Towards a Conditional Theory of Abduction as a Foundation for Artificial Intelligence

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
der Ludwig-Maximilians-Universität München

vorgelegt von
Rolf Bernhard Pfister
aus
München
2025

Referent: Prof. Dr. Stephan Hartmann
Korreferent: Dr. Tom Sterkenburg
Tag der mündlichen Prüfung: 09.07.2025

Contents

Contents i

1 Introduction 1
- 1.1 The Significance of Artificial Intelligence and Its Philosophical Relevance . . . 1
- 1.2 Subject of Research . . . 5
- 1.3 Overview of Methods . . . 8
- 1.4 Structure of the Thesis . . . 9

2 A Representationalist, Functionalist and Naturalistic Conception of Intelligence as a Foundation for AGI 13
- 2.1 Introduction . . . 14
- 2.2 Skills & Intelligence . . . 15
- 2.3 Prediction & Assumption . . . 18
- 2.4 Perception & Experience . . . 19
- 2.5 Representation . . . 22
- 2.6 Phenomena & Appearances . . . 24
- 2.7 Meaning & Understanding . . . 30
- 2.8 World Model & Reasoning . . . 35
- 2.9 Viability & Construction . . . 39
- 2.10 Agentness & Interrelation . . . 41
- 2.11 Conclusion . . . 43

3 An Analysis of Benchmarking Intelligence Based on the Results of OpenAI's o3 on ARC-AGI 47
- 3.1 Introduction . . . 48
- 3.2 Suitability of ARC-AGI as a Benchmark for AGI . . . 49
- 3.3 Towards a new Benchmark for Intelligence . . . 53
- 3.4 Conclusion . . . 56

4 Towards a Theory of Abduction Based on Conditionals 59
- 4.1 Introduction . . . 59
- 4.2 Properties of Abductive Inferences . . . 62

4.3 Conditionals as the Basis of Abduction 70
4.4 Definition of Abduction 74
4.5 Types and Patterns of Abduction 77
4.6 Formalisation of Abductive Inferences 84
4.7 Conclusion 85

5 The Role of Overdetermination and Alternative Implication in the Evaluation of Conditionals 89
5.1 Introduction 89
5.2 Overview of Approaches to Conditionals 91
5.3 Evaluation of Conditionals with Several Mutually Exclusive Antecedents 96
5.4 Evaluation of Conditionals with Several Non-exclusive Antecedents . 100
5.5 Interpretation of the Evaluation Results 103
5.6 Examination of Promising Approaches to Conditionals 107
5.7 Conclusion 112

6 Conclusion 113

7 Summary of the Thesis in German 117

8 Bibliography 125

Chapter 1

Introduction

1.1 The Significance of Artificial Intelligence and Its Philosophical Relevance

Artificial intelligence (AI) has influenced the development of human societies more than almost any other topic in recent decades. It is applied in almost all areas of human life and performs a variety of tasks, many of which people are no longer aware of. This includes, for example, the control and optimisation of networks such as the electricity and water supply and traffic and transport systems. Artificial intelligence is also utilised in many other areas, such as fraud detection, cyber security, financial planning, supply chain management, and predictive maintenance of machines. Although many people have been familiar with the use of artificial intelligence for years, for example through the use of navigation tools, spam filters, recommendation services in online shops, photo filters, and automatic text correction, society's perception of artificial intelligence changed fundamentally with the release of OpenAI's ChatGPT, a generative AI model, in November 2022.

For this, two main factors were decisive. First, ChatGPT – as well as other Large Language Models and Large Multimodal Models – is for the first time no longer a specialised artificial intelligence model that serves only one specific purpose and that can only perform one specific type of task. Instead, the models can answer questions on a wide range of different topics, translate texts, generate a variety of different text types, draw pictures, create videos, and generate code for programming. This allows a single tool to be used in many different areas to provide support for a wide range of everyday digital tasks. Second, most artificial intelligence approaches are functionally designed to automate a task as efficiently as possible. Accordingly, interaction with users was often kept to a minimum, and many applications were only used by experts in their respective fields. ChatGPT and other models, however, are explicitly designed for communication with users and have been humanised accordingly. For example, they welcome users, react emotionally, apologise, and copy

many human behaviours, such as showing joy and curiosity. In addition, interaction takes the form of continuous communication via chat, equivalent to how people often interact with other people on social networks. This imitated human behaviour leads many people to anthropomorphise Large Language Models and perceive them not as a programme but as a being; this can be seen, for example, in the fact that many people say please and thank you when communicating with them.
Both factors together caused many people to feel that they were interacting with another intelligence. In addition to massive economic investments in the further development and application of generative AI, which includes large language models, the models also triggered many social reactions. This includes many positive expectations, such as that the models will automate a large number of human tasks in the future and thus take over tedious work, but also negative fears, such as that the models could lead to job losses, increase social injustice or even become an existential threat to humanity, either through the destructive use of humans or because the models themselves would strive for power.
Despite the extensive expectations and fears, a more in-depth social discussion failed to materialise overall. There are several reasons for this. On the one hand, there were relatively few contributions from professional disciplines that could contribute to the topic, not least because there has only been limited research to date. In particular, practical philosophy should be mentioned here, which could provide not only well-founded insights into ethical aspects but also reflections on what future forms of society in which artificial intelligence plays a fundamental role could look like. On the other hand, there is a great deal of uncertainty, particularly among the general public, but also in the field of artificial intelligence, about what artificial intelligence precisely is, whether it is in fact intelligent, and whether it can or even already possesses consciousness, emotions, will, and other things. For example, Jakob Uszkoreit, one of the authors of the Attention Is All You Need publication, which provided the foundation for the development of generative AI, stated in a personal conversation that he did not know what intelligence was and that he did not consider this question to be important as he was concerned with solving specific challenges.
One reason for the lack of clarity on the nature of artificial intelligence is due to the expression itself. The term, introduced at the Dartmouth Conference in 1956, primarily refers to the automated processing of data; an expression that is much more precise and allows for much less speculation. In the course of the development of the field of artificial intelligence, the term Artificial General Intelligence (AGI) was introduced, which is primarily intended to differentiate it from AI approaches that have the purpose of solving a specific task. AGI generally refers to an artificial intelligence that can solve a variety of different tasks, just as humans can. Numerous other terms have been introduced for differentiation and specification, such as human-level artificial intelligence (HLAI), which is strongly orientated to-

wards human intelligence, and superhuman artificial intelligence (SAI), also known as superintelligence, which describes an artificial intelligence whose capabilities significantly exceed those of humans. However, defining and differentiating the terms is difficult as there are no generally recognised definitions and different representatives of the field of artificial intelligence use the terms differently. A more precise description of the term Artificial General Intelligence (AGI) will be provided in the course of this work.

For a better understanding of artificial intelligence, some of the main events that contributed to its development are described in the following. The first developments in automated data processing can be traced back to the 17th to 19th centuries, when several mechanical machines were developed. Blaise Pascal, for example, developed a machine for addition in 1642, while Wilhem Leibniz developed a machine that can also multiply and divide in 1673. The first control of machines using punched cards was carried out in 1801 by Joseph-Marie Jacquard. In 1837, Charles Babbage developed the concept for a programmable machine with a basic structure consisting of an arithmetic unit, a memory, and a control unit, which corresponds to today's computers.

In the second half of the nineteenth century, developments occurred particularly in the field of logic, which is fundamental to automated data processing. In addition to the publication of the algebraic logic calculus by George Boole in 1847, Gottlob Frege published the basis for modern predicate logic in 1879. Further fundamental publications followed in the first half of the 20th century by Bertrand Russell, Kurt Gödel, and Alan Turing, among others. The latter developed the Turing machine, a model that is fundamental to defining algorithms and their computability. At the same time, the first electromechanical and electronic computers were developed, some of which could be programmed in different ways. In the second half of the twentieth century, fundamental progress was made in the development of microchips, which massively increased computing performance and greatly reduced the cost of computing. At the same time, numerous programming languages and other concepts such as database structures were introduced, which enabled efficient programming and thus the realisation of automated data processing.

