989. ISBN: 978-0-19-503916-0.

[AS94] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules in Large Databases." In: *Proceedings of the 20th International Conference on Very Large Databases (VLDB)*. 1994, pp. 487–499. ISBN: 1558601538.

[Avr+23] K. Avramidis, K. Adsul, D. Bose, and S. Narayanan. "SP Grand Challenge 2023 – E-Prevention: Sleep Behavior as an Indicator of Relapses in Psychotic Patients." In: *Proceedings of the 48th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2023, pp. 1–2. DOI: `10.1109/ICASSP49357.2023.10096044`.

[Awa+18] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. "The Moral Machine Experiment." In: *Nature* 563.7729 (2018), pp. 59–64. ISSN: 0028-0836. DOI: `10.1038/s41586-018-0637-6`.

[Baa+19] J. Baan, J. Leible, M. Nikolaus, D. Rau, D. Ulmer, T. Baumgärtner, D. Hupkes, and E. Bruni. "On the Realization of Compositionality in Neural Networks." In: *Proceedings of the 2nd Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP) @ ACL*. 2019, pp. 127–137. DOI: `10.18653/v1/W19-4814`.

[Bah+18] D. Bahdanau, S. Murty, M. Noukhovitch, T. H. Nguyen, H. de Vries, and A. Courville. "Systematic Generalization: What Is Required and Can It Be Learned?" In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 2018.

[Ban23] S. Banerjee. *Animal Image Dataset*. 2023. URL: `https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals/data` (visited on 01/16/2024).

[Bar20] M. Baroni. "Linguistic Generalization and Compositionality in Modern Artificial Neural Networks." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1791 (2020). ISSN: 0962-8436. DOI: `10.1098/rstb.2019.0307`.

[BC13] T. L. Beauchamp and J. F. Childress. *Principles of Biomedical Ethics*. 7th ed. Oxford University Press, 2013. ISBN: 978-0-19-992458-5.

[BDL13] T. Bergmann, R. Dale, and G. Lupyan. "The Impact of Communicative Constraints on the Emergence of a Graphical Communication System." In: *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (CogSci)*. 2013.

[Bee21] R. D. Beer. “Some Historical Context for Minimal Cognition.” In: *Adaptive Behavior* 29.1 (2021), pp. 89–92. ISSN: 1059-7123. DOI: 10.1177/1059712320931595.

[Bee96] R. D. Beer. “Toward the Evolution of Dynamical Neural Networks for Minimally Cognitive Behavior.” In: *From Animals to Animats 4: Proceedings of the 4th International Conference on Simulation of Adaptive Behavior*. The MIT Press, 1996. ISBN: 9780262291316. DOI: 10.7551/mitpress/3118.003.0051.

[BFG22] N. Biller-Andorno, A. Ferrario, and S. Gloeckler. “In Search of a Mission: Artificial Intelligence in Clinical Ethics.” In: *The American Journal of Bioethics* 22.7 (2022), pp. 23–25. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075055.

[BG74] P. B. Buckley and C. B. Gillman. “Comparisons of Digits and Dot Patterns.” In: *Journal of Experimental Psychology* 103.6 (1974), pp. 1131–1136. DOI: 10.1037/h0037361.

[Bhu+20] A. K. Bhunia, A. Das, U. R. Muhammad, Y. Yang, T. M. Hospedales, T. Xiang, Y. Gryaditskaya, and Y. Song. “Pixelor: A Competitive Sketching AI Agent. So You Think You Can Sketch?” In: *ACM Transactions on Graphics (TOG)* 39.6 (2020), 166:1–166:15. DOI: 10.1145/3414685.3417840.

[BM23] N. Brancazio and R. Meyer. “Minimal Model Explanations of Cognition.” In: *European Journal for Philosophy of Science* 13.41 (2023). ISSN: 1879-4912. DOI: 10.1007/s13194-023-00547-4.

[BMM19] D. G. Barrett, A. S. Morcos, and J. H. Macke. “Analyzing Biological and Artificial Neural Networks: Challenges With Opportunities for Synergy?” In: *Current Opinion in Neurobiology* 55 (2019), pp. 55–64. ISSN: 0959-4388. DOI: 10.1016/j.conb.2019.01.007.

[BP22] A. Barwise and B. Pickering. “The AI Needed for Ethical Decision Making Does Not Exist.” In: *The American Journal of Bioethics* 22.7 (2022), pp. 46–49. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075052.

[Bra73] C. J. Brainerd. “Mathematical and Behavioral Foundations of Number.” In: *The Journal of General Psychology* 88.2 (1973), pp. 221–281. ISSN: 0022-1309. DOI: 10.1080/00221309.1973.9920732.

[BST09] C. L. Baker, R. Saxe, and J. B. Tenenbaum. “Action Understanding as Inverse Planning.” In: *Cognition* 113.3 (2009), pp. 329–349. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2009.07.005.

[BT98] E. M. Brannon and H. S. Terrace. “Ordering of the Numerosities 1 to 9 by Monkeys.” In: *Science* 282.5389 (1998), pp. 746–749. DOI: 0.1126/science.282.5389.746.

[BTA10] D. C. Burr, M. Turi, and G. Anobile. “Subitizing but Not Estimation of Numerosity Requires Attentional Resources.” In: *Journal of Vision* 10.6 (2010). DOI: 10.1167/10.6.20.

[BTT15] C. Boylan, J. C. Trueswell, and S. L. Thompson-Schill. “Compositionality and the Angular Gyrus: A Multi-Voxel Similarity Analysis of the Semantic Composition of Nouns and Verbs.” In: *Neuropsychologia* 78 (2015), pp. 130–141. ISSN: 0028-3932. DOI: 10.1016/j.neuropsychologia.2015.10.007.

[Bun19] Bundesärztekammer. “Hinweise und Empfehlungen zum Umgang mit Vorsorgevollmachten und Patientenverfügungen im ärztlichen Alltag.” In: *Jahrbuch für Wissenschaft und Ethik* 24.1 (2019), pp. 335–354.

[Bur16] J. Burrell. “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms.” In: *Big Data & Society* 3.1 (2016). ISSN: 2053-9517. DOI: 10.1177/2053951715622512.

[BW15] R. D. Beer and P. L. Williams. “Information Processing and Dynamics in Minimally Cognitive Agents.” In: *Cognitive Science* 39.1 (2015), pp. 1–38. DOI: https://doi.org/10.1111/cogs.12142.

[Cad+19] S. A. Cadena, G. H. Denfield, E. Y. Walker, L. A. Gatys, A. S. Tolias, M. Bethge, and A. S. Ecker. “Deep Convolutional Models Improve Predictions of Macaque V1 Responses to Natural Images.” In: *PLOS Computational Biology* 15.4 (2019). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006897.

[Cal+23] S. Calcagno, R. Mineo, D. Giordano, and C. Spampinato. “Ensemble and Personalized Transformer Models for Subject Identification and Relapse Detection in E-Prevention Challenge.” In: *Proceedings of the 48th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–2. ISBN: 978-1-72816-327-7. DOI: 10.1109/ICASSP49357.2023.10095438.

[Cam03] V. Camos. “Counting Strategies From 5 Years to Adulthood: Adaptation to Structural Features.” In: *European Journal of Psychology of Education* 18 (2003), pp. 251–265. DOI: 10.1007/BF03173247.

[Cao+19] N. Cao, X. Yan, Y. Shi, and C. Chen. “AI-Sketcher: A Deep Generative Model for Producing High-Quality Sketches.” In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. Vol. 33. 1. 2019, pp. 2564–2571. DOI: 10.1609/aaai.v33i01.33012564.

[Car00] S. Carey. “The Origin of Concepts.” In: *Journal of Cognition and Development* 1.1 (2000), pp. 37–41. DOI: 10.1093/acprof:oso/9780195367638.001.0001.

[Car11] S. Carey. “Précis of 'The Origin of Concepts'.” In: *Behavioral and Brain Sciences* 34.3 (2011), pp. 113–124. DOI: 10.1017/S0140525X10000919.

[CD22] A. Coin and V. Dubljević. “Using Algorithms to Make Ethical Judgements: METHAD vs. the ADC Model.” In: *The American Journal of Bioethics* 22.7 (2022), pp. 41–43. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075967.

[CGK22] C. Caucheteux, A. Gramfort, and J.-R. King. “Deep Language Algorithms Predict Semantic Comprehension From Brain Activity.” In: *Scientific Reports* 12.1 (2022). ISSN: 2045-2322. DOI: 10.1038/s41598-022-20460-9.

[Cha+06] M. A. Changizi, Q. Zhang, H. Ye, and S. Shimojo. “The Structures of Letters and Symbols Throughout Human History Are Selected to Match Those Found in Objects in Natural Scenes.” In: *The American Naturalist* 167.5 (2006), E117–E139. DOI: 10.1086/502806.

[Cha+18] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov. “Gated-Attention Architectures for Task-Oriented Language Grounding.” In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*. 2018, pp. 2819–2826. ISBN: 978-1-57735-800-8.

[Cha+20] R. Chaabouni, E. Kharitonov, D. Bouchacourt, E. Dupoux, and M. Baroni. “Compositionality and Generalization in Emergent Languages.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2020, pp. 4427–4442. DOI: 10.18653/v1/2020.acl-main.407.

[Cha22a] T. Chambers. “An All-Too-Human Enterprise.” In: *The American Journal of Bioethics* 22.7 (2022), pp. 33–35. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075969.

[Cha22b] D. Char. "Important Design Questions for Algorithmic Ethics Consultation." In: *The American Journal of Bioethics* 22.7 (2022), pp. 38–40. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075054.

[Che+18] S. Chen, Z. Zhou, M. Fang, and J. McClelland. "Can Generic Neural Networks Estimate Numerosity Like Humans?" In: *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*. 2018.

[Che+19] M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, T. H. Nguyen, and Y. Bengio. "BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning." In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019.

[Cic+16] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva. "Comparison of Deep Neural Networks to Spatio-Temporal Cortical Dynamics of Human Visual Object Recognition Reveals Hierarchical Correspondence." In: *Scientific Reports* 6.1 (2016). ISSN: 2045-2322. DOI: 10.1038/srep27755.

[CK19] R. M. Cichy and D. Kaiser. "Deep Neural Networks as Scientific Models." In: *Trends in Cognitive Sciences* 23.4 (2019), pp. 305–317. ISSN: 1364-6613. DOI: 10.1016/j.tics.2019.01.009.

[Cla+19] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. "What Does BERT Look at? An Analysis of BERT's Attention." In: *Proceedings of the 2nd Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP) @ ACL*. 2019, pp. 276–286. DOI: 10.18653/v1/W19-4828.

[CLL18] Y. Chen, Y.-K. Lai, and Y.-J. Liu. "CartoonGAN: Generative Adversarial Networks for Photo Cartoonization." In: *Proceedings of the 31st Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 9465–9474. DOI: 10.1109/CVPR.2018.00986.

[CLS18] S. Chung, D. D. Lee, and H. Sompolinsky. "Classification and Geometry of General Perceptual Manifolds." In: *Physical Review X* 8.3 (2018). ISSN: 2160-3308. DOI: 10.1103/PhysRevX.8.031003.

[Cop+98] D. M. Coppola, H. R. Purves, A. N. McCoy, and D. Purves. "The Distribution of Oriented Contours in the Real World." In: *Proceedings of the National Academy of Sciences* 95.7 (1998), pp. 4002–4006. DOI: 10.1073/PNAS.95.7.4002.

[CPM21] D. S. Chaplot, D. Pathak, and J. Malik. "Differentiable Spatial Planning Using Transformers." In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021, pp. 1484–1495.

[CS06] S. Carey and B. W. Sarnecka. "The Development of Human Conceptual Representations: A Case Study." In: *Processes of Change in Brain and Cognitive Development: Attention and Performance XXI* (2006), pp. 473–496. DOI: 10.1093/oso/9780198568742.003.0020.

[CSS21] C. Creatore, S. Sabathiel, and T. Solstad. "Learning Exact Enumeration and Approximate Estimation in Deep Neural Network Models." In: *Cognition* 215 (2021). ISSN: 0010-0277. DOI: 10.1016/j.cognition.2021.104815.

[CT22] S. Cognolato and A. Testolin. "Transformers Discover an Elementary Calculation System Exploiting Local Attention and Grid-Like Problem Representation." In: *Proceedings of the 32nd International Joint Conference on Neural Networks (IJCNN)*. 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9892619.

[CV10] Q. Chen and T. Verguts. "Beyond the Mental Number Line: A Neural Network Model of Number–Space Interactions." In: *Cognitive Psychology* 60.3 (2010), pp. 218–240. ISSN: 0010-0285. DOI: `10.1016/j.cogpsych.2010.01.001`.

[CWS21] W. Chen, F. Wang, and H. Sun. "S2TNet: Spatio-Temporal Transformer Networks for Trajectory Prediction in Autonomous Driving." In: *Proceedings of the 13th Asian Conference on Machine Learning*. 2021, pp. 454–469.

[Dad+09] M. Dadda, L. Piffer, C. Agrillo, and A. Bisazza. "Spontaneous Number Representation in Mosquitofish." In: *Cognition* 112.2 (2009), pp. 343–348. ISSN: 0010-0277. DOI: `10.1016/j.cognition.2009.05.009`.

[DBG93] S. Dehaene, S. Bossini, and P. Giraux. "The Mental Representation of Parity and Number Magnitude." In: *Journal of Experimental Psychology: General* 122.3 (1993), pp. 371–396. DOI: `10.1037/0096-3445.122.3.371`.

[DC93] S. Dehaene and J.-P. Changeux. "Development of Elementary Numerical Abilities: A Neuronal Model." In: *Journal of Cognitive Neuroscience* 5.4 (1993), pp. 390–407. DOI: `10.1162/jocn.1993.5.4.390`.

[DDC98] S. Dehaene, G. Dehaene-Lambertz, and L. Cohen. "Abstract Representations of Numbers in the Animal and Human Brain." In: *Trends in Neurosciences* 21.8 (1998), pp. 355–361. DOI: `10.1016/S0166-2236(98)01263-6`.

[De +14] V. M. De La Cruz, A. Di Nuovo, S. Di Nuovo, and A. Cangelosi. "Making Fingers and Words Count in a Cognitive Robot." In: *Frontiers in Behavioral Neuroscience* 8 (2014). ISSN: 1662-5153. DOI: `10.3389/fnbeh.2014.00013`.

[DEB12] K. Davidson, K. Eng, and D. Barner. "Does Learning to Count Involve a Semantic Induction?" In: *Cognition* 123.1 (2012), pp. 162–173. DOI: `10.1016/j.cognition.2011.12.013`.

[Deh+06] S. Dehaene, V. Izard, P. Pica, and E. Spelke. "Core Knowledge of Geometry in an Amazonian Indigene Group." In: *Science* 311.5759 (2006), pp. 381–384. DOI: `10.1126/science.1121739`.

[Deh92] S. Dehaene. "Varieties of Numerical Abilities." In: *Cognition* 44.1-2 (1992), pp. 1–42. DOI: `10.1016/0010-0277(92)90049-N`.

[Der+19] T. Deruyttere, S. Vandenhende, D. Grujicic, L. V. Gool, and M. Moens. "Talk2Car: Taking Control of Your Self-Driving Car." In: *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 2088–2098. DOI: `10.18653/V1/D19-1215`.

[Des+22] J. Desai, D. Watson, V. Wang, M. Taddeo, and L. Floridi. "The Epistemological Foundations of Data Science: A Critical Review." In: *Synthese* 200.6 (2022). ISSN: 1573-0964. DOI: `10.1007/s11229-022-03933-2`.

[DES22] J. Demaree-Cotton, B. D. Earp, and J. Savulescu. "How to Use AI Ethically for Ethical Decision-Making." In: *The American Journal of Bioethics* 22.7 (2022), pp. 1–3. ISSN: 1526-5161. DOI: `10.1080/15265161.2022.2075968`.

[DFR22] J. P. DeMarco, P. J. Ford, and S. L. Rose. "Implicit Fuzzy Specifications, Inferior to Explicit Balancing." In: *The American Journal of Bioethics* 22.7 (2022), pp. 21–23. ISSN: 1526-5161. DOI: `10.1080/15265161.2022.2075970`.

[DHP10] D. Dickenson, R. Huxtable, and M. Parker. *The Cambridge Medical Ethics Workbook*. 2nd ed. Cambridge University Press, 2010. ISBN: 978-0-521-73470-7. DOI: 10.1017/CBO9780511910098.

[Di +14] A. Di Nuovo, V. M. De La Cruz, A. Cangelosi, and S. Di Nuovo. "The iCub Learns Numbers: An Embodied Cognition Study." In: *Proceedings of the 24th International Joint Conference on Neural Networks (IJCNN)*. 2014, pp. 692–699. ISBN: 978-1-4799-1484-5. DOI: 10.1109/IJCNN.2014.6889795.

[Di 17] A. Di Nuovo. "An Embodied Model for Handwritten Digits Recognition in a Cognitive Robot." In: *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*. 2017, pp. 1–6. DOI: 10.1109/SSCI.2017.8285274.

[Di 18] A. Di Nuovo. "Long-Short Term Memory Networks for Modelling Embodied Mathematical Cognition in Robots." In: *Proceedings of the 28th International Joint Conference on Neural Networks (IJCNN)*. 2018, pp. 1–7. DOI: 10.1109/IJCNN.2018.8489140.

[Dié15] A. Diéguez. "Scientific Understanding and the Explanatory use of False Models." In: *The Future of Scientific Practice*. Ed. by M. Bertolaso. 1st ed. Routledge, 2015, pp. 161–178. ISBN: 978-1-315-65369-3.

[Dig20] V. Dignum. "Responsibility and Artificial Intelligence." In: *The Oxford Handbook of Ethics of AI*. Ed. by M. D. Dubber, F. Pasquale, and S. Das. Oxford University Press, 2020, pp. 213–231. ISBN: 978-0-19-006739-7. DOI: 10.1093/oxfordhb/9780190067397.013.12.

[DK17] F. Doshi-Velez and B. Kim. "A Roadmap for a Rigorous Science of Interpretability." In: *CoRR* abs/1702.08608 (2017). DOI: 10.48550/arXiv.1702.08608.

[DM19] A. Di Nuovo and J. L. McClelland. "Developing the Knowledge of Number Digits in a Child-Like Robot." In: *Nature Machine Intelligence* 1.12 (2019), pp. 594–605. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0123-3.

[DMP16] R. De Diego-Balaguer, A. Martinez-Alvarez, and F. Pons. "Temporal Attention as a Scaffold for Language Development." In: *Frontiers in Psychology* 7 (2016). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2016.00044.

[DP73] D. H. Douglas and T. K. Peucker. "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature." In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 10.2 (1973), pp. 112–122. DOI: 10.3138/FM57-6770-U75U-7727.

[DS23] G. Delanerolle and J. Q. Shi. *Bayesian Statistics and Machine Learning; Why Should We Pay More Attention to This Relationship?* 2023. URL: https://yoursay.plos.org/2023/08/bayesian-statistics-and-machine-learning-why-should-we-pay-more-attention-to-this-relationship/ (visited on 04/29/2024).

[DT05] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection." In: *Proceedings of the 18th Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2005, pp. 886–893. DOI: 10.1109/CVPR.2005.177.

[Dut+20] E. Dutkiewicz, G. Russo, S. Lee, and C. Bentz. "SignBase, a Collection of Geometric Signs on Mobile Objects in the Paleolithic." In: *Scientific Data* 7.1 (2020), p. 364. DOI: 10.1038/s41597-020-00704-x.

[DVC15] A. Di Nuovo, M. Vivian, and A. Cangelosi. "A Deep Learning Neural Network for Number Cognition: A Bi-cultural Study With the iCub." In: *Proceedings of the 11th Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. 2015, pp. 320–325. DOI: 10.1109/DEVLRN.2015.7346165.

[DWC21] Z. Dulberg, T. Webb, and J. Cohen. "Modelling the Development of Counting With Memory-Augmented Neural Networks." In: *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society (CogSci)*. 2021.

[DYZ21a] H. Du, X. Yu, and L. Zheng. "VTNet: Visual Transformer Network for Object Goal Navigation." In: *Proceedings of the 9th International Conference on Learning Representations, (ICLR)*. 2021.

[DYZ21b] H. Du, X. Yu, and L. Zheng. "VTNet: Visual Transformer Network for Object Goal Navigation." In: *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. 2021.

[Eft+23] N. Efthymiou, G. Retsinas, P. P. Filntisis, C. Garoufis, A. Zlatintsi, E. Kalisperakis, V. Garyfalli, T. Karantinos, M. Lazaridi, N. Smyrnis, and P. Maragos. "From Digital Phenotype Identification To Detection Of Psychotic Relapses." In: *Proceedings of the 11th International Conference on Healthcare Informatics (ICHI)*. 2023, pp. 276–284. DOI: 10.1109/ICHI57859.2023.00045.

[EG18] R. Evans and E. Grefenstette. "Learning Explanatory Rules From Noisy Data." In: *Journal of Artificial Intelligence Research* 61.1 (2018), pp. 1–64. ISSN: 1076-9757.

[Eic+17] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. "Seeing It All: Convolutional Network Layers Map the Function of the Human Visual System." In: *NeuroImage* 152 (2017), pp. 184–194. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2016.10.001.

[Eli21] E. Elizalde. "The Theory of General Relativity and Its Main Solutions." In: *The True Story of Modern Cosmology*. Springer International Publishing, 2021, pp. 81–136. ISBN: 978-3-030-80653-8. DOI: 10.1007/978-3-030-80654-5_4.

[Ell+16] K. C. Elliott, K. S. Cheruvelil, G. M. Montgomery, and P. A. Soranno. "Conceptions of Good Science in Our Data-Rich World." In: *BioScience* 66.10 (2016), pp. 880–889. ISSN: 0006-3568. DOI: 10.1093/biosci/biw115.

[Fan+20] J. E. Fan, R. D. Hawkins, M. Wu, and N. D. Goodman. "Pragmatic Inference and Visual Abstraction Enable Contextual Flexibility during Visual Communication." In: *Computational Brain & Behavior* 3 (2020), pp. 86–101. DOI: 10.1007/s42113-019-00058-7.

[Fay+10] N. Fay, S. Garrod, L. Roberts, and N. Swoboda. "The Interactive Evolution of Human Communication Systems." In: *Cognitive Science* 34.3 (2010), pp. 351–386. DOI: 10.1111/J.1551-6709.2009.01090.X.

[Fay+18] N. Fay, B. Walker, N. Swoboda, and S. Garrod. "How to Create Shared Symbols." In: *Cognitive Science* 42 (2018), pp. 241–269. DOI: 10.1111/COGS.12600.

[FDH19] J. E. Fan, M. Dinculescu, and D. Ha. "Collabdraw: An Environment for Collaborative Sketching with an Artificial Agent." In: *Proceedings of the 12th ACM SIGCHI Conference on Creativity and Cognition (C&C)*. 2019, pp. 556–561. DOI: 10.1145/3325480.3326578.

[FDS04a] L. Feigenson, S. Dehaene, and E. Spelke. "Core Systems of Number." In: *Trends in Cognitive Sciences* 8.7 (2004), pp. 307–314. ISSN: 1364-6613. DOI: 10.1016/j.tics.2004.05.002.

[FDS04b] L. Feigenson, S. Dehaene, and E. Spelke. “Core Systems of Number.” In: *Trends in Cognitive Sciences* 8.7 (2004), pp. 307–314. DOI: 10.1016/j.tics.2004.05.002.

[Fel+19] G. Felix, G. Nápoles, R. Falcon, W. Froelich, K. Vanhoof, and R. Bello. “A Review on Methods and Software for Fuzzy Cognitive Maps.” In: *Artificial Intelligence Review* 52.3 (2019), pp. 1707–1737. ISSN: 0269-2821. DOI: 10.1007/s10462-017-9575-1.

[Fen67] H. Fenichel Pitkin. *The Concept of Representation*. 1967. DOI: 10.1525/9780520340503.

[Fil+24] P. Filntisis, N. Efthymiou, G. Retsinas, A. Zlatintsi, C. Garoufis, T. Sounapoglou, P. Tsanakas, N. Smyrnis, and P. Maragos. “The 2nd e-Prevention challenge: Psychotic and Non-Psychotic Relapse Detection using Wearable-Based Digital Phenotyping.” In: *Proceedings of the 49th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024.

[Fin+21] S. Finke, F. Kemény, F. J. Clayton, C. Banfi, A. F. Steiner, C. M. Perchtold-Stefan, I. Papousek, S. M. Göbel, and K. Landerl. “Cross-Format Integration of Auditory Number Words and Visual-Arabic Digits: An ERP Study.” In: *Frontiers in Psychology* 12 (2021), p. 765709. DOI: 10.3389/fpsyg.2021.765709.

[FM01] J. M. Freeman and K. McDonnell. *Tough Decisions: Cases in Medical Ethics*. 2nd ed. Oxford University Press, 2001. ISBN: 978-0-19-509041-3.

[Fra80] B. C. V. Fraassen. *The Scientific Image*. 1st ed. Oxford University PressOxford, 1980. ISBN: 978-0-19-824427-1. DOI: 10.1093/0198244274.001.0001.

[FRB82] K. C. Fuson, J. Richards, and D. J. Briars. “The Acquisition and Elaboration of the Number Word Sequence.” In: *Children's Logical and Mathematical Cognition: Progress in Cognitive Development Research*. Ed. by C. J. Brainerd. Springer New York, 1982, pp. 33–92. ISBN: 978-1-4613-9466-2. DOI: 10.1007/978-1-4613-9466-2_2.

[Fri15] M. Frické. “Big Data and Its Epistemology.” In: *Journal of the Association for Information Science and Technology* 66.4 (2015), pp. 651–661. ISSN: 2330-1635. DOI: 10.1002/asi.23212.

[Fuk+22] R. Fukushima, K. Ota, A. Kanezaki, Y. Sasaki, and Y. Yoshiyasu. “Object Memory Transformer for Object Goal Navigation.” In: *Proceedings of the 39th International Conference on Robotics and Automation (ICRA)*. 2022, pp. 11288–11294. DOI: 10.1109/ICRA46639.2022.9812027.

[Gag68] R. M. Gagne. “Presidential Address of Division 15 – Learning Hierarchies.” In: *Educational Psychologist* 6.1 (1968), pp. 1–9. DOI: 10.1080/00461526809528968.

[Gal05] B. Galantucci. “An Experimental Study of the Emergence of Human Communication Systems.” In: *Cognitive Science* 29.5 (2005), pp. 737–767. DOI: 10.1207/S15516709COG0000_34.

[Gan+21] K. Gandhi, G. Stojnic, B. M. Lake, and M. R. Dillon. “Baby Intuitions Benchmark (BIB): Discerning the Goals, Preferences, and Actions of Others.” In: *Proceedings of the 35th Annual Conference of Neural Information Processing Systems (NeurIPS)* (2021), pp. 9963–9976.

[Gar+07] S. Garrod, N. Fay, J. Lee, J. Oberlander, and T. MacLeod. “Foundations of Representation: Where Might Graphical Symbol Systems Come From?” In: *Cognitive Science* 31.6 (2007), pp. 961–987. DOI: 10.1080/03640210701703659.

[Gar+20] M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, et al. "Evaluating Models' Local Decision Boundaries via Contrast Sets." In: *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 1307–1323. DOI: 10.18653/v1/2020.findings-emnlp.117.

[Gar+93] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. "DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1." In: *NASA STI/Recon Technical Report* 93 (1993).

[Gau+21] D. J. Gauthier, E. Bollt, A. Griffith, and W. A. S. Barbosa. "Next Generation Reservoir Computing." In: *Nature Communications* 12.1 (2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-25801-2.

[GB22a] A. Goyal and Y. Bengio. "Inductive Biases for Deep Learning of Higher-Level Cognition." In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 478.2266 (2022). ISSN: 1364-5021. DOI: 10.1098/rspa.2021.0068.

[GB22b] T. Gundersen and K. Bærøe. "Ethical Algorithmic Advice: Some Reasons to Pause and Think Twice." In: *The American Journal of Bioethics* 22.7 (2022), pp. 26–28. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075053.

[GC03] G. Gergely and G. Csibra. "Teleological Reasoning in Infancy: The Naïve Theory of Rational Action." In: *Trends in Cognitive Sciences* 7.7 (2003), pp. 287–292. ISSN: 1364-6613. DOI: 10.1016/S1364-6613(03)00128-1.

[Gei+20] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. "Shortcut Learning in Deep Neural Networks." In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673. DOI: 10.1038/s42256-020-00257-z.