Automated data processing was performed using numerous different methods. Initially, symbolic approaches, in which symbolic representations were processed according to explicit rules, were used in particular, as these could be performed with low computing capacities. In the 1960s, neuronal approaches were developed that are based on highly simplified neuron models and form artificial neuronal networks by connecting a large number of artificial neurons in different configurations. However, initial successes led to great expectations and, when these could not be met, to the first so-called AI winter in the 1970s. In the 1980s, further developments were made in the field of artificial neural networks, leading to new successes. Significant contributions were made among others by John Hopfield and Geoffrey Hinton, who

both received the Nobel Prize in Physics in 2024 for this.
Due to technical advances, but also due to the constant conceptual development of artificial neural networks, there has been continuous progress in the application of these networks since the 2000s, which are summarised under the term deep learning due to their size. The developments allow numerous applications; for example convolutional neural networks are particularly successful in image recognition, and generative adversarial networks consist of two mutually improving neural networks. In 2017, the Attention Is All You Need publication introduced the Transformer architecture, which led to the development of generative AI and thus of large multimodal models, which are currently considered the most powerful approaches in the field of artificial intelligence. These conceptual developments were accompanied by extensive progress and investment in the area of hardware, which led to the availability of enormous computing capacities, and without which the developments would not have been possible. The continuous digitalisation of human societies generates huge amounts of data, which also play an important role, particularly for training models.
Despite the numerous advances and the large multimodal models, which are considered impressive by many, the field of artificial intelligence is still far from its actual goal: artificial general intelligence, i.e., a system of automated data processing that can cope at least as well as an average human being in the human world and can perform all type of tasks that arise there. Although a large number of superhuman successes have been achieved in specific tasks, there is no system that is capable of solving such a wide range of tasks as humans can. This is despite the fact that the development of artificial intelligence in general is the overarching goal of the field of artificial intelligence, which has been communicated by numerous representatives since its establishment. Also currently, artificial general intelligence, or variations of it, is widely pursued and its imminent arrival frequently announced; for example, by OpenAI with the releases of ChatGPT and newer models, as well as by Elon Musk, who has been promising the near availability of self-driving cars for many years. This raises the questions of exactly what artificial general intelligence is and how it can be created.
This thesis aims to analyse this problem from a philosophical perspective, incorporating both methods and insights from philosophy, in particular from the philosophy of science. The philosophical examination of artificial intelligence offers numerous mutual advantages for both the field of artificial intelligence and philosophy.
For example, a fundamental element of artificial intelligence is logic, which originates from philosophy and which continues to be of significant importance in philosophy. The knowledge gained in philosophy can be transferred to the field of artificial intelligence to investigate, among other things, the theoretical limitations of artificial intelligence and how different methods of knowledge generation can be formalised. Philosophy has been concerned for thousands of years with the question of how human reasoning works and how humans acquire knowledge and insights – abilities

that are central to the creation of artificial intelligence, which also processes data in order to gain insights from it. Philosophy already has an extensive collection of methods for gaining knowledge, the transferability of which to artificial intelligence can be investigated; in particular, abduction should be mentioned here, which enables the creation of new concepts, for example theoretical concepts such as gravity, and which will play an important role in this thesis. Furthermore, philosophy of science in particular is concerned with the question of how research can be carried out so that it is successful and leads to a gain in insight; a question that is also relevant to the field of artificial intelligence with regard to the path to the development of artificial general intelligence. This also includes the application of classical methods such as conceptual analysis, for example, to determine what intelligence is and how it can be measured – both topics that are examined in the course of the thesis.
Conversely, artificial intelligence also offers numerous advantages for philosophy. Philosophy can benefit from the formalisation – and thus the explication – of the methods it employs. For example, the concept of abduction has so far only been defined to a limited extent, and there are only a few, limited approaches to formalising it; these originate from the field of artificial intelligence. Furthermore, artificial intelligence is not only a means that philosophy can use for its own research. Artificial intelligence also creates a completely new field of research and allows existing research topics to be explored in greater depth. Examples of this include the question of whether artificial intelligence requires consciousness, a question that is also addressed in this thesis. In addition, artificial intelligence raises many other questions that are relevant from a philosophical perspective but are not discussed in this thesis. This includes, for example, the assessment of the fairness and performance of algorithms, the ethical aspects mentioned above, and the question of what consequences the realisation of artificial general intelligence will have for humans, both in terms of society and, in particular, the humanities.
Overall, artificial intelligence plays an important role in society and its importance is expected to continue to grow. Philosophy has the possibility – and the responsibility – to contribute to this development and to provide its own insights to ensure that the development is as positive as possible from a societal perspective. The aim of this work is to fulfil this purpose and to contribute to the positive development of artificial intelligence with the help of the findings and methods of philosophy.

1.2 Subject of Research

The fundamental problem the thesis is orientated towards is the question of how artificial general intelligence can be created. However, this research question is of a very comprehensive nature, demonstrated not only by the lack of a solution in recent decades despite great efforts but also by the numerous works that have been published to answer it and which have led to progress, but not to success. This

work therefore restricts the examination of the problem in several respects. First, it is specifically concerned with analysing the problem from a philosophical perspective and applying philosophical methods and insights. Insights from various other disciplines, including mathematics and biology, are taken into account, but serve merely as a means. Second, the thesis is concerned only with the investigation of fundamental principles that need to be considered for the creation of artificial general intelligence; the thesis does not address the practical implementation or specific AI procedures, unless these are relevant to the investigation. Third, the aim of the work is not to provide an all-encompassing answer to the topic. Instead, the work begins by clarifying the necessary fundamental questions, in particular what intelligence is and how it can be measured. Subsequently, the work is orientated towards a cross-sectional examination of the topic of how artificial general intelligence can be created. In this respect, the work examines a detailed aspect of intelligence, the inference method abduction, and then again a detailed aspect of this method, the evaluation of different theories of conditionals.

With this cross-sectional study, two objectives are being pursued. First, the aim is to show that the investigation and use of philosophical methods in the field of artificial intelligence enriches both artificial intelligence and philosophy and leads to a mutual gain in knowledge. Second, the aim is to develop an approach that shows how abduction can be used in the field of artificial intelligence, which can lead to the development of new approaches to artificial intelligence that are able to introduce new theoretical concepts in a controlled way and thus to develop more powerful theories. Due to the complexity of the topic, the thesis does not provide a complete elaboration of how abduction can be successfully implemented in the field of artificial intelligence. However, it shows the general feasibility, elaborates the necessary foundations, and shows which further research work is necessary for implementation. The thesis thus aims to enable the development of an important component in the realisation of intelligence and thereby to support the development of artificial general intelligence.

On the basis of this structure, the work covers four different, yet interrelated main topics: the nature of intelligence, the measurement of intelligence, the inference method abduction, and theories of conditionals. Each of these topics is presented in the following with a brief overview.

The nature of intelligence is still an unresolved issue, despite the fact that intelligence plays an important role in human life as, for example, success is correlated with intelligence. Intelligence also plays a fundamental role in the field of artificial intelligence, although, as shown above, there is no uniform understanding of it and the term is sometimes used less as a characterisation and more as a phrase. From a psychological perspective, human intelligence can be measured using IQ tests, which analyse how well people perform in different areas of thinking, for example in solving logical or mathematical tasks. However, the informative value is limited

and IQ tests are not suitable for measuring the intelligence of algorithms, as they can give the answers to the questions without understanding them. Another aspect is already apparent here: It is unclear what intelligence is and which associated characteristics – such as understanding, consciousness, intentionality – are part of or necessary for intelligence. The definition of intelligence is also difficult in other respects. Although there are numerous definitions, no generally recognised definition has yet been agreed. This is partly due to the fact that some people understand intelligence as the ability to solve specific tasks, while others understand it as the ability to develop new solution approaches.
The problem that the nature of intelligence has not yet been determined also leads to problems in the measurement of intelligence. This is illustrated by the fact that in the past, achieving various tasks, such as winning at chess or Go, or writing texts, was regarded as proof of intelligence. However, whenever an AI programme was able to solve these tasks, it was not considered intelligent for various reasons, for example because it tried many different solutions or because the solution was not provided by the AI approach but by the programmer and was only implemented by the approach. This is not just an abstract, theoretical problem, as the current successes of large language models show. A large number of tests are successfully solved by them and the developers point out, for instance, that in some tests the models exceed the level of post-docs or of medical doctors with extensive professional experience. This raises the socially relevant question of whether the tests are informative and the models actually demonstrate a high human-level peformance, or whether the models are successful in the tests but still fail the tasks in practice and also do not exhibit intelligence. The development of tests to measure the intelligence of AI approaches is still in its early stages from a research perspective, despite its great importance for the field of artificial intelligence, and there are still many unanswered questions. This concerns not only the nature of intelligence, but also, for example, how it can be quantified and how a human-biased view can be avoided.
Abduction is an inference method, i.e., a method that allows new knowledge to be derived from existing knowledge. Besides abduction, additional inference methods are deduction and induction. Deduction allows the inference of certain conclusions; for example, from the knowledge that every swan is white, it can be certainly inferred that the next swan to be seen will also be white. The truth of the conclusion is, however, only given if the premises are true. Induction allows the generalisation of statements. For instance, if all the squirrels one sees are brown, one can conclude that all squirrels that exist on Earth are brown. Generalisations are uncertain as there may be, for instance, other areas where squirrels have a different colour. Abduction allows one to infer from a given fact, such as an observation, to a fact that implies it, e.g., a cause. For instance, it can be inferred from the observation that the road is wet that it has recently rained. As with induction, abduction is uncertain and there are often several possible causes. Abduction is also the most powerful method