[Gel16a] A. Gelfert. "Between Theory and Phenomena: What are Scientific Models?" In: *How to Do Science with Models*. Series Title: SpringerBriefs in Philosophy. Springer International Publishing, 2016, pp. 1–24. ISBN: 978-3-319-27954-1. DOI: 10.1007/978-3-319-27954-1_1.

[Gel16b] A. Gelfert. "Exploratory Uses of Scientific Models." In: *How to Do Science with Models*. Series Title: SpringerBriefs in Philosophy. Springer International Publishing, 2016, pp. 71–99. ISBN: 978-3-319-27952-7. DOI: 10.1007/978-3-319-27954-1_4.

[Gel16c] A. Gelfert. *How to Do Science with Models*. SpringerBriefs in Philosophy. Springer International Publishing, 2016. ISBN: 978-3-319-27954-1. DOI: 10.1007/978-3-319-27954-1.

[Gel16d] A. Gelfert. "Models as Mediators, Contributors, and Enablers of Scientific Knowledge." In: *How to Do Science with Models*. Series Title: SpringerBriefs in Philosophy. Springer International Publishing, 2016, pp. 101–129. ISBN: 978-3-319-27952-7. DOI: 10.1007/978-3-319-27954-1_5.

[Gel16e] A. Gelfert. "Scientific Representation and the Uses of Scientific Models." In: *How to Do Science with Models*. Series Title: SpringerBriefs in Philosophy. Springer International Publishing, 2016, pp. 25–42. ISBN: 978-3-319-27954-1. DOI: 10.1007/978-3-319-27954-1_2.

[Gel17] A. Gelfert. "The Ontology of Models." In: *Springer Handbook of Model-Based Science*. Ed. by L. Magnani and T. Bertolotti. Springer International Publishing, 2017, pp. 5–23. ISBN: 978-3-319-30526-4. DOI: 10.1007/978-3-319-30526-4_1.

[Gel19] A. Gelfert. "Probing Possibilities: Toy Models, Minimal Models, and Exploratory Models." In: *Model-Based Reasoning in Science and Technology*. Ed. by Á. Nepomuceno-Fernández, L. Magnani, F. J. Salguero-Lamillar, C. Barés-Gómez, and M. Fontaine. Vol. 49. Springer International Publishing, 2019, pp. 3–19. ISBN: 978-3-030-32722-4. DOI: 10.1007/978-3-030-32722-4_1.

[Gev+06] W. Gevers, T. Verguts, B. Reynvoet, B. Caessens, and W. Fias. "Numbers and Space: A Computational Model of the SNARC Effect." In: *Journal of Experimental Psychology: Human Perception and Performance* 32.1 (2006), p. 32. DOI: 10.1037/0096-1523.32.1.32.

[GG04] R. Gelman and C. R. Gallistel. "Language and the Origin of Numerical Concepts." In: *Science* 306.5695 (2004), pp. 441–443. DOI: 10.1126/science.110514.

[GG86] R. Gelman and C. R. Gallistel. *The Child's Understanding of Number*. Harvard University Press, 1986. ISBN: 9780674116375.

[GHM20] T. Gao, Q. Huang, and R. Mooney. "Systematic Generalization on gSCAN with Language Conditioned Embedding." In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 9th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*. 2020, pp. 491–503.

[Gie99] R. N. Giere. "Using Models to Represent Reality." In: *Model-Based Reasoning in Scientific Discovery*. Ed. by L. Magnani, N. J. Nersessian, and P. Thagard. Springer US, 1999, pp. 41–57. ISBN: 978-1-4615-4813-3. DOI: 10.1007/978-1-4615-4813-3_3.

[Gig20] G. Gigerenzer. "How to Explain Behavior?" In: *Topics in Cognitive Science* 12.4 (2020), pp. 1363–1381. ISSN: 1756-8757. DOI: 10.1111/tops.12480.

[Giu+21] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso. "Transformer Networks for Trajectory Forecasting." In: *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*. 2021, pp. 10335–10342. DOI: 10.1109/ICPR48806.2021.9412190.

[GLS11] A. R. Girshick, M. S. Landy, and E. P. Simoncelli. "Cardinal Rules: Visual Orientation Perception Reflects Knowledge of Environmental Statistics." In: *Nature Neuroscience* 14.7 (2011), pp. 926–932. DOI: 10.1038/NN.2831.

[GM92] M. A. Goodale and A. Milner. "Separate Visual Pathways for Perception and Action." In: *Trends in Neurosciences* 15.1 (1992), pp. 20–25. ISSN: 0166-2236. DOI: 10.1016/0166-2236(92)90344-8.

[GMS07] C. K. Gilmore, S. E. McCarthy, and E. S. Spelke. "Symbolic Arithmetic Knowledge Without Instruction." In: *Nature* 447.7144 (2007), pp. 589–591. DOI: 10.1038/nature05850.

[Goë+13] H. Goëau, A. Joly, P. Bonnet, V. Bakic, D. Barthélémy, N. Boujemaa, and J.-F. Molino. "The ImageCLEF Plant Identification Task 2013." In: *Proceedings of the 2nd ACM International Workshop on Multimedia Analysis for Ecological Data*. 2013, pp. 23–28. DOI: 10.1145/2509896.2509902.

[Gol22] M. Goldrick. "An Impoverished Epistemology Holds Back Cognitive Science Research." In: *Cognitive Science* 46.9 (2022), e13199. ISSN: 0364-0213. DOI: 10.1111/cogs.13199.

[Goo68] N. Goodman. *Languages of Art: An Approach to a Theory of Symbols*. Bobbs-Merrill, 1968.

[Gor04] P. Gordon. "Numerical Cognition Without Words: Evidence From Amazonia." In: *Science* 306.5695 (2004), pp. 496–499. DOI: 10.1126/science.1094492.

[GP17] J. Gouvea and C. Passmore. "'Models of' versus 'Models for': Toward an Agent-Based Conception of Modeling in the Science Classroom." In: *Science & Education* 26.1-2 (2017), pp. 49–63. ISSN: 0926-7220. DOI: 10.1007/s11191-017-9884-4.

[Gra+17] A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu. "Automated Curriculum Learning for Neural Networks." In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Vol. 70. 2017, pp. 1311–1320.

[Gri+16] "Enlightening Falsehoods: A Modal View of Scientific Understanding." In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Ed. by S. R. Grimm, C. Baumberger, S. Ammon, and S. Le Bihan. Routledge, Taylor & Francis Group, 2016, pp. 111–136. ISBN: 978-1-315-68611-0.

[GV15] U. Guclu and M. A. J. Van Gerven. "Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream." In: *Journal of Neuroscience* 35.27 (2015), pp. 10005–10014. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.5023-14.2015.

[Haa+20] A. Haahr, A. Norlyk, B. Martinsen, and P. Dreyer. "Nurses Experiences of Ethical Dilemmas: A Review." In: *Nursing Ethics* 27.1 (2020), pp. 258–272. ISSN: 0969-7330. DOI: 10.1177/0969733019832941.

[Hab+21] T. Habuza, A. N. Navaz, F. Hashim, F. Alnajjar, N. Zaki, M. A. Serhani, and Y. Statsenko. "AI Applications in Robotics, Diagnostic Image Analysis and Precision Medicine: Current Limitations, Future Trends, Guidelines on Cad Systems for Medicine." In: *Informatics in Medicine Unlocked* 24 (2021), p. 100596. ISSN: 2352-9148. DOI: 10.1016/j.imu.2021.100596.

[Han+10] N. Hansen, A. Auger, R. Ros, S. Finck, and P. Posík. "Comparing Results of 31 Algorithms From the Black-Box Optimization Benchmarking BBOB-2009." In: *Proceedings of the 11th Genetic and Evolutionary Computation Conference (GECCO) (Companion)*. ACM, 2010, pp. 1689–1696. DOI: 10.1145/1830761.1830790.

[Har+13] B. M. Harvey, B. P. Klein, N. Petridou, and S. O. Dumoulin. "Topographic Representation of Numerosity in the Human Parietal Cortex." In: *Science* 341.6150 (2013), pp. 1123–1126. DOI: 10.1126/science.1239052.

[Has+17] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. "Neuroscience-Inspired Artificial Intelligence." In: *Neuron* 95.2 (2017), pp. 245–258. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2017.06.011.

[Has14] M. Hast. "Collaborating With the 'More Capable' Self: Achieving Conceptual Change in Early Science Education Through Underlying Knowledge Structures." In: *ReflectEd, St Mary's Journal of Education* 3 (2014), pp. 18–25. ISSN: 2046-6978.

[Hau+03] M. D. Hauser, F. Tsao, P. Garcia, and E. S. Spelke. "Evolutionary Foundations of Number: Spontaneous Representation of Numerical Magnitudes by Cotton–Top Tamarins." In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270.1523 (2003), pp. 1441–1446. ISSN: 0962-8452. DOI: 10.1098/rspb.2003.2414.

[Haw+23] R. D. Hawkins, M. Sano, N. D. Goodman, and J. E. Fan. "Visual Resemblance and Interaction History Jointly Constrain Pictorial Meaning." In: *Nature Communications* 14.1 (2023), p. 2199. DOI: 10.1038/s41467-023-37737-w.

[HB20] C. Heinze-Deml and D. Bouchacourt. "Think Before You Act: A Simple Baseline for Compositional Generalization." In: *CoRR* abs/2009.13962 (2020). DOI: 10.48550/arXiv.2009.13962.

[HC17] B. Hommel and L. S. Colzato. "The Grand Challenge: Integrating Nomothetic and Ideographic Approaches to Human Cognition." In: *Frontiers in Psychology* 8 (2017). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.00100.

[HCH00] M. D. Hauser, S. Carey, and L. B. Hauser. "Spontaneous Number Representation in Semi–Free–Ranging Rhesus Monkeys." In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267.1445 (2000), pp. 829–833. DOI: 10.1098/rspb.2000.1078.

[HD22a] A. Hein and K. Diepold. "A Minimal Model for Compositional Generalization on gSCAN." In: *Proceedings of the 5th Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP) @ EMNLP*. 2022, pp. 1–15. DOI: 10.18653/V1/2022.BLACKBOXNLP-1.1.

[HD22b] A. Hein and K. Diepold. "Comparing Intuitions about Agents' Goals, Preferences and Actions in Human Infants and Video Transformers." In: *Shared Visual Representations in Human and Machine Intelligence Workshop (SVRHM) @ NeurIPS*. 2022.

[HD22c] A. Hein and K. Diepold. "Winning Solution of the BIB MVCS Challenge 2022." In: *The First Challenge on Machine Visual Common Sense: Perception, Prediction, Planning @ ECCV*. 2022.

[HD23] A. Hein and K. Diepold. "Comparing Intuitions about Agents' Goals, Preferences and Actions in Human Infants and Video Transformers." In: *Proceedings of the 45th Annual Meeting of the Cognitive Science Society (CogSci)*. 2023.

[HD24a] A. Hein and K. Diepold. "Exploring Early Number Abilities with Multimodal Transformers." In: *Cognitive Science* 48.9 (2024), e13492.

[HD24b] A. Hein and K. Diepold. "Modeling the Emergence of Letter Shapes." In: *Proceedings of the 46th Annual Meeting of the Cognitive Science Society (CogSci)*. 2024.

[He+16] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.90. URL: http://ieeexplore.ieee.org/document/7780459/.

[HE18] D. Ha and D. Eck. "A Neural Representation of Sketch Drawings." In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 2018.

[Hei+22] A. Hein, L. J. Meier, A. M. Buyx, and K. Diepold. "A Fuzzy-Cognitive-Maps Approach to Decision-Making in Medical Ethics." In: *Proceedings of the International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2022, pp. 1–8. ISBN: 978-1-66546-710-0. DOI: 10.1109/FUZZ-IEEE55066.2022.9882615.

[Hem98] C. Hempel. "Two Basic Types of Scientific Explanation." In: *Philosophy of Science: The Central Issues*. Ed. by M. Curd and J. A. Cover. 1st ed. W.W. Norton, 1998, pp. 685–694. ISBN: 978-0-393-97175-0.

[Her+17] K. M. Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin, M. Wainwright, C. Apps, D. Hassabis, and P. Blunsom. "Grounded Language Learning in a Simulated 3D World." In: *CoRR* abs/1706.06551 (2017). DOI: 10.48550/arXiv.1706.06551.

[Hey+17] D. K. Heyland, P. Dodek, J. J. You, T. Sinuff, T. Hiebert, C. Tayler, X. Jiang, J. Simon, and J. Downar. "Validation of Quality Indicators for End-of-Life Communication: Results of a Multicentre Survey." en. In: *Canadian Medical Association Journal* 189.30 (July 2017), E980–E989. ISSN: 0820-3946, 1488-2329. DOI: 10.1503/cmaj.160515. URL: http://www.cmaj.ca/lookup/doi/10.1503/cmaj.160515 (visited on 05/03/2024).

[HG15] D. F. Harwath and J. R. Glass. "Deep Multimodal Semantic Embeddings for Speech and Images." In: *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. 2015, pp. 237–244. DOI: 10.1109/ASRU.2015.7404800.

[HGD24] A. Hein, S. Gronauer, and K. Diepold. "Patient-Specific Modeling of Daily Activity Patterns for Unsupervised Detection of Psychotic and Non-Psychotic Relapses." In: *Proceedings of the 49th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024.

[HHS17] E. Hoffer, I. Hubara, and D. Soudry. "Train Longer, Generalize Better: Closing the Generalization Gap in Large Batch Training of Neural Networks." In: *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2017, pp. 1731–1741.

[HK17] T. Horikawa and Y. Kamitani. "Generic Decoding of Seen and Imagined Objects Using Hierarchical Visual Features." In: *Nature Communications* 8.1 (2017). ISSN: 2041-1723. DOI: 10.1038/ncomms15037.

[HM19] M. Hohol and M. Miłkowski. "Cognitive Artifacts for Geometric Reasoning." In: *Foundations of Science* 24.4 (2019), pp. 657–680. ISSN: 1233-1821. DOI: 10.1007/s10699-019-09603-w.

[HMK03] N. Hansen, S. D. Müller, and P. Koumoutsakos. "Reducing the Time Complexity of the Derandomized Evolution Strategy With Covariance Matrix Adaptation (CMA-ES)." In: *Evolutionary Computation* 11.1 (2003), pp. 1–18. DOI: 10.1162/106365603321828970.

[HO01a] N. Hansen and A. Ostermeier. "Completely Derandomized Self-Adaptation in Evolution Strategies." In: *Evolutionary Computation* 9.2 (2001), pp. 159–195. DOI: 10.1162/106365601750190398.

[HO01b] N. Hansen and A. Ostermeier. "Completely Derandomized Self-Adaptation in Evolution Strategies." In: *Evolutionary Computation* 9.2 (2001), pp. 159–195.

[Hom20] B. Hommel. "Pseudo-mechanistic Explanations in Psychology and Cognitive Neuroscience." In: *Topics in Cognitive Science* 12.4 (2020), pp. 1294–1305. ISSN: 1756-8757. DOI: 10.1111/tops.12448.

[HS04] M. D. Hauser and E. Spelke. "Evolutionary and Developmental Foundations of Human Knowledge." In: *The Cognitive Neurosciences* 3 (2004), pp. 853–864.

[HS09] D. C. Hyde and E. S. Spelke. "All Numbers Are Not Equal: An Electrophysiological Investigation of Small and Large Number Representations." In: *Journal of Cognitive Neuroscience* 21.6 (2009), pp. 1039–1053. DOI: 10.1162/jocn.2009.21090.

[HS21] M. Henderson and J. T. Serences. "Biased Orientation Representations Can Be Explained by Experience With Nonuniform Training Set Statistics." In: *Journal of Vision* 21.8 (2021), pp. 1–22. DOI: 10.1167/JOV.21.8.10.

[HS44] F. Heider and M. Simmel. "An Experimental Study of Apparent Behavior." In: *The American Journal of Psychology* 57.2 (1944), pp. 243–259. DOI: 10.2307/1416950.

[HSP18] C. B. Hornburg, S. A. Schmitt, and D. J. Purpura. "Relations Between Preschoolers' Mathematical Language Understanding and Specific Numeracy Skills." In: *Journal of Experimental Child Psychology* 176 (2018), pp. 84–100. DOI: 10.1016/j.jecp.2018.07.005.

[Hsu+21] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3451–3460. DOI: 10.1109/TASLP.2021.3122291.

[Hub+05] E. M. Hubbard, M. Piazza, P. Pinel, and S. Dehaene. "Interactions Between Number and Space in Parietal Cortex." In: *Nature Reviews Neuroscience* 6.6 (2005), pp. 435–448. ISSN: 1471-003X. DOI: 10.1038/nrn1684.

[Hum04] P. Humphreys. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press, 2004. ISBN: 978-0-19-515870-0. DOI: 10.1093/0195158709.001.0001.

[Hum09] P. Humphreys. "The Philosophical Novelty of Computer Simulation Methods." In: *Synthese* 169.3 (2009), pp. 615–626. DOI: 10.1007/s11229-008-9435-2.

[Hup+18] D. Hupkes, A. Singh, K. Korrel, G. Kruszewski, and E. Bruni. "Learning Compositionally Through Attentive Guidance." In: *CoRR* abs/1805.09657 (2018). DOI: 10.48550/arXiv.1805.09657.

[Hup+20] D. Hupkes, V. Dankers, M. Mul, and E. Bruni. "Compositionality Decomposed: How do Neural Networks Generalise?" In: *Journal of Artificial Intelligence Research* 67 (2020), pp. 757–795. DOI: 10.1613/jair.1.11674.

[IK20] E. Ilkou and M. Koutraki. "Symbolic Vs Sub-symbolic AI Methods: Friends or Enemies?" In: *Proceedings of the International Conference on Information and Knowledge Management (CIKM) Workshops*. CEUR Workshop Proceedings. 2020. DOI: 10.1145/3340531.3414072.

[Imb17] C. Imbert. "Computer Simulations and Computational Models in Science." In: *Springer Handbook of Model-Based Science*. Ed. by L. Magnani and T. Bertolotti. Springer International Publishing, 2017, pp. 735–781. ISBN: 978-3-319-30526-4. DOI: 10.1007/978-3-319-30526-4_34.

[Jad20] S. Jadon. "A Survey of Loss Functions for Semantic Segmentation." In: *Proceedings of the 17th Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2020, pp. 1–7. DOI: 10.1109/CIBCB48159.2020.9277638.

[JB16] C. Johnston and P. Bradbury. *100 Cases in Clinical Ethics and Law*. 2nd ed. 100 Cases. CRC Press, 2016. ISBN: 978-1-4987-3933-7. DOI: 10.1201/b19192.

[JB21] Y. Jiang and M. Bansal. "Inducing Transformer's Compositional Generalization Ability via Auxiliary Sequence Prediction Tasks." In: *Proceedings of the 16th Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021, pp. 6253–6265. DOI: 10.18653/v1/2021.emnlp-main.505.

[Jes78] P. J. Jeske. *The Effects of Modeling, Imitative Performance, and Modeling Feedback on Hierarchical Seriation Learning*. The University of Arizona, 1978.

[JK17] E. Jonas and K. P. Kording. "Could a Neuroscientist Understand a Microprocessor?" In: *PLOS Computational Biology* 13.1 (2017). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005268.

[JPS22] H. Ji, T. Patel, and O. Scharenborg. "Predicting Within and Across Language Phoneme Recognition Performance of Self-Supervised Learning Speech Pre-trained Models." In: *CoRR* abs/2206.12489 (2022). DOI: 10.48550/ARXIV.2206.12489.

[JS06] D. H. Jonassen and J. Strobel. "Modeling for Meaningful Learning." In: *Engaged Learning with Emerging Technologies*. Ed. by D. Hung and M. S. Khine. Springer-Verlag, 2006, pp. 1–27. ISBN: 978-1-4020-3668-2. DOI: 10.1007/1-4020-3669-8_1.

[JST20] J. Jara-Ettinger, L. E. Schulz, and J. B. Tenenbaum. "The Naïve Utility Calculus as a Unified, Quantitative Framework for Action Understanding." In: *Cognitive Psychology* 123 (2020), p. 101334. ISSN: 0010-0285. DOI: 10.1016/j.cogpsych.2020.101334.

[JTS22] H. Jee, M. Tamariz, and R. Shillcock. "Systematicity in Language and the Fast and Slow Creation of Writing Systems: Understanding Two Types of Non-arbitrary Relations Between Orthographic Characters and Their Canonical Pronunciation." In: *Cognition* 226 (2022), p. 105197. DOI: 10.1016/J.COGNITION.2022.105197.

[Kay+17] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. "The Kinetics Human Action Video Dataset." In: *CoRR* abs/1705.06950 (2017).

[Kay18] K. N. Kay. "Principles for Models of Neural Information Processing." In: *NeuroImage* 180 (2018), pp. 101–109. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2017.08.016.

[KB15a] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization." In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. 2015.

[KB15b] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization." In: *Proceedings of the 3rd International Conference on Learning Representations, (ICLR)*. 2015.

[KB23] S. Katz and Y. Belinkov. "VISIT: Visualizing and Interpreting the Semantic Information Flow of Transformers." In: *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023, pp. 14094–14113. DOI: 10.18653/v1/2023.findings-emnlp.939.

[KBO16] J. Kubilius, S. Bracci, and H. P. Op De Beeck. "Deep Neural Networks as a Computational Model for Human Shape Sensitivity." In: *PLOS Computational Biology* 12.4 (2016), e1004896. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004896.

[KD18] N. Kriegeskorte and P. K. Douglas. "Cognitive Computational Neuroscience." In: *Nature Neuroscience* 21.9 (2018), pp. 1148–1160. ISSN: 1097-6256. DOI: 10.1038/s41593-018-0210-5.

[Kel+21] P. Kelly, J. Winters, H. Miton, and O. Morin. "The Predictable Evolution of Letter Shapes: An Emergent Script of West Africa Recapitulates Historical Change in Writing Systems." In: *Current Anthropology* 62.6 (2021), pp. 669–691. DOI: 10.1086/717779.

[Kes+17] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima." In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. 2017.

[Key+19] D. Keysers, N. Schärli, N. Scales, H. Buisman, D. Furrer, S. Kashubin, N. Momchev, D. Sinopalnikov, L. Stafiniak, T. Tihon, et al. "Measuring Compositional Generalization: A Comprehensive Method on Realistic Data." In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. 2019.

[KG22] C. M. Klugman and S. Gerke. "Rise of the Bioethics AI: Curse or Blessing?" In: *The American Journal of Bioethics* 22.7 (2022), pp. 35–37. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075056.

[Kim+21] G. Kim, J. Jang, S. Baek, M. Song, and S.-B. Paik. "Visual Number Sense in Untrained Deep Neural Networks." In: *Science Advances* 7.1 (2021). ISSN: 2375-2548. DOI: 10.1126/sciadv.abd6127.

[Kir+23] R. Kirk, A. Zhang, E. Grefenstette, and T. Rocktäschel. "A Survey of Zero-shot Generalisation in Deep Reinforcement Learning." In: *Journal of Artificial Intelligence Research* 76 (2023), pp. 201–264. DOI: 10.1613/JAIR.1.14174.

[Kit14] R. Kitchin. "Big Data, New Epistemologies and Paradigm Shifts." In: *Big Data & Society* 1.1 (2014). ISSN: 2053-9517. DOI: 10.1177/2053951714528481.

[Kit89] P. Kitcher. "Explanatory Unification and the Causal Structure of the World." In: *Scientific Explanation*. University of Minnesota Press, 1989, pp. 410–505.

[KK14] S.-M. Khaligh-Razavi and N. Kriegeskorte. "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation." In: *PLoS Computational Biology* 10.11 (2014). Ed. by J. Diedrichsen, e1003915. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003915.

[KKB21] Y.-L. Kuo, B. Katz, and A. Barbu. "Compositional Networks Enable Systematic Generalization for Grounded Language Understanding." In: *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021, pp. 216–226. DOI: 10.18653/v1/2021.findings-emnlp.21.

[KKL13] M. E. Kolkman, E. H. Kroesbergen, and P. P. Leseman. "Early Numerical Development and the Role of Non-symbolic and Symbolic Skills." In: *Learning and Instruction* 25 (2013), pp. 95–103. DOI: 10.1016/j.learninstruc.2012.12.001.

[KM19] A. J. Kell and J. H. McDermott. "Deep Neural Network Models of Sensory Systems: Windows Onto the Role of Task Constraints." In: *Current Opinion in Neurobiology* 55 (2019), pp. 121–132. ISSN: 0959-4388. DOI: 10.1016/j.conb.2019.02.003.

[KMK19] T. C. Kietzmann, P. McClure, and N. Kriegeskorte. "Deep Neural Networks in Computational Neuroscience." In: *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press, 2019. ISBN: 978-0-19-026408-6. DOI: 10.1093/acrefore/9780190264086.013.46.

[KMM23] A. Koshevoy, H. Miton, and O. Morin. "Zipf's Law of Abbreviation Holds for Individual Characters Across a Broad Range of Writing Systems." In: *Cognition* 238 (2023). DOI: doi.org/10.1016/J.COGNITION.2023.105527.

[Knu05] T. Knuuttila. "Models, Representation, and Mediation." In: *Philosophy of Science* 72.5 (2005), pp. 1260–1271. ISSN: 0031-8248. DOI: 10.1086/508124.

[Knu11] T. Knuuttila. "Modelling and Representing: An Artefactual Approach to Model-Based Representation." In: *Studies in History and Philosophy of Science Part A* 42.2 (2011), pp. 262–271. ISSN: 0039-3681. DOI: 10.1016/j.shpsa.2010.11.034.

[Knu21] T. Knuuttila. “Epistemic Artifacts and the Modal Dimension of Modeling.” In: *European Journal for Philosophy of Science* 11.3 (2021), p. 65. ISSN: 1879-4912. DOI: 10.1007/s13194-021-00374-5.

[Kos86] B. Kosko. “Fuzzy Cognitive Maps.” In: *International Journal of Man-Machine Studies* 24.1 (1986), pp. 65–75. ISSN: 0020-7373. DOI: 10.1016/S0020-7373(86)80040-2.

[Kre+23] M. Krenn, L. Buffoni, B. Coutinho, S. Eppel, J. G. Foster, A. Gritsevskiy, H. Lee, Y. Lu, J. P. Moutinho, N. Sanjabi, R. Sonthalia, N. M. Tran, F. Valente, Y. Xie, R. Yu, and M. Kopp. “Forecasting the Future of Artificial Intelligence With Machine Learning-Based Link Prediction in an Exponentially Growing Knowledge Network.” In: *Nature Machine Intelligence* 5.11 (2023), pp. 1326–1335. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00735-0.

[Kri15] N. Kriegeskorte. “Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing.” In: *Annual Review of Vision Science* 1.1 (2015), pp. 417–446. ISSN: 2374-4642. DOI: 10.1146/annurev-vision-082114-035447.

[Kri20] M. Krishnan. “Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning.” In: *Philosophy & Technology* 33.3 (2020), pp. 487–502. ISSN: 2210-5433. DOI: 10.1007/s13347-019-00372-9.

[Kru96] P. Krugman. “What Economists Can Learn From Evolutionary Theorists.” In: *A talk given to the European Association for Evolutionary Political Economy* (1996).

[KT13] D. Kahneman and A. Tversky. “Prospect Theory: An Analysis of Decision Under Risk.” In: *World Scientific Handbook in Financial Economics Series*. Vol. 4. 2013, pp. 99–127. ISBN: 978-981-4417-34-1. DOI: 10.1142/9789814417358_0006.

[Kuh97] T. S. Kuhn. *The Structure of Scientific Revolutions*. Vol. 962. University of Chicago Press, 1997.

[Kum+10] M. Kumar, M. Husain, N. Upreti, and D. Gupta. “Genetic Algorithm: Review and Application.” In: *SSRN Electronic Journal* (2010). ISSN: 1556-5068. DOI: 10.2139/ssrn.3529843.

[Kur31] L. Kurt. “The Conflict Between Aristotelian and Galileian Modes of Thought in Contemporary Psychology.” In: *Journal of General Psychology* 5 (1931), pp. 141–177. DOI: 10.1080/00221309.1931.9918387.

[KV03] T. Knuuttila and A. Voutilainen. “A Parser as an Epistemic Artifact: A Material View on Models.” In: *Philosophy of Science* 70.5 (2003), pp. 1484–1495. DOI: 10.1086/377424.