of inference, as it allows the introduction of new theoretical concepts. For instance, from the observation of falling apples, the concept of gravity can be inferred. Due to its complexity, abduction is the least researched inference method so far. Among others, it is unclear to what extent the generation of new hypotheses is rule-based and can therefore be formalised, and to what extent it is intuitive and cannot be captured in the form of a scientific method accordingly. It is also not clear yet how abductive hypotheses can be justified and, if there are several, how the best one can be selected.
Conditionals have an important function in science, which is shown not only by the fact that abduction is based on them, but also, for example, by the fact that many laws are presented in the form of conditionals. Although conditionals are simple in their if-then-structure, they possess a high degree of complexity and many questions remain unanswered. Among others, there are different approaches on how to categorise the relation between the condition and the consequence. Some theories assume, for instance, a purely probability-based relation, while others assume a relevance relation, which may be argumentative or causal in nature. Overall, there are a large number of different theories on conditionals, all of which have different strengths and weaknesses. In addition, numerous studies have shown that people use conditionals in different ways and evaluate their degree of truth and acceptability differently. Additional complexity arises from the fact that there are many different types of conditionals, for example subjunctive conditionals, as well as so-called biscuit conditionals ("If you like biscuits, they are in the kitchen."), which do not express a condition.

1.3 Overview of Methods

To achieve these goals, the thesis follows a functional approach, i.e. it does not apply a specific philosophical method but applies the methods that are considered most promising. In consequence, the work draws on many different methods, some of which are described in more detail in the following. The investigation of the nature of intelligence in Chapter 2 involves the analysis and definition of concepts, the application of thought experiments, the critical examination, explication and comparison of arguments, and the consideration of the assumptions underlying the various positions. In addition, a phenomenological analysis as well as an analysis of phenomenology itself are performed to examine different methodical approaches in the creation of artificial intelligence.
In the examination of how intelligence can be measured in Chapter 3, the focus is on current developments in the field of artificial intelligence. Although philosophy is more concerned with fundamental and, therefore, more theoretical issues, current developments are important not only insofar as they offer new insights for philosophy, but also because philosophy can provide valuable contributions to their further

advancement. For this reason, the chapter analyses OpenAI's generative AI model o3, which had just been released at the time of writing, and investigates whether it represents a form of AGI, as claimed by the company – a question that is highly relevant not only from a philosophical point of view, but also from a social perspective. The chapter also discusses the validity of a benchmark, which has received a lot of attention in the field of artificial intelligence. In this way, a contribution can be made to its improvement and to the development of new, better benchmarks by the application of philosophical methods.
In the development of a theory of abduction based on conditionals in Chapter 4, a variety of methods are applied to achieve this goal. Existing theories are critically reviewed and checked for coherence, evaluated using examples, and compared with each other. Argumentation procedures are systematised and formalised to achieve greater clarity. In addition, an evaluation of the own and other approaches is performed on the basis of a historical case to test the empirical coherence and applicability of the theories.
Chapter 5 finally focuses on the evaluation of different theories of conditionals on the basis of a specific type of conditional. This comparative approach includes analysing various concrete examples of conditionals, as well as analysing the underlying assumptions of the different theories and the extent to which they are suitable for application in more complex cases. For the comparison of the different theories, they are systematised and explicated, and their coherence and consequences are examined.
Overall, the work employs a larger variety of methods with the aim of maximising the potential gain in findings. Although the work is in many cases critical of existing concepts, arguments and theories, it has to be emphasised that the work is fundamentally based on these and that it was only possible to develop the own thoughts and insights by engaging with them. The analysis of the various contents has therefore always to be understood not only as a criticism but also as an appreciation.

1.4 Structure of the Thesis

The aim of the work is to develop an approach that allows to introduce abduction as an inference method in the field of artificial intelligence to support the development of AGI. Based on this goal and taking into account the previously discussed considerations, such as the unclarity of what exactly intelligence is, the work is based on the following structure: Chapter 2 analyses the nature of intelligence and the foundational principles that have to be considered for the creation of AGI. Chapter 3 examines the performance of OpenAI's generative AI model o3 on the ARC-AGI benchmark to evaluate the validity of the benchmark and to outline a new benchmark for measuring intelligence based on the understanding of intelligence devel-

oped in the preceding chapter. Chapter 4 analyses existing theories of the inference method abduction and develops a new theory of abduction based on conditionals. Chapter 5 evaluates different theories of conditionals based on a specific type of conditionals and identifies two approaches that come to the correct results. Chapter 6 presents an overview of the insights gained in the thesis, and Chapter 7 offers a summary of the work in German.

In detail, Chapter 2 addresses the following aspects: Section 2.1 discusses current developments in the field of artificial intelligence and highlights the underlying problem that while AI approaches can solve specific tasks, they are not capable of generalisation. Section 2.2 analyses different conceptions of intelligence and concludes that intelligence is the ability to create novel skills that allow one to achieve goals under previously unknown conditions. Section 2.3 discusses the role of prediction and the necessity for intelligence to be based on assumptions about the world in which it is to be applied. Section 2.4 is concerned with perception, its indirect and representational nature, and its distinction from conscious experience. Section 2.5 examines the nature of representations and shows that they are an inherent aspect of grasping a world to determine goal-directed actions. Section 2.6 explores how a world is grasped and, based on the phenomenological approaches of Heidegger and others, outlines the dichotomy between a world itself and the interpreted conception of it. Section 2.7 analyses the conceptions of meaning and understanding and argues for a functional definition of them, which allows for a naturalistic interpretation of intelligence that does not require assumptions of mental features such as consciousness. Section 2.8 describes how intelligence utilises reasoning methods such as deduction, induction and abduction, as well as abstraction and classification for the development of world models. Section 2.9 discusses the assessment of world models on the basis of their functional usefulness, i.e., viability, rather than their depiction of truth, and discusses their constructivist character, which results from the uncertainty and contingency of the reasoning methods. Section 2.10 addresses the subjective perspective through which an agent perceives a world and examines the numerous interrelations between an agent and the rest of the world. Section 2.11 concludes with an overview of the approach developed in the article, outlining the foundational characteristics that have to be considered to enable the creation of AGI.

Chapter 3 builds directly on the preceding chapter and examines the following topics in detail: Section 3.1 discusses the recent success of OpenAI's AI model o3 on the ARC-AGI benchmark and introduces the benchmark in its structure and in terms of its successes to date. Section 3.2 analyses the suitability of ARC-AGI as a benchmark for intelligence and for measuring progress towards AGI. This includes an analysis of the type of problem structure that ARC-AGI tasks represent, as well as the weaknesses that ARC-AGI possesses and the extent to which these can be overcome. Section 3.3 outlines a new benchmark for intelligence that is based on the