[LA15] I. M. Lyons and D. Ansari. “Foundations of Children’s Numerical and Mathematical Skills: The Roles of Symbolic and Nonsymbolic Representations of Numerical Magnitude.” In: *Advances in Child Development and Behavior* 48 (2015), pp. 93–116. DOI: 10.1016/bs.acdb.2014.11.003.

[LAB12] I. M. Lyons, D. Ansari, and S. L. Beilock. “Symbolic Estrangement: Evidence Against a Strong Association Between Numerical Symbols and the Quantities They Represent.” In: *Journal of Experimental Psychology: General* 141.4 (2012), pp. 635–641. DOI: 10.1037/a0027248.

[Lak+17] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. “Building Machines That Learn and Think Like People.” In: *Behavioral and Brain Sciences* 40 (2017). DOI: 10.1017/S0140525X16001837.

[LB18] B. M. Lake and M. Baroni. "Generalization Without Systematicity: On the Compositional Skills of Sequence-To-Sequence Recurrent Networks." In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 2018, pp. 2873–2882.

[LB23] B. M. Lake and M. Baroni. "Human-Like Systematic Generalization Through a Meta-Learning Neural Network." In: *Nature* 623.7985 (2023), pp. 115–121. ISSN: 0028-0836. DOI: 10.1038/s41586-023-06668-3.

[LBL18] J. Loula, M. Baroni, and B. M. Lake. "Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks." In: *Proceedings of the 1st Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP) @ EMNLP*. 2018. DOI: 10.18653/v1/w18-5413.

[Lei+17] T. Leibovich, N. Katzin, M. Harel, and A. Henik. "From "Sense of Number" to "Sense of Magnitude": The Role of Continuous Magnitudes in Numerical Cognition." In: *Behavioral and Brain Sciences* 40 (2017), e164. DOI: 10.1017/S0140525X16000960.

[Leo14] S. Leonelli. "What Difference Does Quantity Make? On the Epistemology of Big Data in Biology." In: *Big Data & Society* 1.1 (2014). ISSN: 2053-9517. DOI: 10.1177/2053951714534395.

[LH17] I. Loshchilov and F. Hutter. "SGDR: Stochastic Gradient Descent with Warm Restarts." In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. 2017.

[Li+18] Y. Li, M. Zhang, Y. Chen, Z. Deng, X. Zhu, and S. Yan. "Children's Non-symbolic and Symbolic Numerical Representations and Their Associations With Mathematical Ability." In: *Frontiers in Psychology* 9 (2018). DOI: 10.3389/fpsyg.2018.01035.

[Li+20] L. L. Li, B. Yang, M. Liang, W. Zeng, M. Ren, S. Segal, and R. Urtasun. "End-To-End Contextual Perception and Prediction With Interaction Transformer." In: *Proceedings of the 33rd International Conference on Intelligent Robots and Systems (IROS)*. 2020, pp. 5784–5791. DOI: 10.1109/IROS45743.2020.9341392.

[Lin+14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. "Microsoft COCO: Common Objects in Context." In: *Proceedings of the 13th European Conference on Computer Vision (ECCV)*. Springer. 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48.

[Lin+19] Y.-C. Lin, Y.-L. Chao, C.-H. Hsu, H.-M. Hsu, P.-T. Chen, and L.-C. Kuo. "The Effect of Task Complexity on Handwriting Kinetics." In: *Canadian Journal of Occupational Therapy* 86.2 (2019), pp. 158–168. DOI: 10.1177/0008417419832327.

[Lin+22] T. Lin, Y. Wang, X. Liu, and X. Qiu. "A Survey of Transformers." In: *AI Open* 3 (2022), pp. 111–132. ISSN: 26666510. DOI: 10.1016/j.aiopen.2022.10.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666651022000146.

[Lin63] J. C. Lingoes. "Multiple Scalogram Analysis: A Set-Theoretic Model for Analyzing Dichotomous Items." In: *Educational and Psychological Measurement* 23.3 (1963), pp. 501–524. DOI: 10.1177/001316446302300307.

[Lip18] Z. C. Lipton. "The Mythos of Model Interpretability." In: *Communications of the ACM* 61.10 (2018), pp. 36–43. ISSN: 0001-0782. DOI: 10.1145/3233231.

[Liu+15] Z. Liu, P. Luo, X. Wang, and X. Tang. "Deep Learning Face Attributes in the Wild." In: *Proceedings of the 15th International Conference on Computer Vision (ICCV)*. 2015, pp. 3730–3738. DOI: 10.1109/ICCV.2015.425.

[Liu+20] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. "On the Variance of the Adaptive Learning Rate and Beyond." In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. 2020.

[LKB18] A. Liška, G. Kruszewski, and M. Baroni. "Memorize or Generalize? Searching for a Compositional RNN in a Haystack." In: *Proceedings of the FAIM Joint Workshop on Architectures and Evaluation for Generality, Autonomy and Progress in AI @ ICLR*. 2018.

[LKC20] W. Lotter, G. Kreiman, and D. Cox. "A Neural Network Trained for Prediction Mimics Diverse Features of Biological Neurons and Perception." In: *Nature Machine Intelligence* 2.4 (2020), pp. 210–219. ISSN: 2522-5839. DOI: 10.1038/s42256-020-0170-9.

[LL22] C. Li and Y. Liu. "Rethinking Query-Key Pairwise Interactions in Vision Transformers." In: *CoRR* abs/2207.00188 (2022). DOI: 10.48550/ARXIV.2207.00188.

[Lon73] H. C. Longuet-Higgins. "Comments on the Lighthill Report and the Sutherland Reply." In: *Artificial Intelligence: A Paper Symposium*. 1973, pp. 35–37.

[LS20] K. Lee and H. H. Schertz. "Brief Report: Analysis of the Relationship Between Turn Taking and Joint Attention for Toddlers with Autism." In: *Journal of Autism and Developmental Disorders* 50.7 (2020), pp. 2633–2640. ISSN: 0162-3257. DOI: 10.1007/s10803-019-03979-1.

[LSP06] S. Lazebnik, C. Schmid, and J. Ponce. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories." In: *Proceedings of the 19th Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. 2006, pp. 2169–2178. DOI: 10.1109/CVPR.2006.68.

[LSR96] S. E. Lea, A. M. Slater, and C. M. Ryan. "Perception of Object Unity in Chicks: A Comparison With the Human Infant." In: *Infant Behavior and Development* 19.4 (1996), pp. 501–504. ISSN: 0163-6383. DOI: 10.1016/S0163-6383(96)90010-7.

[MA06] N. Mateou and A. Andreou. "An Evolutionary Methodology to Eliminate the Limit Cycle Phenomenon in FCM-Based Models." In: *Proceedings of the 2nd International Conference on Information & Communication Technologies*. Vol. 1. 2006, pp. 1651–1656. ISBN: 978-0-7803-9521-3. DOI: 10.1109/ICTTA.2006.1684632.

[MA16] A. A. Matejko and D. Ansari. "Trajectories of Symbolic and Nonsymbolic Magnitude Processing in the First Year of Formal Schooling." In: *PloS one* 11.3 (2016), e0149863. DOI: 10.1371/journal.pone.0149863.

[Mäk05] U. Mäki. "Models Are Experiments, Experiments Are Models." In: *Journal of Economic Methodology* 12.2 (2005), pp. 303–315. DOI: 10.1080/13501780500086255.

[Mäk09] U. Mäki. "MISSing the World. Models as Isolations and Credible Surrogate Systems." In: *Erkenntnis* 70.1 (2009), pp. 29–43. ISSN: 0165-0106. DOI: 10.1007/s10670-008-9135-9.

[Mar+00] C. Marshuetz, E. E. Smith, J. Jonides, J. DeGutis, and T. L. Chenevert. "Order Information in Working Memory: fMRI Evidence for Parietal and Prefrontal Mechanisms." In: *Journal of Cognitive Neuroscience* 12.Supplement 2 (2000), pp. 130–144. DOI: 10.1162/08989290051137459.

[Mar06] H. Markram. "The Blue Brain Project." In: *Proceedings of the ACM/IEEE Conference on Supercomputing (SC)*. ACM Press, 2006, p. 53. ISBN: 978-0-7695-2700-0. DOI: 10.1145/1188455.1188511.

[Mar82] D. Marr. *Vision*. W. H. Freeman, 1982.

[Mas19] M. Massimi. "Two Kinds of Exploratory Models." In: *Philosophy of Science* 86.5 (2019), pp. 869–881. ISSN: 0031-8248. DOI: 10.1086/705494.

[Mat92] I. G. Mattingly. "Linguistic Awareness and Orthographic Form." In: *Advances in Psychology*. Vol. 94. 1992, pp. 11–26. DOI: 10.1016/S0166-4115(08)62786-7.

[MC23] N. C. for Mind, Brain and Consciousness. *Panel: What Can Deep Learning Do for Cognitive Science and Vice Versa?* 2023 Workshop on The Philosophy of Deep Learning. 2023. URL: https://www.youtube.com/watch?v=IaifsZV2mXI (visited on 02/05/2024).

[McC+16] J. L. McClelland, K. Mickey, S. Hansen, A. Yuan, and Q. Lu. "A Parallel-Distributed Processing Approach to Mathematical Cognition." In: *Manuscript, Stanford University* (2016).

[McC09] J. L. McClelland. "The Place of Modeling in Cognitive Science." In: *Topics in Cognitive Science* 1.1 (2009), pp. 11–38. ISSN: 1756-8757. DOI: 10.1111/j.1756-8765.2008.01003.x.

[McM85] E. McMullin. "Galilean Idealization." In: *Studies in History and Philosophy of Science Part A* 16.3 (1985), pp. 247–273. ISSN: 0039-3681. DOI: https://doi.org/10.1016/0039-3681(85)90003-2.

[Mei+22a] L. J. Meier, A. Hein, K. Diepold, and A. Buyx. "Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept." In: *The American Journal of Bioethics* 22.7 (2022), pp. 4–20. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2040647.

[Mei+22b] L. J. Meier, A. Hein, K. Diepold, and A. Buyx. "Clinical Ethics – To Compute, or Not to Compute?" In: *The American Journal of Bioethics* 22.12 (2022), W1–W4. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2127970.

[Mes+22] N. Messina, G. Amato, F. Carrara, C. Gennaro, and F. Falchi. "Recurrent Vision Transformer for Solving Visual Reasoning Problems." In: *Proceedings of the 21st International Conference in Image Analysis and Processing (ICIAP)*. Vol. 13233. Springer, 2022, pp. 50–61. DOI: 10.1007/978-3-031-06433-3_5.

[MG22] C. S. Mekik and C. M. Galang. "Cognitive Science in a Nutshell." In: *Cognitive Science* 46.8 (2022), e13179. ISSN: 0364-0213. DOI: 10.1111/cogs.13179.

[MH21] D. Mihai and J. S. Hare. "Learning to Draw: Emergent Communication through Sketching." In: *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2021, pp. 7153–7166.

[Mil03] G. A. Miller. "The Cognitive Revolution: A Historical Perspective." In: *Trends in Cognitive Sciences* 7.3 (2003), pp. 141–144. ISSN: 1364-6613. DOI: 10.1016/S1364-6613(03)00029-9.

[Mil07] R. Miller. "Another Slant on the Oblique Effect in Drawings and Paintings." In: *Empirical Studies of the Arts* 25.1 (2007), pp. 41–61. DOI: 10.2190/E737-V374-0640-6766.

[Mil10] H. J. Miller. "The Data Avalanche Is Here. Shouldn't We Be Digging?" In: *Journal of Regional Science* 50.1 (2010), pp. 181–201. ISSN: 0022-4146. DOI: 10.1111/j.1467-9787.2009.00641.x.

[Mis+18] D. K. Misra, A. Bennett, V. Blukis, E. Niklasson, M. Shatkhin, and Y. Artzi. "Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction." In: *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2018, pp. 2667–2678. DOI: 10.18653/v1/d18-1287.

[MK19] M. McGonigle-Chalmers and I. Kusel. "The Development of Size Sequencing Skills: An Empirical and Computational Analysis." In: *Monographs of the Society for Research in Child Development* 84.4 (2019). Number: 4, pp. 7–202. ISSN: 0037-976X. DOI: 10.1111/mono.12411.

[MKW20] O. Morin, P. Kelly, and J. Winters. "Writing, Graphic Codes, and Asynchronous Communication." In: *Topics in Cognitive Science* 12.2 (2020), pp. 727–743. DOI: 10.1111/TOPS.12386.

[ML18] D. Masters and C. Luschi. "Revisiting Small Batch Training for Deep Neural Networks." In: *CoRR* abs/1804.07612 (2018). DOI: 10.48550/arXiv.1804.07612.

[MM19] H. Miton and O. Morin. "When Iconicity Stands in the Way of Abbreviation: No Zipfian Effect for Figurative Signals." In: *PLoS One* 14.8 (2019), e0220793. DOI: 10.1371/JOURNAL.PONE.0220793.

[MM21] H. Miton and O. Morin. "Graphic Complexity in Writing Systems." In: *Cognition* 214 (2021), p. 104771. DOI: 10.1016/J.COGNITION.2021.104771.

[MM99] M. Morrison and M. S. Morgan. "Models as Mediating Instruments." In: *Models as Mediators*. 1st ed. Cambridge University Press, 1999, pp. 10–37. ISBN: 978-0-511-66010-8. DOI: 10.1017/CBO9780511660108.003.

[MML20] R. T. McCoy, J. Min, and T. Linzen. "BERTs of a Feather Do Not Generalize Together: Large Variability in Generalization Across Models With Similar Test Set Performance." In: *Proceedings of the 3rd Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP) @ EMNLP*. 2020, pp. 217–227. DOI: 10.18653/v1/2020.blackboxnlp-1.21.

[Mom23] I. Momennejad. "A Rubric for Human-Like Agents and NeuroAI." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 378.1869 (2023). ISSN: 0962-8436. DOI: 10.1098/rstb.2021.0446.

[Mor07] M. Morrison. "Where Have All the Theories Gone?" In: *Philosophy of Science* 74.2 (2007), pp. 195–228. ISSN: 0031-8248. DOI: 10.1086/520778.

[Mor16] O. Morin. "The Spontaneous Emergence of Functional Complexity in Writing Systems: The Case of Cardinal Lines." In: *The Idea of Writing* (2016), p. 13.

[Mor18] O. Morin. "Spontaneous Emergence of Legibility in Writing Systems: The Case of Orientation Anisotropy." In: *Cognitive Science* 42.2 (2018), pp. 664–677. DOI: 10.1111/COGS.12550.

[Mor22a] O. Morin. "The Piecemeal Evolution of Writing." In: *Lingue e Linguaggio* 21.2 (2022), pp. 217–237. DOI: 10.1418/105963.

[Mor22b] O. Morin. "The Puzzle of Ideography." In: *Behavioral and Brain Sciences* (2022), pp. 1–69. DOI: 10.1017/S0140525X22002801.

[MP20] W. J. Ma and B. Peters. "A Neural Network Walks Into a Lab: Towards Using Deep Nets as Models for Human Behavior." In: *CoRR* abs/2005.02181 (2020). DOI: 10.48550/arXiv.2005.02181.

[MS82] G. Mandler and B. J. Shebo. "Subitizing: An Analysis of Its Component Processes." In: *Journal of Experimental Psychology* 111.1 (1982), pp. 1–22. DOI: 10.1037/0096-3445.111.1.1.

[MS99] D. Mareschal and T. R. Shultz. "Development of Children's Seriation: A Connectionist Approach." In: *Connection Science* 11.2 (1999), pp. 149–186. ISSN: 0954-0091. DOI: 10.1080/095400999116322.

[Muh+18] U. R. Muhammad, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales. "Learning Deep Sketch Abstraction." In: *Proceedings of the 31st Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8014–8023. DOI: 10.1109/CVPR.2018.00836.

[Mus+14] C. Mussolin, J. Nys, A. Content, and J. Leybaert. "Symbolic Number Abilities Predict Later Approximate Number System Acuity in Preschool Children." In: *PLoS One* 9.3 (2014), e91839. DOI: 10.1371/journal.pone.0091839.

[Nak20] K. Nakajima. "Physical Reservoir Computing—an Introductory Perspective." In: *Japanese Journal of Applied Physics* 59.6 (2020). ISSN: 0021-4922. DOI: 10.35848/1347-4065/ab8d4f.

[NFG87] R. S. Newman, C. A. Friedman, and D. R. Gockley. "Children's Use of Multiple-Counting Skills: Adaptation to Task Factors." In: *Journal of Experimental Child Psychology* 44.2 (1987), pp. 268–282. DOI: 10.1016/0022-0965(87)90034-8.

[NFM02] A. Nieder, D. J. Freedman, and E. K. Miller. "Representation of the Quantity of Visual Items in the Primate Prefrontal Cortex." In: *Science* 297.5587 (2002), pp. 1708–1711. DOI: 10.1126/science.1072493.

[Nie+22] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper. "A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion." In: *Proceedings of the 47th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 6562–6566. DOI: 10.1109/ICASSP43922.2022.9746484.

[Nie05] A. Nieder. "Counting on Neurons: The Neurobiology of Numerical Competence." In: *Nature Reviews Neuroscience* 6.3 (2005), pp. 177–190. ISSN: 1471-003X. DOI: 10.1038/nrn1626.

[NMT03] Y. Ninokura, H. Mushiake, and J. Tanji. "Representation of the Temporal Order of Visual Objects in the Primate Lateral Prefrontal Cortex." In: *Journal of Neurophysiology* 89.5 (2003), pp. 2868–2873. DOI: 10.1152/jn.00647.2002.

[NMT04] Y. Ninokura, H. Mushiake, and J. Tanji. "Integration of Temporal Order and Object Information in the Monkey Lateral Prefrontal Cortex." In: *Journal of Neurophysiology* 91.1 (2004), pp. 555–560. DOI: 10.1152/jn.00694.2003.

[NN21] K. Nasr and A. Nieder. "Spontaneous Representation of Numerosity Zero in a Deep Neural Network for Visual Object Recognition." In: *iScience* 24.11 (2021). ISSN: 2589-0042. DOI: 10.1016/j.isci.2021.103301.

[nos20] nostalgebraist. *interpreting GPT: The Logit Lens*. 2020. URL: https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens (visited on 08/05/2023).

[NPS14] D. Napoletani, M. Panza, and D. C. Struppa. "Is Big Data Enough? A Reflection on the Changing Role of Mathematics in Applications." In: *Notices of the American Mathematical Society* 61.5 (2014), pp. 485–490. DOI: 10.1090/noti1102.

[Núñ+19] R. Núñez, M. Allen, R. Gao, C. Miller Rigoli, J. Relaford-Doyle, and A. Semenuks. "What Happened to Cognitive Science?" In: *Nature Human Behaviour* 3.8 (2019), pp. 782–791. ISSN: 2397-3374. DOI: 10.1038/s41562-019-0626-2.

[Nye+21] M. Nye, M. Tessler, J. Tenenbaum, and B. M. Lake. "Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro-Symbolic Reasoning." In: *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)* (2021), pp. 25192–25204.

[Ont+22] S. Ontañón, J. Ainslie, Z. Fisher, and V. Cvicek. "Making Transformers Solve Compositional Tasks." In: *Proceedings of the 60th Annual Meeting of the Association for Computational (ACL)*. Association for Computational Linguistics, 2022, pp. 3591–3607. DOI: 10.18653/V1/2022.ACL-LONG.251.

[Ope23] OpenAI. "GPT-4 Technical Report." In: *CoRR* abs/2303.08774 (2023). DOI: 10.48550/arXiv.2303.08774.

[PA16] S. Piantadosi and R. Aslin. "Compositional Reasoning in Early Childhood." In: *PLOS ONE* 11.9 (2016). Ed. by A. Bruce, e0147734. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0147734.

[PAG17] J. C. Peterson, J. T. Abbott, and T. L. Griffiths. "Adapting Deep Network Features to Capture Psychological Representations: An Abridged Report." In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 2017, pp. 4934–4938. ISBN: 978-0-9992411-0-3. DOI: 10.24963/ijcai.2017/697.

[Pan+15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. "Librispeech: An ASR Corpus Based on Public Domain Audio Books." In: *Proceedings of the 40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[Pan+22] M. Panagiotou, A. Zlatintsi, P. P. Filntisis, A. J. Roumeliotis, N. Efthymiou, and P. Maragos. "A Comparative Study of Autoencoder Architectures for Mental Health Analysis Using Wearable Sensors Data." In: *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*. 2022, pp. 1258–1262. ISBN: 978-90-827970-9-1. DOI: 10.23919/EUSIPCO55093.2022.9909697.

[Par20] S. W. Park. "Generating Novel Glyph without Human Data by Learning to Communicate." In: *Machine Learning for Creativity and Design Workshop @ NeurIPS*. 2020.

[Pas+19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Proceedings of the 32nd Annual Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2019, pp. 8024–8035.

[PB22] B. Pilkington and C. Binkley. "Disproof of Concept: Resolving Ethical Dilemmas Using Algorithms." In: *The American Journal of Bioethics* 22.7 (2022), pp. 81–83. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2087789.

[Per92] M. L. Perlin. "Fatal Assumption: A Critical Evaluation of the Role of Counsel in Mental Disability Cases." In: *Law and Human Behavior* 16.1 (1992), pp. 39–59. ISSN: 1573-661X. DOI: 10.1007/BF02351048.

[PF12] S. R. Powell and L. S. Fuchs. "Early Numerical Competencies and Students with Mathematics Difficulty." In: *Focus on Exceptional Children* 44.5 (2012), pp. 1–16.

[PG23] G. Paaß and S. Giesselbach. “Foundation Models for Speech, Images, Videos, and Control.” In: *Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media*. Springer, 2023, pp. 313–382. DOI: 10.1007/978-3-031-23190-2_7.

[PGH52] J. Piaget, C. Gattegno, and F. Hodgson. *The Child's Conception of Number*. London: Routledge & Kegan Paul Ltd, 1952. ISBN: 9789120004266.

[PGP22] B. Pitt, E. Gibson, and S. T. Piantadosi. “Exact Number Concepts Are Limited to the Verbal Count Range.” In: *Psychological Science* (2022). ISSN: 0956-7976. DOI: 10.1177/09567976211034502.

[Pia+02] M. Piazza, A. Mechelli, B. Butterworth, and C. J. Price. “Are Subitizing and Counting Implemented as Separate or Functionally Overlapping Processes?” In: *Neuroimage* 15.2 (2002), pp. 435–446. DOI: 10.1006/nimg.2001.0980.

[Pia16] S. T. Piantadosi. “A Rational Analysis of the Approximate Number System.” In: *Psychonomic Bulletin & Review* 23 (2016), pp. 877–886. DOI: 10.3758/s13423-015-0963-8.

[Pia61] J. Piaget. “The Genetic Approach to the Psychology of Thought.” In: *Journal of Educational Psychology* 52.6 (1961), pp. 275–281. DOI: 10.1037/h0042963.

[Pic+04a] P. Pica, C. Lemer, V. Izard, and S. Dehaene. “Exact and Approximate Arithmetic in an Amazonian Indigene Group.” In: *Science* 306.5695 (2004), pp. 499–503. ISSN: 0036-8075. DOI: 10.1126/science.1102085.

[Pic+04b] P. Pica, C. Lemer, V. Izard, and S. Dehaene. “Exact and Approximate Arithmetic in an Amazonian Indigene Group.” In: *Science* 306.5695 (2004), pp. 499–503. DOI: 10.1126/science.1102085.

[Pin+21] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle. “Improving Reproducibility in Machine Learning Research (A Report From the NeurIPS 2019 Reproducibility Program).” In: *Journal of Machine Learning Research* 22.1 (2021). ISSN: 1532-4435.

[Pit+21] B. Pitt, S. Ferrigno, J. F. Cantlon, D. Casasanto, E. Gibson, and S. T. Piantadosi. “Spatial Concepts of Number, Size, and Time in an Indigenous Culture.” In: *Science Advances* 7.33 (2021). DOI: 10.1126/sciadv.abg414.

[PL68] M. C. Potter and E. I. Levy. “Spatial Enumeration without Counting.” In: *Child Development* 39.1 (1968), pp. 265–272. DOI: 10.2307/1127377.

[Pol+87] A. Pollatsek, A. D. Well, C. Konold, P. Hardiman, and G. Cobb. “Understanding Conditional Probabilities.” In: *Organizational Behavior and Human Decision Processes* 40.2 (1987), pp. 255–269. ISSN: 07495978. DOI: 10.1016/0749-5978(87)90015-X.

[Pow+21] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. “Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets.” In: *Mathematical Reasoning in General Artificial Intelligence Workshop (MATHAI) @ ICLR*. 2021.

[PP10] G. E. Pence and G. E. Pence. *Medical Ethics: Accounts of Ground-Breaking Cases*. 6th ed. McGraw-Hill, 2010. ISBN: 978-0-07-340749-4.

[PP20] P. Perconti and A. Plebe. “Deep Learning and Cognitive Science.” In: *Cognition* 203 (2020), p. 104365. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2020.104365.

[PR16] D. J. Purpura and E. E. Reid. "Mathematics and Language: Individual and Group Differences in Mathematical Language Skills in Young Children." In: *Early Childhood Research Quarterly* 36 (2016), pp. 259–268. DOI: 10.1016/j.ecresq.2015.12.020.

[Pra+22] R. Pradhan, J. Zhu, B. Glavic, and B. Salimi. "Interpretable Data-Based Explanations for Fairness Debugging." In: *Proceedings of the 47th International Conference on Management of Data (SIGMOD)*. ACM, 2022, pp. 247–261. ISBN: 978-1-4503-9249-5. DOI: 10.1145/3514221.3517886.

[PS00] S. A. Peterson and T. J. Simon. "Computational Evidence for the Subitizing Phenomenon as an Emergent Property of the Human Cognitive Architecture." In: *Cognitive Science* 24.1 (2000), pp. 93–122. DOI: 10.1016/S0364-0213(99)00022-1.

[PS13] E. I. Papageorgiou and J. L. Salmeron. "A Review of Fuzzy Cognitive Maps Research During the Last Decade." In: *IEEE Transactions on Fuzzy Systems* 21.1 (2013), pp. 66–79. ISSN: 1063-6706. DOI: 10.1109/TFUZZ.2012.2201727.

[PTG12] S. T. Piantadosi, J. B. Tenenbaum, and N. D. Goodman. "Bootstrapping in a Language of Thought: A Formal Model of Numerical Concept Learning." In: *Cognition* 123.2 (2012), pp. 199–217. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2011.11.005.

[Qiu+21] L. Qiu, H. Hu, B. Zhang, P. Shaw, and F. Sha. "Systematic Generalization on gSCAN: What is Nearly Solved and What is Next?" In: *Proceedings of the 16th Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021, pp. 2180–2188. DOI: 10.18653/v1/2021.emnlp-main.166.

[Qiu+22] S. Qiu, S. Xie, L. Fan, T. Gao, J. Joo, S. Zhu, and Y. Zhu. "Emergent Graphical Conventions in a Visual Communication Game." In: *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2022, pp. 13119–13131.

[Rah+22] V. Rahimzadeh, J. Lawson, J. Baek, and E. S. Dove. "Automating Justice: An Ethical Responsibility of Computational Bioethics." In: *The American Journal of Bioethics* 22.7 (2022), pp. 30–33. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075051.

[RCB11] M. Rucinski, A. Cangelosi, and T. Belpaeme. "An Embodied Developmental Robotic Model of Interactions Between Numbers and Space." In: *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci)*. 2011.

[RCB12] M. Rucinski, A. Cangelosi, and T. Belpaeme. "Robotic Model of the Contribution of Gesture to Learning to Count." In: *Proceedings of the 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. 2012, pp. 1–6. ISBN: 978-1-4673-4965-9. DOI: 10.1109/DevLrn.2012.6400579.

[RCW15] A. M. Rush, S. Chopra, and J. Weston. "A Neural Attention Model for Abstractive Sentence Summarization." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015, pp. 379–389. DOI: 10.18653/v1/D15-1044. URL: http://aclweb.org/anthology/D15-1044.

[RD99] P. J. Rousseeuw and K. V. Driessen. "A Fast Algorithm for the Minimum Covariance Determinant Estimator." In: *Technometrics* 41.3 (1999), pp. 212–223. ISSN: 0040-1706. DOI: 10.1080/00401706.1999.10485670.

[Res73] L. B. Resnick. "Hierarchies in Children's Learning: A Symposium." In: *Instructional Science* 2.3 (1973), pp. 311–361. ISSN: 00204277. DOI: 10.1007/BF00119059.