definition of intelligence introduced in Section 2.2 and which is intended to enable a more comprehensive assessment of intelligence. Section 3.4 concludes with an evaluation of the performance of OpenAI's o3 on ARC-AGI.
Chapter 4 examines abduction in the following way: Section 4.1 provides an introduction to the central ideas of abduction and provides a historical outline that shows the current state of development. Section 4.2 examines various important properties of abduction based on an analysis of Peirce's retroduction and Inference to the Best Explanation. Section 4.3 offers a discussion of conditionals and, in particular, inferentialism. Building on all this, a definition of abductive inferences founded on conditionals is given in Section 4.4. The different types of abductive inferences are discussed in Section 4.5, in which moreover the use of analogies in patterns is explored. Section 4.6 examines the conditions under which abductive inferences can be formalised, and finally a conclusion is drawn in Section 4.7.
Chapter 5 considers the following aspects in the evaluation of conditional theories: Section 5.1 highlights fundamental differences between various conditional theories and defines which types of conditionals are analysed subsequently. Section 5.2 offers an overview of recently and widely discussed approaches to conditionals. Section 5.3 provides an analysis of the various approaches on conditionals whose consequents are implied by several mutually exclusive and exhaustive antecedents. Section 5.4 presents an analysis of the various approaches on conditionals whose consequents are implied by several non-exclusive antecedents. Section 5.5 discusses how the conditionals from sections 5.3 and 5.4 are ideally evaluated and compares this with the actual results. Section 5.6 examines the most promising approaches to conditionals in this respect in more detail for their general applicability.
With regard to the structure of the work, there are two aspects to be noted. First, due to the cross-sectional approach, the chapters build on each other, and in the later chapters, specific aspects of the previous chapters are explored in greater depth. Chapter 3, which is concerned with the measurement of intelligence, is directly based on Chapter 2, which analyses the nature of intelligence, thus deepening a specific aspect that is relevant to intelligence. Equally, Chapter 4 on abduction is related to Chapter 2 in that it examines and develops an important component of intelligence in detail by analysing an inference method. Chapter 5, which focuses on the investigation of theories of conditionals, in turn addresses a specific aspect of Chapter 4, as conditionals are an inherent aspect of the theory of abduction developed there.
Second, the publications to which the chapters are related were created over the course of several years and not in the order presented here. Instead, the publication on abduction from Chapter 4 was created first, next the publication on conditionals from Chapter 5, and subsequently the publication on the measurement of intelligence from Chapter 3 as well as the publication on nature of intelligence from Chapter 2. As a result, the articles are compatible with each other, but differ in some details.

The chapter on intelligence, for example, develops a constructivist understanding of the world, while the chapter on abduction (still) advocates a truth-based understanding of the world. However, this does not constitute an incompatibility, as the approach of abduction can also be applied with a constructivist understanding of the world.

Chapter 2

A Representationalist, Functionalist and Naturalistic Conception of Intelligence as a Foundation for AGI[1]

The article analyses foundational principles relevant to the creation of artificial general intelligence (AGI). Intelligence is understood as the ability to create novel skills that allow to achieve goals under previously unknown conditions. To this end, intelligence utilises reasoning methods such as deduction, induction and abduction as well as other methods such as abstraction and classification to develop a world model. The methods are applied to indirect and incomplete representations of the world, which are obtained through perception, for example, and which do not depict the world but only correspond to it. Due to these limitations and the uncertain and contingent nature of reasoning, the world model is constructivist. Its value is functionally determined by its viability, i.e., its potential to achieve the desired goals. In consequence, meaning is assigned to representations by attributing them a function that makes it possible to achieve a goal. This representational and functional conception of intelligence enables a naturalistic interpretation that does not presuppose mental features, such as intentionality and consciousness, which are regarded as independent of intelligence. Based on a phenomenological analysis, it is shown that AGI can gain a more fundamental access to the world than humans, although it is limited by the No Free Lunch theorems, which require assumptions to be made.

[1] This chapter will be submitted as an article for publication in Mind & Machines (Pfister, 2025b).

2.1 Introduction

In recent years, extensive developments have taken place in the field of artificial intelligence (AI). These include in particular generative AI approaches that use transformer or diffusion architectures and lead to contributions in many areas such as text and image generation (Touvron et al., 2023), protein structure prediction (Abramson et al., 2024) and autonomous driving (Seff et al., 2023). However, although these approaches achieve results that are considered impressive, they are unreliable and fail in many tasks that appear simple from a human perspective (Nezhurina et al., 2024; Dziri et al., 2023; Berglund et al., 2023). They also fail the more frequently the less similar the tasks are to those on which they were trained (McCoy et al., 2023; Wu et al., 2023). Such weaknesses do not occur only in specific approaches, but constitute a general problem in the field of AI (Dohare et al., 2024; Shanahan and Mitchell, 2022).

As a consequence, AI applications can be used reliably in specific, controlled domains for which they have been designed and evaluated. But AI applications often fail in more complex and practical tasks in which uncertainties occur; for instance, in autonomous driving (Suk et al., 2024; Cummings and Bauchwitz, 2024). Currently, there is no artificial general intelligence (AGI), i.e., AI models that can solve a wide range of everyday tasks as reliably as humans can (Mitchell, 2021). The development of AGI is considered a desirable goal, as AGI could relieve humans of tasks they do not want to perform. Furthermore, with AGI, a single AI model could be used for all types of tasks instead of having to develop a separate model for each specific use case, as at present.

The aim of the article is to identify and analyse principles that have to be considered for the creation of AGI. The analysis focuses in particular on understanding intelligence and how AGI can perceive and interpret a world in such a way that it can reliably fulfil a wide range of goals. The analysis is not about the evaluation of a specific AI approach such as symbolic, embodied or generative AI, but about the foundational characteristics of AGI.

Section 2.2 analyses different conceptions of intelligence and concludes that intelligence is the ability to create novel skills that allow one to achieve goals under previously unknown conditions. Section 2.3 discusses the role of prediction and the necessity for intelligence to be based on assumptions about the world in which it is to be applied. Section 2.4 is concerned with perception, its indirect and representational nature, and its distinction from conscious experience. Section 2.5 examines the nature of representations and shows that they are an inherent aspect of grasping a world to determine goal-directed actions. Section 2.6 explores how a world is grasped and, based on the phenomenological approaches of Heidegger and others, outlines the dichotomy between a world itself and the interpreted conception of it. Section 2.7 analyses the conceptions of meaning and understanding and argues

for a functional definition of them, which allows for a naturalistic interpretation of intelligence that does not require assumptions of mental features such as consciousness. Section 2.8 describes how intelligence utilises reasoning methods such as deduction, induction and abduction, as well as abstraction and classification for the development of world models. Section 2.9 discusses the assessment of world models on the basis of their functional usefulness, i.e., viability, rather than their depiction of truth, and discusses their constructivist character, which results from the uncertainty and contingency of the reasoning methods. Section 2.10 addresses the subjective perspective through which an agent perceives a world and examines the numerous interrelations between an agent and the rest of the world. Section 2.11 concludes with an overview of the approach developed in the article, outlining the foundational characteristics that have to be considered to enable the creation of AGI.

2.2 Skills & Intelligence

For the development of AGI, it is important to understand its nature precisely. This includes in particular the concept of intelligence. Human intelligence is explained by the Cattell-Horn-Carroll theory as an interaction between crystallised intelligence and fluid intelligence (Schneider and McGrew, 2018, pp. 73-75): Crystallised intelligence consists of several broad cognitive abilities, such as reasoning, processing visual information, and remembering information. Fluid intelligence is a general ability whose performance affects all broad abilities and describes the general cognitive capacity. In the field of AI, a variety of definitions of intelligence are used (Legg et al., 2007), which can be broadly categorised into two groups: Process-oriented definitions name required abilities such as learning, abstraction, logical thinking, and problem solving. Result-oriented definitions focus on the outcome and define intelligence as the ability to achieve specific goals; for instance, to adjust to an environment, to create products, or to grasp truths.
To determine whether an AI approach is intelligent, it is usually tested on tasks that fulfil the requirements of the definitions. In the course of the history of AI, numerous tasks whose solutions were assumed to require extensive cognitive abilities, and therefore intelligence, were proposed. The proposed tasks included for example playing chess, playing Go, image recognition, translating texts, or creating meaningful texts. However, when AI approaches were able to solve any of the problems, they were considered not intelligent. One reason for this is that the methods used by the approaches to solve a task, for example trying out a large number of possibilities, are not considered intelligent. It is also argued that the tasks are not solved by the intelligence of the AI approaches but by the intelligence of the programmers embedded in the approaches. Moreover, it is argued that an approach cannot be intelligent if it can solve a task but fails if the task is modified; a problem

that concerns many approaches. This leads some to conclude that AI approaches are making major progress in terms of performance but not in terms of intelligence (Hernández-Orallo, 2017, pp. 396-404, 421-423, 434; Chollet, 2019, pp. 7-9, 16f).
Chollet (2019, pp. 3-7) explains this contradictory development by the fact that two different interpretations of intelligence are used and that they are not distinguished sufficiently clearly. The first interpretation understands intelligence as a collection of task-specific skills, as advocated by Darwin and Minsky, for example. The second interpretation understands intelligence as the ability to create novel skills for solving tasks, as advocated by Turing and McCarthy, among others. Accordingly, while the first interpretation classifies solving tasks known to an AI approach as intelligent, the second interpretation classifies solving tasks hitherto unknown to an approach as intelligent. Chollet (2019, pp. 18-20) argues that the first interpretation of intelligence as task-specific skills is misleading because it does not describe intelligence but only its output: Skills are specific solutions to specific problems that are created by intelligence but that are not intelligence itself. In contrast, the second interpretation describes intelligence as a process, as an ability that creates skills.
A further reason in favour of the second interpretation of intelligence is that only that one is suitable for the development of AGI. This, as skills can be applied to specific tasks for which they were created, i.e., tasks that are known and well-defined, such as mastering games. But skills cannot be reliably applied to tasks outside the well-defined domain for which they were created: Skills do not include specifications on how to handle unfamiliar conditions[2] that occur outside the well-defined domain. Everyday tasks from the human domain, which AGI is supposed to solve, often have unfamiliar conditions: The future development of the world is only partially predictable for humans – and thus also for skill-based AI approaches created by humans – and future conditions remain partially unknown. Accordingly, AGI cannot be realised via a skills-based approach, as it would not be able to handle the constantly arising new, unknown conditions. Instead, AGI must be able to create novel skills to cover the unknown conditions, i.e., AGI must be able to fulfil the second interpretation of intelligence.
The foregoing considerations allow for a more precise definition of skill and intelligence: A skill is the ability to achieve a specific goal under specific known conditions. Intelligence is the ability to create novel skills that allow to achieve goals under previously unknown conditions. As such, intelligence is also a skill: it is a skill that allows to create other skills. Intelligence is not a fixed ability that is only either present or absent, but one that can also be stronger or weaker: An agent is the more intelligent, the more efficiently it can achieve the more diverse goals in the

[2]In the field of AI, conditions are often called states. The two terms are used interchangeably in this article.

more diverse worlds[3] with the less knowledge. Knowledge is understood pragmatically here: It does not have to be true statements about the worlds, but it includes all the information the agent has, including skills. The negative consideration of knowledge in the definition of intelligence entails that only the ability to generate skills but not skills themselves falls under intelligence. The definition thus corresponds to the second interpretation of intelligence discussed by Chollet above and excludes the first interpretation. Simply put, intelligence describes how well an agent can achieve goals in novel, unknown conditions.

The juxtaposition of the application of existing skills on the one hand, and the generation of skills, i.e., intelligence, on the other, reveals a fundamental relationship between the two: Tasks can be solved either by skills or by intelligence. This means skills and intelligence can be substituted for each other, provided that all conditions are known. Intelligence is only necessary to the extent conditions are unknown or skills are not available for other reasons; for example, because skills cannot be provided for all possible known conditions. Beyond that, the assessment of the degree of intelligence is abstract in that it does not permit a quantitative assessment without further specification of how this is to be carried out. For example, the assessment does not describe how exactly efficiency or diversity are quantified, or how the individual factors are weighed against each other. However, the provision of such specifications is not necessary for the further course of the article. Chollet (2019, pp. 27-42), who provides a measurable definition of intelligence, states that many possible ways of measuring intelligence may be valid. Which specific quantitative valuation is the best requires further research and may depend on epistemic as well as ontological assumptions.

The above definition describes intelligence as an ability of an agent. An agent is defined in this article as a system that is able to perform specific actions depending on specific conditions to achieve specific goals. Understanding AI approaches as agents is a fundamental perspective within the field of AI (Russell and Norvig, 2022, pp. 7, 21f). With regard to AGI, the aim is to develop an AI agent that is intelligent, i.e., an agent that is able to fulfil goals under partially unknown conditions. The goals are specified by the creator of the agent, i.e., by humans. They can be of a more specific nature, such as controlling a vehicle, or of a more general nature, such as developing scientific theories. For an agent, skills provide specifications under which conditions which actions are appropriate to achieve a specific goal; for example, in which chess position which move is appropriate to win the game. As shown above, if an agent encounters conditions that are at least partially not covered by skills, the specifications provided may be insufficient to achieve its goals; this, because it is uncertain how the uncovered conditions will affect the achievement. Consequently,

[3]In this definition, a world is seen as a self-contained and independent system that can have different conditions, some of which are accessible to the agent and some of which may be manipulable by the agent. Instances of individual worlds are the universe in which humanity is situated, games such as Go and computer games, and mathematical and logical systems.

the agent must utilise intelligence to create a skill, i.e., provide the specifications on how to achieve its goals under the unknown conditions. To determine which possible actions are appropriate, the agent must determine how they affect the achievement of the goal. This means that the agent has to make a prediction: It has to determine how a specific action influences the achievement of its goals without performing the action.

2.3 Prediction & Assumption

A prediction is a specification of unknown conditions. Conditions can be unknown to an agent, for example, because they occur in the future or because the agent cannot perceive them for other reasons. To be successful, a prediction requires knowledge of the world, i.e., of some of its conditions. Furthermore, a prediction requires knowledge of how the conditions of the world develop; i.e., it requires a model of the world that describes the development of the conditions to be predicted on the basis of the current conditions of the world. The applicability of such a world model requires that the world is subject to at least some regularities. If all conditions of a world were irregular, for example because they were completely random, there would be no regularities that could be part of the world model and used to specify unknown conditions. Consequently, predictions – and therefore intelligence – can succeed only in worlds that exhibit at least some regularities (cf. Ma et al., 2022, pp. 1300f).

The No Free Lunch (NFL) theorems show that across all possible optimisation problems any algorithm has the same average performance as every other. Consequently, there is no algorithm that is better than others at solving all optimisation problems: If an algorithm performs better than another on one set of optimisation problems, it performs worse than the other on the set of all other optimisation problems (Wolpert and Macready, 1997, pp. 69-71; Wolpert, 2013, pp. 4f). This can be seen as a counterargument to the formalisation of intelligence: Intelligence is about solving unknown optimisation problems with above-average performance, but the NFL theorems indicate that there cannot be such an algorithm. However, as shown above, intelligence can only be beneficial in worlds that have at least some regularities. This means that intelligence does not have to be adapted for all possible optimisation problems but only for the subset of optimisation problems that occur in worlds with regularities (cf. Hernández-Orallo, 2017, pp. 402f). Consequently, it is possible to find an algorithm that performs better than others on this subset of problems – and worse on the remaining optimisation problems of completely irregular worlds.

For an algorithm to be better than others on a subset of optimisation problems, the characteristics of the subset must be incorporated into the algorithm (Wolpert and Macready, 1997, pp. 71f). In the case of intelligence, the algorithm has to be optimised with respect to regularities (cf. Ma et al., 2022, pp. 1300f). The regular-

ities considered are thereby not necessarily truths of the worlds but assumptions. The formalisation of intelligence thus faces a dilemma in regard to determining to what extent regularities – and possible other assumptions – should be considered: The more assumptions are considered, the smaller the subset of optimisation problems covered and the more performant the algorithm, all else being equal. However, the more assumptions are considered, the greater the chance that they do not correspond to the worlds to which the algorithm is applied, and its performance decreases accordingly.

2.4 Perception & Experience

Skills and intelligence both require knowledge of at least some conditions of a world to determine appropriate actions to achieve a goal. Conditions can be determined through perception. For example, humans and animals can perceive stimuli that can be divided into three different types: Chemical stimuli include molecules and are experienced as odour and taste; mechanical stimuli include forces transmitted by matter and are experienced as touch, sound and heat; electromagnetic stimuli include electrical and magnetic radiation and are experienced as vision, for example. Stimuli are detected by receptors located in sensory organs, such as eyes. Together with the nerves that transmit and process their signals, sensory organs are referred to as sensory systems. For example, the human visual system includes the eyes, the connected nerves, and the visual cortex of the brain (Yong, 2022, pp. 7-11, 191, 213f). Sensory organs can vary in performance, e.g., regarding the type and detail of stimuli that can be perceived. Eyes, for example, can be divided into four stages of functional efficiency: In the first stage only the presence of light can be perceived, in the second stage also the rough direction from which the light comes. The third stage allows the perception of more detailed directions and therefore contrasts; and the fourth stage, through the use of lenses, allows sharp spatial vision at distance (Nilsson, 2009, pp. 2837-2843).
Sensory organs and their performance thus represent a limitation as to which conditions of a world can be perceived and in what detail. The limitations of sensory organs can lead not only to a lack of perception but also to distorted perceptions. For example, flickering light is experienced as continuous light above a particular flickering speed due to the limited temporal resolution of the sensory system. Another example is the human perception of the sky as blue: This occurs because the shorter the wavelength of light is, the more it is scattered and therefore the better it is perceived. Accordingly, the shorter-wave, blue light component of sunlight is scattered more strongly in the Earth's atmosphere than the longer-wave, red light component. However, the violet light component is even shorter in wavelength and is therefore scattered even more strongly. Yet, as human receptors perceive blue light more strongly than violet light, the sky still appears blue from a human per-

spective (Schaffer, 2005, p. 253). Moreover, many optical illusions demonstrate that conscious experiences do not correspond to what is perceived and some optical illusions persist even when one is aware of their incorrectness (e.g. Frith, 2007, pp. 40-50, 127-134).