[Rev+08] S. K. Revkin, M. Piazza, V. Izard, L. Cohen, and S. Dehaene. "Does Subitizing Reflect Numerical Estimation?" In: *Psychological Science* 19.6 (2008), pp. 607–614. DOI: 10.1111/j.1467-9280.2008.02130.x.

[RFJ01] S. A. Rose, J. F. Feldman, and J. J. Jankowski. "Visual Short-Term Memory in the First Year of Life: Capacity and Recency Effects." In: *Developmental Psychology* 37.4 (2001), p. 539. DOI: 10.1037/0012-1649.37.4.539.

[RHH18] A. Reutlinger, D. Hangleiter, and S. Hartmann. "Understanding (with) Toy Models." In: *The British Journal for the Philosophy of Science* 69.4 (2018), pp. 1069–1099. ISSN: 0007-0882. DOI: 10.1093/bjps/axx005.

[Rip17] L. J. Rips. "Core Cognition and Its Aftermath." In: *Philosophical Topics* 45.1 (2017), pp. 157–180. ISSN: 02762080.

[RK22] M. Rost and T. Knuuttila. "Models as Epistemic Artifacts for Scientific Reasoning in Science Education Research." In: *Education Sciences* 12.4 (2022), p. 276. ISSN: 2227-7102. DOI: 10.3390/educsci12040276.

[RL18] E. Ratti and E. López-Rubio. "Mechanistic Models and the Explanatory Limits of Machine Learning." In: *Mechanism Meets Big Data: Different Strategies for Machine Learning in Cancer Research @ PSA*. 2018.

[RLG15] G. Roberts, J. Lewandowski, and B. Galantucci. "How Communication Changes When We Cannot Mime the World: Experimental Evidence for the Effect of Iconicity on Combinatoriality." In: *Cognition* 141 (2015), pp. 52–66. DOI: 10.1016/J.COGNITION.2015.04.001.

[RN10] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. London, 2010.

[ROL03] S. Ross-Sheehy, L. M. Oakes, and S. J. Luck. "The Development of Visual Short-Term Memory Capacity in Infants." In: *Child Development* 74.6 (2003), pp. 1807–1822. DOI: 10.1046/j.1467-8624.2003.00639.x.

[Rud19] C. Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

[Rui+20] L. Ruis, J. Andreas, M. Baroni, D. Bouchacourt, and B. M. Lake. "A Benchmark for Systematic Generalization in Grounded Language Understanding." In: *Proceedings of the 33th Annual Conference on Neural Information Processing Systems (NeurIPS)* (2020), pp. 19861–19872.

[RV95] L. Regolin and G. Vallortigara. "Perception of Partly Occluded Objects by Young Chicks." In: *Perception & Psychophysics* 57.7 (1995), pp. 971–976. ISSN: 0031-5117. DOI: 10.3758/BF03205456.

[RWK73] L. B. Resnick, M. C. Wang, and J. Kaplan. "Task Analysis in Curriculum Design: A Hierarchically Sequenced Introductory Mathematics Curriculum." In: *Journal of Applied Behavior Analysis* 6.4 (1973), pp. 679–709. ISSN: 0021-8855. DOI: 10.1901/jaba.1973.6-679.

[SA19] J. Symons and R. Alvarado. "Epistemic Entitlements and the Practice of Computer Simulation." In: *Minds and Machines* 29.1 (2019), pp. 37–60. ISSN: 0924-6495. DOI: 10.1007/s11023-018-9487-0.

[Sab+21] M. Sablé-Meyer, J. Fagot, S. Caparos, T. Van Kerkoerle, M. Amalric, and S. Dehaene. "Sensitivity to Geometric Shape Regularity in Humans and Baboons: A Putative Signature of Human Singularity." In: *Proceedings of the National Academy of Sciences* 118.16 (2021). ISSN: 0027-8424. DOI: 10.1073/pnas.2023123118.

[Sab22] M. Sabatello. "Wrongful Birth: AI-Tools for Moral Decisions in Clinical Care in the Absence of Disability Ethics." In: *The American Journal of Bioethics* 22.7 (2022), pp. 43–46. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075971.

[Sal20] W. C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, 2020. ISBN: 978-0-691-10170-5. DOI: 10.2307/j.ctv173f2gh.

[SC08] B. W. Sarnecka and S. Carey. "How Counting Represents Number: What Children Must Learn and When They Learn It." In: *Cognition* 108.3 (2008), pp. 662–674. DOI: 10.1016/j.cognition.2008.05.007.

[Sch+18] M. Schneider, S. Merz, J. Stricker, B. De Smedt, J. Torbeyns, L. Verschaffel, and K. Luwel. "Associations of Number Line Estimation With Mathematical Competence: A Meta-analysis." In: *Child Development* 89.5 (2018), pp. 1467–1484. ISSN: 0009-3920. DOI: 10.1111/cdev.13068.

[Sch+21] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. "The Neural Architecture of Language: Integrative Modeling Converges on Predictive Processing." In: *Proceedings of the National Academy of Sciences* 118.45 (2021), e2105646118. ISSN: 0027-8424. DOI: 10.1073/pnas.2105646118.

[SES74] B. Schaeffer, V. H. Eggleston, and J. L. Scott. "Number Development in Young Children." In: *Cognitive Psychology* 6.3 (1974), pp. 357–379. DOI: 10.1016/0010-0285(74)90017-6.

[SG08] J. E. Snyder and C. C. Gauthier. *Evidence-Based Medical Ethics: Cases for Practice-Based Learning*. Humana Pr, 2008. ISBN: 978-1-60327-245-2.

[SH19] C. Steppa and T. L. Holch. "HexagDLy—Processing Hexagonally Sampled Data with CNNs in PyTorch." In: *SoftwareX* 9 (2019), pp. 193–198. ISSN: 2352-7110. DOI: 10.1016/J.SOFTX.2019.02.010.

[Sha+20] G. Shala, A. Biedenkapp, N. Awad, S. Adriaensen, M. Lindauer, and F. Hutter. "Learning Step-Size Adaptation in CMA-ES." In: *Proeedings of the 15th International Conference on Parallel Problem Solving from Nature*. Springer. 2020, pp. 691–706. DOI: 10.1007/978-3-030-58112-1_48.

[Sha48] C. E. Shannon. "A Mathematical Theory of Communication." In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/J.1538-7305.1948.tb01338.X.

[Sha78] L. Shannon. "Spatial Strategies in the Counting of Young Children." In: *Child Development* 49.4 (1978), pp. 1212–1215. ISSN: 0009-3920. DOI: 10.2307/1128762.

[She+21] H. Sheahan, F. Luyckx, S. Nelli, C. Teupe, and C. Summerfield. "Neural State Space Alignment for Magnitude Generalization in Humans and Recurrent Networks." In: *Neuron* 109.7 (2021), pp. 1214–1226. DOI: 10.1016/j.neuron.2021.02.004.

[SHK22] A. Sauerbrei, N. Hallowell, and A. Kerasidou. "AIgorithmic Ethics: A Technically Sweet Solution to a Non-Problem." In: *The American Journal of Bioethics* 22.7 (2022), pp. 28–30. ISSN: 1526-5161. DOI: 10.1080/15265161.2022.2075050.

[Shr+20] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox. "ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks." In: *Proceedings of the 33rd Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 10737–10746. DOI: 10.1109/CVPR42600.2020.01075.

[Sie71] L. S. Siegel. "The Development of the Understanding of Certain Number Concepts." In: *Developmental Psychology* 5.2 (1971), pp. 362–363. ISSN: 0012-1649. DOI: 10.1037/h0031425.

[SK07] E. S. Spelke and K. D. Kinzler. "Core Knowledge." In: *Developmental Science* 10.1 (2007), pp. 89–96. ISSN: 1363-755X. DOI: 10.1111/j.1467-7687.2007.00569.x.

[SK20] K. R. Storrs and N. Kriegeskorte. "Deep Learning for Cognitive Neuroscience." In: *The Cognitive Neurosciences*. Ed. by D. Poeppel, G. R. Mangun, and M. S. Gazzaniga. 6th ed. The MIT Press, 2020, pp. 703–716. ISBN: 978-0-262-35617-6. DOI: 10.7551/mitpress/11442.003.0077.

[SL12] E. S. Spelke and S. A. Lee. "Core Systems of Geometry in Animal Minds." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1603 (2012), pp. 2784–2793. ISSN: 0962-8436. DOI: 10.1098/rstb.2012.0210.

[SL18] S. L. Smith and Q. V. Le. "A Bayesian Perspective on Generalization and Stochastic Gradient Descent." In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 2018.

[SM14] G. S. Starkey and B. D. McCandliss. "The Emergence of "Groupitizing" in Children's Numerical Cognition." In: *Journal of Experimental Child Psychology* 126 (2014), pp. 120–137. DOI: 10.1016/j.jecp.2014.03.006.

[SMH22] B. Shi, A. Mohamed, and W. Hsu. "Learning Lip-Based Audio-Visual Speaker Embeddings With AV-HuBERT." In: *Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2022, pp. 4785–4789. DOI: 10.21437/INTERSPEECH.2022-885.

[Smi+18] S. L. Smith, P. Kindermans, C. Ying, and Q. V. Le. "Don't Decay the Learning Rate, Increase the Batch Size." In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 2018.

[SMP12] A. M. Slowther, L. McClimans, and C. Price. "Development of Clinical Ethics Services in the UK: A National Survey." In: *Journal of Medical Ethics* 38.4 (2012), pp. 210–214. ISSN: 0306-6800. DOI: 10.1136/medethics-2011-100173.

[SMS20a] S. Sabathiel, J. L. McClelland, and T. Solstad. "Emerging Representations for Counting in a Neural Network Agent Interacting with a Multimodal Environment." In: *Proceedings of the 6th Conference on Artificial Life (ALife)*. 2020, pp. 736–743. DOI: 10.1162/isal_a_00333.

[SMS20b] S. Sabathiel, J. McClelland, and T. Solstad. "A Computational Model of Learning to Count in a Multimodal, Interactive Environment." In: *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci)*. 2020.

[SNS21] A. Saxe, S. Nelli, and C. Summerfield. "If Deep Learning Is the Answer, What Is the Question?" In: *Nature Reviews Neuroscience* 22.1 (2021), pp. 55–67. ISSN: 1471-003X. DOI: 10.1038/s41583-020-00395-8.

[Soc] C. S. Society. *Our History*. URL: https://cognitivesciencesociety.org/about/ (visited on 04/29/2024).

[Son+18] J. Song, K. Pang, Y.-Z. Song, T. Xiang, and T. M. Hospedales. “Learning to Sketch with Shortcut Cycle Consistency.” In: *Proceedings of the 31st Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 801–810. DOI: 10.1109/CVPR.2018.00090.

[Spe22] E. S. Spelke. *What Babies Know: Core Knowledge and Composition Volume 1*. Vol. 1. Oxford University Press, 2022.

[SPG91] G. Salomon, D. N. Perkins, and T. Globerson. “Partners in Cognition: Extending Human Intelligence with Intelligent Technologies.” In: *Educational Researcher* 20.3 (1991), pp. 2–9. ISSN: 0013-189X. DOI: 10.2307/1177234.

[SPW95] E. S. Spelke, A. Phillips, and A. L. Woodward. “Infants’ Knowledge of Object Motion and Human Action.” In: *Causal Cognition: A Multidisciplinary Debate*. Symposia of the Fyssen Foundation. Clarendon Press/Oxford University Press, 1995, pp. 44–78. ISBN: 0-19-852314-9. DOI: 10.1093/acprof:oso/9780198524021.003.0003.

[SSG19] S. Subramanian, S. Singh, and M. Gardner. “Analyzing Compositionality in Visual Question Answering.” In: *Proceedings of the Visually Grounded Interaction and Language Workshop (VIGIL) @ NeurIPS* 7 (2019).

[Sto+23] G. Stojnić, K. Gandhi, S. Yasuda, B. M. Lake, and M. R. Dillon. “Commonsense Psychology in Human Infants and Machines.” In: *Cognition* 235 (2023). DOI: 10.1016/j.cognition.2023.105406.

[Sui+21] Z. Sui, Y. Zhou, X. Zhao, A. Chen, and Y. Ni. “Joint Intention and Trajectory Prediction Based on Transformer.” In: *Proceedings of the 34th International Conference on Intelligent Robots and Systems (IROS)*. 2021, pp. 7082–7088. DOI: 10.1109/IROS51168.2021.9636241.

[Sul22] E. Sullivan. “Understanding from Machine Learning Models.” In: *The British Journal for the Philosophy of Science* 73.1 (2022), pp. 109–133. ISSN: 0007-0882. DOI: 10.1093/bjps/axz035.

[Sup60] P. Suppes. “A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences.” In: *Synthese* 12.2/3 (1960), pp. 287–301. ISSN: 00397857. DOI: 10.1007/BF00485107.

[SV93] E. S. Spelke and G. Van de Walle. “Perceiving and Reasoning About Objects: Insights From Infants.” In: *Spatial Representation: Problems in Philosophy and Psychology* (1993), pp. 132–161.

[SZ12] I. Stoianov and M. Zorzi. “Emergence of a ‘Visual Number Sense’ in Hierarchical Generative Models.” In: *Nature Neuroscience* 15.2 (2012), pp. 194–196. ISSN: 1097-6256. DOI: 10.1038/nn.2996.

[SZS12] K. Soomro, A. R. Zamir, and M. Shah. “UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild.” In: *CoRR* abs/1212.0402 (2012). DOI: 10.48550/arXiv.1212.0402.

[TC07] M. Tomasello and M. Carpenter. “Shared Intentionality.” In: *Developmental Science* 10.1 (2007), pp. 121–125. ISSN: 1363-755X. DOI: 10.1111/j.1467-7687.2007.00573.x.

[TH18] E. Y. Toomarian and E. M. Hubbard. “On the Genesis of Spatial-Numerical Associations: Evolutionary and Cultural Factors Co-construct the Mental Number Line.” In: *Neuroscience and Biobehavioral Reviews* 90 (2018), pp. 184–199. ISSN: 1873-7528. DOI: 10.1016/j.neubiorev.2018.04.010.

[TK97] W. Tomic and J. Kingma. “The Relationship between Seriation and Number Line Comprehension: A Validation Study.” In: *Curriculum and Teaching* 12.2 (1997), pp. 59–69. ISSN: 0726-416X. DOI: 10.7459/ct/12.2.06.

[TNH20] Y. Tang, D. Nguyen, and D. Ha. “Neuroevolution of Self-Interpretable Agents.” In: *Proceedings of the 21st Genetic and Evolutionary Computation Conference (GECCO)*. 2020, pp. 414–424. DOI: 10.1145/3377930.3389847.

[Too+21] A. Toosi, A. G. Bottino, B. Saboury, E. Siegel, and A. Rahmim. “A Brief History of AI: How to Prevent Another Winter (A Critical Review).” In: *PET Clinics* 16.4 (2021), pp. 449–469. ISSN: 1556-8598. DOI: 10.1016/j.cpet.2021.07.001.

[TSZ17] A. Testolin, I. Stoianov, and M. Zorzi. “Letter Perception Emerges from Unsupervised Deep Learning and Recycling of Natural Image Features.” In: *Nature Human Behaviour* 1.9 (2017), pp. 657–664. DOI: 10.1038/s41562-017-0186-2.

[TV14] S. W. Toll and J. E. Van Luit. “The Developmental Relationship Between Language and Low Early Numeracy Skills Throughout Kindergarten.” In: *Exceptional Children* 81.1 (2014), pp. 64–78. DOI: 10.1177/0014402914532233.

[Tyn+17] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik. “The Influences of Emotion on Learning and Memory.” In: *Frontiers in Psychology* 8 (2017), p. 1454. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01454.

[TZM20] A. Testolin, W. Y. Zou, and J. L. McClelland. “Numerosity Discrimination in Deep Neural Networks: Initial Competence, Developmental Refinement and Experience Statistics.” In: *Developmental Science* 23.5 (2020), e12940. ISSN: 1467-7687. DOI: 10.1111/desc.12940.

[Val+06] E. Valenza, I. Leo, L. Gava, and F. Simion. “Perceptual Completion in Newborn Human Infants.” In: *Child Development* 77.6 (2006), pp. 1810–1821. ISSN: 0009-3920. DOI: 10.1111/j.1467-8624.2006.00975.x.

[Var+18] K. Varelas, A. Auger, D. Brockhoff, N. Hansen, O. A. ElHara, Y. Semet, R. Kassab, and F. Barbaresco. “A Comparative Study of Large-Scale Variants of CMA-ES.” In: *Parallel Problem Solving from Nature – PPSN XV*. Ed. by A. Auger, C. M. Fonseca, N. Lourenço, P. Machado, L. Paquete, and D. Whitley. Vol. 11101. Series Title: Lecture Notes in Computer Science. Springer International Publishing, 2018, pp. 3–15. ISBN: 978-3-319-99252-5. DOI: 10.1007/978-3-319-99253-2_1.

[Vas+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. “Attention Is All You Need.” In: *Proceedings of the 30th Annual Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 5998–6008.

[Vas+21] R. K. Vasudevan, M. Ziatdinov, L. Vlcek, and S. V. Kalinin. “Off-The-Shelf Deep Learning Is Not Enough, and Requires Parsimony, Bayesianity, and Causality.” In: *npj Computational Materials* 7.1 (2021), p. 16. ISSN: 2057-3960. DOI: 10.1038/s41524-020-00487-0.

[VC22] S. Vijayabaskaran and S. Cheng. “Navigation Task and Action Space Drive the Emergence of Egocentric and Allocentric Spatial Representations.” In: *PLOS Computational Biology* 18.10 (2022), e1010320. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010320.

[Vea20] R. M. Veatch. “Reconciling Lists of Principles in Bioethics.” In: *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 45.4-5 (2020), pp. 540–559. ISSN: 0360-5310. DOI: 10.1093/jmp/jhaa017.

[Vel20] A. Vellido. "The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care." In: *Neural Computing and Applications* 32.24 (2020), pp. 18069–18083. ISSN: 0941-0643. DOI: `10.1007/s00521-019-04051-w`.

[VF04] T. Verguts and W. Fias. "Representation of Number in Animals and Humans: A Neural Model." In: *Journal of Cognitive Neuroscience* 16.9 (2004), pp. 1493–1504. ISSN: 0898-929X. DOI: `10.1162/0898929042568497`.

[VFS05] T. Verguts, W. Fias, and M. Stevens. "A Model of Exact Small-Number Representation." In: *Psychonomic Bulletin & Review* 12 (2005), pp. 66–80. DOI: `10.3758/BF03196349`.

[Vin+22] Y. Vinker, E. Pajouheshgar, J. Y. Bo, R. C. Bachmann, A. H. Bermano, D. Cohen-Or, A. Zamir, and A. Shamir. "CLIPasso: Semantically-Aware Object Sketching." In: *ACM Transactions on Graphics (TOG)* 41.4 (2022), 86:1–86:11. DOI: `10.1145/3528223.3530068`.

[Von55] J. Von Neumann. "Method in the Physical Sciences." In: *Collected Works* 6 (1955), pp. 491–498.

[VR18] S. Verma and J. Rubin. "Fairness Definitions Explained." In: *Proceedings of the International Workshop on Software Fairness (FairWare)*. 2018, pp. 1–7. ISBN: 978-1-4503-5746-3. DOI: `10.1145/3194770.3194776`.

[VV82] M. P. Van Oeffelen and P. G. Vos. "A Probabilistic Model for the Discrimination of Visual Number." In: *Perception & Psychophysics* 32.2 (1982), pp. 163–170. ISSN: 0031-5117. DOI: `10.3758/BF03204275`.

[Wai+17] M. L. Wainberg, P. Scorza, J. M. Shultz, L. Helpman, J. J. Mootz, K. A. Johnson, Y. Neria, J.-M. E. Bradford, M. A. Oquendo, and M. R. Arbuckle. "Challenges and Opportunities in Global Mental Health: A Research-to-Practice Perspective." In: *Current Psychiatry Reports* 19 (2017), pp. 1–10. DOI: `10.1007/s11920-017-0780-z`.

[Wan+21] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik. "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization." In: *The Journal of Machine Learning Research* 22.1 (2021), pp. 9129–9201.

[Wan73] M. C. Wang. "Psychometric Studies in the Validation of an Early Learning Curriculum." In: *Child Development* 44.1 (1973), pp. 54–60. ISSN: 0009-3920. DOI: `10.2307/1127679`.

[Wat21] D. Watson. "Explaining Black Box Algorithms: Epistemological Challenges and Machine Learning Solutions." PhD thesis. University of Oxford, 2021.

[Wat94] W. Watt. "Curves as Angles." In: *Writing Systems and Cognition: Perspectives from Psychology, Physiology, Linguistics, and Semiotics*. Springer, 1994, pp. 215–246.

[WCB19] K. Wagner, J. Chu, and D. Barner. "Do Children's Number Words Begin Noisy?" In: *Developmental Science* 22.1 (2019), e12752. DOI: `10.1111/desc.12752`.

[WDN21] X. Wu, E. Dyer, and B. Neyshabur. "When Do Curricula Work?" In: *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. 2021.

[Wei07] M. Weisberg. "Three Kinds of Idealization." In: *The Journal of Philosophy* 104.12 (2007), pp. 639–659. ISSN: 0022362X.

[WF21] D. S. Watson and L. Floridi. "The Explanation Game: A Formal Framework for Interpretable Machine Learning." In: *Synthese* 198.10 (2021), pp. 9211–9242. ISSN: 0039-7857. DOI: 10.1007/s11229-020-02629-9.

[WFC87] H. M. Wellman, W. V. Fabricius, and W. Chuan-Wen. "Considering Every Available Instance: The Early Development of a Fundamental Problem Solving Skill." In: *International Journal of Behavioral Development* 10.4 (1987), pp. 485–500. DOI: 10.1177/016502548701000407.

[Wim87] W. C. Wimsatt. "False Models as Means to Truer Theories." In: *Neutral Models in Biology*. Ed. by M. H. Nitecki and A. Hoffman. Oxford University Press, 1987, pp. 23–55. ISBN: 978-0-19-505099-8.

[Win01] E. Winsberg. "Simulations, Models, and Theories: Complex Physical Systems and Their Representations." In: *Philosophy of Science* 68.S3 (2001), S442–S454. DOI: 10.1086/392927.

[Woo99] A. L. Woodward. "Infants' Ability to Distinguish Between Purposeful and Nonpurposeful Behaviors." In: *Infant Behavior and Development* 22.2 (1999), pp. 145–160. ISSN: 0163-6383. DOI: 10.1016/S0163-6383(99)00007-7.

[WP22] B. Wang and C. R. Ponce. "High-Performance Evolutionary Algorithms for Online Neuron Control." In: *Proceedings of the 23rd Genetic and Evolutionary Computation Conference (GECCO)*. 2022, pp. 1308–1316. DOI: 10.1145/3512290.3528725.

[WRB71] M. C. Wang, L. B. Resnick, and R. F. Boozer. "The Sequence of Development of Some Early Mathematics Behaviors." In: *Child Development* 42.6 (1971), p. 1767. ISSN: 0009-3920. DOI: 10.2307/1127583.

[WSB18] N. Weber, L. Shekhar, and N. Balasubramanian. "The Fine Line between Linguistic Generalization and Failure in Seq2Seq-Attention Models." In: *Generalization in the Age of Deep Learning Workshop @ NAACL*. 2018, pp. 24–27. DOI: 10.18653/v1/W18-1004.

[Wyn92] K. Wynn. "Children's Acquisition of the Number Words and the Counting System." In: *Cognitive Psychology* 24.2 (1992), pp. 220–251. DOI: 10.1016/0010-0285(92)90008-P.

[XS00] F. Xu and E. S. Spelke. "Large Number Discrimination in 6-Month-Old Infants." In: *Cognition* 74.1 (2000), B1–B11. ISSN: 0010-0277. DOI: 10.1016/S0010-0277(99)00066-9.

[Xu03] F. Xu. "Numerosity Discrimination in Infants: Evidence for Two Systems of Representations." In: *Cognition* 89.1 (2003), B15–B25. DOI: 10.1016/s0010-0277(03)00050-7.

[Yam+14] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. "Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex." In: *Proceedings of the National Academy of Sciences* 111.23 (2014), pp. 8619–8624. ISSN: 0027-8424. DOI: 10.1073/pnas.1403112111.

[YD16] D. L. K. Yamins and J. J. DiCarlo. "Using Goal-Driven Deep Learning Models to Understand Sensory Cortex." In: *Nature Neuroscience* 19.3 (2016), pp. 356–365. ISSN: 1097-6256. DOI: 10.1038/nn.4244.

[Yos+15] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. "Understanding Neural Networks through Deep Visualization." In: *Deep Learning Workshop @ ICML*. 2015.

[You+83] S. J. Youngner, D. L. Jackson, C. Coulton, B. W. Juknialis, and E. M. Smith. "A National Survey of Hospital Ethics Committees." In: *Critical Care Medicine* 11.11 (1983), pp. 902–905. ISSN: 0090-3493. DOI: 10.1097/00003246-198311000-00013.

[Yu+18] H. Yu, X. Lian, H. Zhang, and W. Xu. "Guided Feature Transformation (GFT): A Neural Language Grounding Module for Embodied Agents." In: *Proceedings of the 2nd Conference on Robot Learning (CoRL)*. Vol. 87. 2018, pp. 81–98.

[Yu+20] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi. "Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction." In: *Proceedings of the 16th European Conference on Computer Vision (ECCV)*. Springer. 2020, pp. 507–523.

[Yua+21] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani. "AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting." In: *Proceedings of the 18th International Conference on Computer Vision (ICCV)*. 2021, pp. 9813–9823. DOI: 10.1109/ICCV48922.2021.00967.

[YZX18] H. Yu, H. Zhang, and W. Xu. "Interactive Grounded Language Acquisition and Generalization in a 2D World." In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 2018.

[Zal20] "Scientific Research and Big Data." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Metaphysics Research Lab, Stanford University, 2020. URL: https://plato.stanford.edu/archives/sum2020/entries/science-big-data/ (visited on 04/29/2024).

[Zed21] C. Zednik. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." In: *Philosophy & Technology* 34.2 (2021), pp. 265–288. ISSN: 2210-5433. DOI: 10.1007/s13347-019-00382-7.

[Zer+19] J. Zerilli, A. Knott, J. Maclaurin, and C. Gavaghan. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" In: *Philosophy & Technology* 32.4 (2019), pp. 661–683. ISSN: 2210-5433. DOI: 10.1007/s13347-018-0330-6.

[Zha16] X. Zhang. "Linking Language, Visual-Spatial, and Executive Function Skills to Number Competence in Very Young Chinese Children." In: *Early Childhood Research Quarterly* 36 (2016), pp. 178–189. DOI: 10.1016/j.ecresq.2015.12.010.

[Zhi+22] T. Zhi-Xuan, N. Gothoskar, F. Pollok, D. Gutfreund, J. B. Tenenbaum, and V. K. Mansinghka. "Solving the Baby Intuitions Benchmark with a Hierarchically Bayesian Theory of Mind." In: *Robotics: Science and Systems Workshop on Social Intelligence in Humans and Robots*. 2022.

[Zho+22] J. Zhou, B. Lamichhane, D. Ben-Zeev, A. Campbell, and A. Sano. "Predicting Psychotic Relapse in Schizophrenia With Mobile Sensor Data: Routine Cluster Analysis." In: *JMIR mHealth and uHealth* 10.4 (2022). DOI: 10.2196/31006.

[Zip49] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, 1949.

[Zla+22] A. Zlatintsi, P. P. Filntisis, C. Garoufis, N. Efthymiou, P. Maragos, A. Menychtas, I. Maglogiannis, P. Tsanakas, T. Sounapoglou, E. Kalisperakis, T. Karantinos, M. Lazaridi, V. Garyfalli, A. Mantas, L. Mantonakis, and N. Smyrnis. "E-Prevention: Advanced Support System for Monitoring and Relapse Prevention in Patients with Psychotic Disorders Analyzing Long-Term Multimodal Data from Wearables and Video Captures." In: *Sensors* 22.19 (2022), p. 7544. DOI: 10.3390/S22197544.

[ZPU02] M. Zorzi, K. Priftis, and C. Umiltà. “Neglect Disrupts the Mental Number Line.” In: *Nature* 417.6885 (2002), pp. 138–139. ISSN: 0028-0836. DOI: 10.1038/417138a.

[ZZ18] Q. Zhang and S. Zhu. “Visual Interpretability for Deep Learning: A Survey.” In: *Frontiers of Information Technology & Electronic Engineering* 19.1 (2018), pp. 27–39. ISSN: 2095-9184. DOI: 10.1631/FITEE.1700808.