In subjective human conscious experience[4], stimuli seem to be experienced directly, as if one experiences the stimuli themselves. Nevertheless, the relationship between stimuli and human experiences can be indirect and varying. Some sensory experiences are not generated directly by specific stimuli, but are generated by sensory systems. The colour yellow, for example, is perceived as a direct and genuine experience of a stimulus, just like the colour red. However, the colour yellow is not experienced because a colour receptor for yellow light is activated. Instead, it is experienced when green and red colour receptors are activated simultaneously (cf. Kelber et al., 2003, pp. 88-91). Hence, although the colour yellow appears as a direct perception of a stimulus, it is a generated experience without a corresponding stimulus of its own.

Furthermore, an individual stimulus can be experienced as perceptions of several sensory systems simultaneously. Synaesthetes experience, for example, sonic waves not only as sound but also visually as colours, whereby the experienced colours can differ depending on the person. Equally, their perception of light can lead not only to visual experiences, but also to experiences of taste (Ward, 2013, pp. 50-56). Conversely, stimuli of different types can trigger the same sensory system. For instance, capsaicin in chilli and menthol in mint produce the experience of heat and cold respectively, as the molecules activate temperature receptors (Hoffstaetter et al., 2018, pp. 746f, 751). The joint processing of stimuli of different types within a sensory system is widespread among animals: Platypuses combine signals from receptors for electric fields and mechanical forces, mosquitoes have neurons that react simultaneously to both temperature and chemicals, and migratory birds process the perception of both light and magnetic fields in the visual centre (Yong, 2022, pp. 314f, 323f). In addition, different sensory systems can be activated by the same stimuli. For example, odour and taste are partially activated by the same chemicals, such as esters and amino acids. Odour and taste thus do not differ primarily in that they perceive different types of stimuli; rather, their difference is functional: Reactions to taste are reflexive and innate, whereas those to odour are learnt and depend on experience (Valentinčič et al., 1994).

In summary, the same stimuli can trigger different experiences, and, conversely, different stimuli can trigger the same experiences. This shows that human subjective conscious experience is not a direct and unaltered experience of stimuli but an indirect and varying one. One of the reasons for this lies in how stimuli are perceived. In their basic functioning, all sensory systems are structured in the same way, re-

[4]This experience is often referred to as qualia. For a detailed discussion of qualia, see Tye (2021).

gardless of the type of stimuli they perceive: A stimulus triggers in complementary receptors a chemical or electrical reaction that leads to an electro-chemical activity of the receptors' neurons, which in turn results in neural activity in the sensory system (Dusenbery, 1992, pt. 2).

Odours, for example, are experienced when receptors are activated by specific chemical molecules. When molecules activate corresponding receptors, the receptors send a signal and release or destroy the molecule. However, there is not a specific receptor for every particular odour. Instead, many types of molecules activate several different receptors at once, and depending on which receptors are activated simultaneously, different odours are experienced. The characteristics of the receptors and their interaction depends on genes; for instance, the OR7D4 gene determines whether androsterone, a male sex hormone, is experienced as repulsive, vanilla-scented, or odourless (Keller et al., 2007). Visual perception relies on the same process, except that the relevant receptors, opsins, do not hold and repel molecules but are permanently connected to a chromophore molecule. When a photon hits a chromophore molecule, its energy changes the shape of the molecule, which in turn leads to neural activity of the receptor (Porter et al., 2012, pp. 3f, 11f). In hearing, hair cells are involved which, depending on the movements caused by sonic waves, release chemical substances that then lead to neuronal activity (Dusenbery, 1992, ch. 9).

Common to all these and other sensory systems is that stimuli themselves are not retained (Glasersfeld, 1996, pp. 115f).[5] Instead, a stimulus leads to a neural activity of an electro-chemical nature, which is dependent on various aspects of the stimulus; in the simplest case on its presence. The conscious experience of perception in humans is therefore not a direct experience of stimuli themselves, but is based on neural electrochemical activities caused by the stimuli. Overall, this shows that perception in humans and animals is the ability to convert stimuli, i.e., conditions of a world into neural activity. Generalised, perception can be defined as the ability to form states in dependence on conditions of a world. As such, the formed states are representations of conditions of the world. However, as shown above, the connection between the conditions of the world and their representations can be incomplete, ambiguous, and inaccurate due to the limitations of the sensory organs.

Since sensory organs provide only representations and not the stimuli themselves, representations can also be provided by other means. Accordingly, although intelligence requires knowledge of at least some conditions of a world, this does not necessarily have to be obtained through perception. Instead, knowledge can also be provided in other forms, such as a database. Examples of worlds for which knowledge is provided in this way, both for AI systems and for humans, are games or logical and

[5] The electrical sense, which allows to perceive electrical fields, could be considered an exception: Both stimuli and neural activity are electrical in nature. Yet, here too, the electrical stimulus is not continuously preserved, but its presence triggers chemical activity, which in turn leads to neural electrical activity that differs from that of the input (Baker et al., 2013).

mathematical systems for which axioms are provided instead of perceptions. Nevertheless, in principle, a comprehensive and precise perception is favourable: The more conditions of a world are known, the more precisely it is possible to determine which actions are appropriate, all else being equal. Furthermore, perception allows one to continuously obtain conditions of a world, allowing, for instance, the consequences of actions or previously unknown states to be determined.

2.5 Representation

Brooks (1991b, pp. 149-158) describes a robot called Herbert, which is an intermediate result of his approach to creating intelligent systems. The robot can move around in a regular office environment to collect empty soda cans. It is controlled by fourteen activity modules, each designed for a specific function; for example to avoid obstacles, to recognise tables, or to grasp objects. Accordingly, the robot is not based on classic AI approaches, such as symbolic AI reasoning systems, or neural networks. Instead, the robot is controlled by the activity modules and their interaction. The modules are interconnected and different modules take over control at different times depending on their states. For example, by default the robot wanders around. Yet, if the avoidance module recognises an obstacle, it takes over and changes direction. Equally, when a soda can is discovered, the grasping module takes over to stop the robot and to grasp the can.[6]
Brooks (1991b, pp. 148f, 140, 154; cf. Shapiro, 2019, pp. 175-180) takes the seemingly strict position that the approach does not rely on representations because there are no "tokens which have any semantics that can be attached to them". Subsequently, Brooks (1991a, pp. 18-20) takes a more nuanced position, which does not entirely deny the presence of representations, but rejects the presence of "explicit representations", "symbolic representations", and "traditional Artificial Intelligence representations schemes". The divergence seems to be primarily due to an insufficient differentiation between various kinds of representations, and the attempt to demarcate from traditional AI approaches that are based on logical systems involving natural language. At least implicitly, Brooks (1991b, p. 157) functional description of the robot refers to representations: "For instance the grasp behaviour can cause the manipulator to grasp any object of the appropriate size seen by the hand sensors." In other words, in case the hand sensors perceive stimuli typical for a soda can, the grasp behaviour module sends a signal to the manipulator. This signal is thus a representation of the perception of a soda can; as Brooks states

[6]Each activity module is based on a hardwired fixed-topology-network of simple finite state machines. As such, each module represents a specific skill. The robot is not able to adapt to novel, unknown circumstances; for example, it would not be able to learn to grasp soda cans of different shapes or bottles. Consequently, the robot fulfils only the first interpretation of intelligence outlined in Section 2.2, but not the definition of intelligence advocated in this article.

it: ”aspects of the world are extracted”. The same applies to other modules: For example, the ultrasonic sensors of the obstacle module send a signal when they are activated by an object. The signal thus represents a state of the world in which there is something in front of the ultrasonic sensor that activates it.
Representations vary in how vaguely or specifically they describe states of a world and how simple or complex they are. An example of rather simple, yet functional representations provides the water flea Daphnia. Its visual sensory system is not able to perceive details but can only recognise the presence of four different wavelengths of light. Depending on the wavelength of the perceived light, one of four different types of opsins is activated. Each type of opsin sends a specific signal, which thus represents the presence of light of the corresponding wavelength. Based on these representations, specific actions are triggered. For example, the signal representing the presence of green light, which indicates the presence of food, triggers the action to swim in the direction of the light. Similarly, the signal representing the presence of UV light, which indicates damaging UV radiation, causes the insect to swim away from the light source.[7]
In both examples, the signals represent very simple states, namely the presence of light of particular wavelengths. However, representations can also be more vague as well as more complex. For example, the representation of a forest has a higher level of complexity, as it includes a larger number of trees, other plants, and animals as well as a terrain. In addition, it has greater vagueness, as forests can include a wide variety of plant and animal species and can be of different kinds, all aspects that are not specified in the representation (cf. Shapiro, 2019, pp. 81f). Consequently, there are many different possible sets of states of the world that can lead to the same representation, and for all of which the representation stands accordingly. The complexity of a representation is of a gradual nature and depends, among other things, on the number of possible states represented, as well as on the variety in which they can be combined.
How easily and precisely a representation can be described depends not only on its complexity but also on the availability of suitable linguistic expressions. For example, the German term ‘Regenschirm’ refers to an umbrella that is used specifically to protect against rain; consequently, the representation of such an umbrella is easier to describe in German than in English. While the complexity of a representation is an inherent property, its describability depends on the language used and is consequently independent of the representation itself. Accordingly, Brooks (1991b, pp. 148f) distinction between ‘implicit’ representations and ‘explicit’ or ‘symbolic’ representations cannot be upheld: Explicitations and symbolic connotations of representations are only assignments, but not inherent aspects of the representations.

[7] In detail, the reactions are more complex than described here. For example, the insect’s reactions are also influenced by the circadian rhythm and genetic dispositions. In addition, very intense green light also causes the insect to move away from the light source rather than towards it. For reasons of illustration, these additional influencing factors are not taken into account here.

Based on the above, representations can be defined in the following way: A representation is a state that is dependent on one or more other states. Representations do not have to reflect other states completely, but can reflect only specific aspects of them. For example, the representations of the water flea Daphnia indicate only the presence of light of a particular wavelength, but not the polarisation or spatial distribution of the light. Moreover, representations can be indeterminate insofar as they can stand for several possible combinations of states, as the example of forests shows. Shapiro (2019, p. 182) argues that a definition of representations based only on dependencies is too broad; instead, representations "must be used as stand-ins by someone or something to count as representations". However, like the assignment of linguistic terms to representations, their use is something extrinsic – whether a representation is used or not is not part of the representation itself. As an illustration can serve a water flea whose opsins function normally but whose nervous system fails to process signals and thus to trigger actions. In that case, the representation of the perceived light generated by the opsins is not used; however, it is the same representation that a functional water flea would have that would use the representation.

2.6 Phenomena & Appearances

Based on the considerations in the last section, it therefore appears that representations are a fundamental component in the implementation of intelligence, as they provide information about states of a world. This view is widely held, particularly in the field of AI, where representations are assumed to be necessary for human and animal behaviour, as well as for AI approaches (Russell and Norvig, 2022, pp. 31, 76-78, 226f). However, Dreyfus (2007, pp. 249-251) argues on the basis of the relevance problem that AI approaches which are applied in dynamically changing worlds cannot be based on representations: AI approaches have to determine in specific situations which states of a world are relevant and which consequences result from changing states. Yet, representations of states are meaningless and, as part of this, provide no information about their significance. The meaning of a state of a world could be determined by knowing the concrete situation in which it occurs. For example, the significance of a red traffic light for cars depends on whether one is participating in the situation as a driver or as a pedestrian, as well as on the direction one intends to take. In order to determine the meaning of a represented state, an AI approach would therefore have to determine the situation in which it is applied. But to do so, it would have to determine which states of the world form the situation, in other words, which states are relevant. This leads to an infinite regress that cannot be overcome, as both the meaning and the situation can be determined only on the basis of the other. Although Dreyfus' criticism is directed against symbolic AI approaches, he also applies the argument to other approaches

that use explicit rules to manipulate representations (Dreyfus and Dreyfus, 1986, p. 99).

To overcome the relevance problem, Dreyfus (2007, pp. 252-255) argues, AI approaches must not be based on representations but must be able to perceive solicitations: "In coping in a particular context, say a classroom, we learn to ignore most of what is in the room, but, if it gets too warm, the windows solicit us to open them. We ignore the chalk dust in the corners and the chalk marks on the desks but we attend to the chalk marks on the blackboard. We take for granted that what we write on the board doesn't affect the windows, even if we write, 'open windows,' and what we do with the windows doesn't affect what's on the board" (Dreyfus, 2007, p. 263). In conclusion, solicitations arise from concrete situations and provide meaning. They disclose the world and offer a flexible response based on the significance of the current situation. In contrast to representations, which are part of an AI approach and are only assigned to a world, solicitations are the world itself (Dreyfus, 2007, p. 249). Accordingly, the meaningful is provided to an agent by the world, and appropriate actions do not have to be determined by the agent, but are offered as dispositions to respond to the solicitations of situations (Dreyfus, 2002, p. 367).[8]

Whereas Dreyfus' account is intended to overcome the relevance problem, it requires a strong ontological commitment: The approach presupposes the existence of solicitations for each agent in each situation. It is not clear where the solicitations originate and what nature they are.

Furthermore, Dreyfus' account is in contradiction to the above findings from the analysis of Brooks' robot Herbert. Dreyfus (2007, pp. 249f) does not see Brooks' approach as a solution to the relevance problem, since the approach is not able to learn and thus cannot deal with changing meanings in novel situations. Nevertheless, he considers Brooks' approach to be an important advance, as it is not based on representational, symbolic AI approaches, but on activity modules that react directly to the environment.[9] However, as shown above, Brook's approach is based on representations that, although not annotated with symbols or expressions of natural language, are processed according to explicit rules. This raises the question of whether Dreyfus' approach of solicitations is, at least partially, based on representations, too. Wheeler (2008, pp. 333-342), who also regards Brooks' approach as a major advance, presents an account which comes close to the one of Dreyfus, but relies on representations. Dreyfus (2007, p. 263), however, explicitly rejects this account, arguing that any representational state precludes meaning; instead, it

[8] This perspective is similar to the non-representational and non-computational account of Gibson (2014, pp. 119-121): Gibson, speaking of affordances instead of solicitations, argues that possibilities for action are offered to animals by their environment. Affordances are not part of the agent but part of the environment, and perception is not about perceiving and processing information but about receiving guidance for action.

[9] A similar assessment is provided by Shapiro (2019, pp. 175-180).

is necessary to directly sense and respond to the world. This raises the question, if solicitations are entirely non-representational, how can they be recognised by AI approaches, as well as by humans and animals, if not by means of their representational sensory systems (cf. Dreyfus, 2007, pp. 249-251, 256-265; Merleau-Ponty, 2012, pp. 364-369).

Beyond that, Dreyfus' approach seems not suitable as a basis for intelligence. According to Dreyfus (2007, p. 250), with increasing experience, we are presented with more and more finely discriminated situations that solicit increasingly detailed responses. As background know-how is refined, states of the world take on more and more significance. Additionally, Dreyfus (2007, p. 263) explains: "[W]henever there is a change in the current context we respond to it only if in the past it has turned out to be significant, and when we sense a significant change we treat everything else as unchanged except what our familiarity with the world suggests might also have changed and so needs to be checked out." Yet, with this statement, Dreyfus does not describe how the world provides solicitations, and thus meaning in situations. Instead, he describes how we cope with situations by applying our existing knowledge – in other words, by applying skills. In contrast, Dreyfus' approach does not allow for the application of intelligence, as it does not explain how we are able to perceive the meaning and significance of states of the world that are unknown to us.[10]

Overall, therefore, it seems that Dreyfus' account is not suitable as a foundation for intelligence. Nevertheless, the question arises as to whether insights can be gained from his approach and the underlying considerations that are helpful for understanding intelligence and for the creation of AGI. Dreyfus draws largely on considerations from phenomenology, in particular from the works of Heidegger and Merleau-Ponty.

Phenomenology focuses on phenomena and appearances and their conditions of possibility. Appearances, i.e., conscious experiences of phenomena, play an important role in phenomenology, since they are the most immediate to which one has access (cf. Kant, 1968, pp. 45f). Yet, phenomena are not, as is often mistakenly assumed, equal to the immediate appearances that one consciously experiences (Gallagher and Zahavi, 2020, pp. 11, 21-23, 251). Instead, phenomena are the essential structures that characterise appearances. Phenomenology is therefore not primarily concerned with the investigation of appearances as such, but with the investigation of phenomena, of appearances as their correlates, and of the connection between the two (Gallagher and Zahavi, 2020, pp. 23-28). As Heidegger (1967, pp. 36f) describes: just because phenomena are proximally and for the most part not given, there is a need for phenomenology. He argues, the idea of grasping and explicating phenom-

[10] Dreyfus (2007, p. 264) describes that we are made aware of new situations and states by having our attention drawn, "summoned", to them. However, this can only explain how to switch from one skill to another, but not how to create new skills that can address new, unknown situations and determine the meaning of unknown states.

ena in a way which is original and intuitive is directly opposed to the naïveté of a haphazard, immediate, and unreflective beholding. The aim of phenomenology is thus not the description of subjective content of experience, but the determination of necessary and invariant features and the answering of questions related to truth, reason, reality, being, ontology, science, and objectivity (Gallagher and Zahavi, 2020, p. 28).

However, this does not mean that appearances and phenomena are distinct from each other. Phenomena are not represented by appearances but unfold in them; appearances are thus part of phenomena (Gallagher and Zahavi, 2020, pp. 23-28). Consequently, no distinction can be made between subjective experience on the one side and objective reality on the other. Phenomenology is thus directed against the assumption of scientific realism that there is an objective reality that can be understood by removing all subjective elements of perception. Instead, the objective, necessary, and invariant features can be understood only if conscious experiences, i.e., appearances, are part of the investigation (Gallagher and Zahavi, 2020, pp. 108-114). Phenomenology hence reflects that science is carried out by someone and thus from a specific theoretical stance, which has its own presuppositions and origins. These presuppositions and origins need to be examined, which is why phenomenology is concerned, for example, with what the primitive modes of understanding are that precede beliefs in objectivity and how objectivity is constituted. In this way, phenomenology aims to provide a new epistemological foundation for science (Gallagher and Zahavi, 2020, pp. 23-28).

One of the main representatives of phenomenology is Heidegger, to whom Dreyfus refers most strongly. Heidegger (2012, pp. 10f) emphasises that our primary relationship to being, i.e., to the world in its entirety, is not in theoretical contemplation and investigation, but in immediate experience: For example, we do not hear a sequence of sounds, but we experience the closing of a door. In order to hear the sounds, we first have to reflect ourselves out of the situation and listen abstractly. The sounds thus not only represent an abstracted and hence reduced view, their characterisation is also based on theoretical assumptions, such as the existence of an objective world. Such views are therefore not suitable as a foundation for scientific investigations and insights, as they are already incomplete and may be based on erroneous assumptions. Instead, investigations have to start in the immediate experience: Only in the realisation of existence, called being-in-the-world, phenomena have the opportunity to reveal themselves and to disclose meaning. For example, we only recognise the meaning of music when we not only perceive it as a sequence of sounds but experience it as music (Heidegger, 1994, pp. 171-179; cf. Gallagher and Zahavi, 2020, pp. 177-182).

Heidegger's change from the assumption of an objective world to immediate experience entails a different understanding of the role of cognitive abilities. These no longer serve to establish the relation between the self and the world. Instead,

the world unfolds within the being-of-the-world, and relations between phenomena result from this. Cognition thus becomes a secondary modification of being-in-the-world and is only possible because that is already present (Gallagher and Zahavi, 2020, p. 178). Heidegger's approach is thus closely related to Dreyfus' approach. In consequence, Heidegger's approach faces the same limitations as Dreyfus' in relation to skills: The perception of appearances is immediate, but at the same time it is already a matter of specific interpretation; for example of sounds as music (Gallagher and Zahavi, 2020, p. 8). An assessment that is also supported by Husserl (1984, pp. 801f), who states: It belongs to experience that something appears in it, but the interpretation makes up what we call appearance – be it correct or not, anticipatory or exaggerated. Heidegger's approach is therefore in some respects more direct and less presuppositional than, for example, scientific realism, but at the same time it is also based on interpretations and thus on assumptions.

In Heidegger's works, a clear change in perspective can be recognised between his earlier and later writings, which he himself describes as a turn. While all the above considerations derive from his earlier writings, Heidegger (2001, pp. 173-185) argues in his later writings that an understanding of the world requires, moreover, an engagement with the openness of unconcealment. In his earlier writings, to which Dreyfus refers, Heidegger is concerned primarily with the question of how one can experience the world directly by being-in-the-world. In his later writings, Heidegger focuses more on being, which, he states, he did not sufficiently consider in his earlier writings, as he focused too much on being-in-the-world (Heidegger, 2000, pp. 49f). Being can be contrasted with being-in-the-world: In being-in-the-world, in the realisation of existence, phenomena manifest themselves in the form of appearances and have meaning. In the experience of being, however, one transcends concrete existence and experiences phenomena without interpretation: One experiences the inexhaustibility of the world and discovers the possibility that existence can also be different. In concrete terms, being contains all practised, all conceivable, and all as yet inconceivable possibilities of being-in-the-world. At the same time, being eludes definition; the moment it is defined, it becomes being-in-the-world and is no longer being. Accordingly, the experience of being is not present when one is trapped in one's own being-in-the-world; i.e., when one experiences the world in a specific way that is determined by a particular interpretation and from which one cannot free oneself. Being can thus be understood as a game of possibilities that allows one to see the world as it is without interpretation, without a particular world view (Heidegger, 1997, pp. 153-169). Yet, this world is not graspable since it consists of that very play of possibilities without ever adopting a particular one (Heidegger, 1999, pp. 224-227). Metaphorically speaking, concrete realisations of being-in-the-world, such as religions and cultures, can be understood as fragile rafts that humans build on the open sea and on which they drift through time for a while, while modifying

and sometimes rebuilding the rafts (Safranski, 2014, pp. 341-343, 406-409, 473f).[11] Heidegger's earlier approach and his later approach show clear parallels to the concepts of skills and intelligence. Heidegger's earlier approach, to which Dreyfus also refers, describes the perception of and interaction with a world from a specific perspective, i.e., skill, whereby things have a specific meaning. One example of this is Brook's robot Herbert, whose task is to identify objects shaped like soda cans and to pick them up. Another example is the game of chess, in which pieces have a specific function and the game follows specific rules.

Heidegger's later approach, by contrast, describes the experience of phenomena without them being subject to any particular interpretation. This corresponds to situations in which one is confronted with unknown states of the world and in which one therefore has to apply intelligence to be able to interpret them. Both Heidegger's later approach and intelligence are therefore concerned with assigning meaning to uninterpreted phenomena in order to gain new insights. Thereby it becomes apparent that Heidegger's uninterpreted experience of phenomena is subject to the same restrictions as intelligence with regard to the necessity of assumptions: The existence of being reveals itself only in non-interpretation. Yet, being-in-the-world presupposes that the world is interpreted in a specific way. Any specific interpretation thus hinders access to being. The same applies to intelligence, which, as the NFL theorems show, can be successfully applied only if assumptions are made, such as that the world exhibits regularities. At the same time, however, these assumptions already represent an initial interpretation of the world and hinder other interpretations of the world that are not in accordance with them.

The significant similarities between Heidegger's later approach and intelligence lead to several implications that arise from phenomenology with regard to the creation of intelligence: First, phenomenology shows that a subject is not independent of the world but is part of it and that there are close interactions; the separation between subject and world is therefore artificial and depends on the respective interpretation. The water flea Daphnia can successfully consume food because it is in a world in which food is of such a nature that the available light stimulates the sensory system in such a way that it triggers the corresponding action. Nevertheless, the question arises as to whether one can therefore speak of solicitations as Dreyfus and Gibson do. Algae and their properties are indeed necessary, as is light and its properties. However, some of Dreyfus' and Gibson's statements seem too strongly focused on the world and thus insufficiently consider the role of the agent; this, for example, when it is said that the world offers solicitations, provides guidance, and summons the agent. Although the agent is part of the world from a phenomenological point of view, the specific characteristics of the agent, such as the degree to which it can perceive the world and gain insights from it, have to be considered as well. This is

[11] A similar position is held by Nietzsche (1982). For a comparison of both positions, see e.g. Safranski (2014, pp. 66-68, 276-278, 336-343).

especially because the experience of the world depends very much on the agent – the same states of the world are perceived and, in particular, interpreted very differently by different animals, even by every human being. The experience of the world thus also depends fundamentally on the agent itself, and it can only be understood if its active role is sufficiently taken into account.
Second, phenomenology shows that the pursuit of insight is carried out by subjects and that their presuppositions and origins must be taken into account. This can be seen, for example, in the necessary consideration of which perceptions an agent can have, as shown in Section 2.4. Sensory systems determine not only which aspects of the world can be perceived at all and to what degree of detail, but also how they are processed and whether they are subject to distortions, for instance. Likewise, it is necessary to consider on which assumptions the intelligence of an agent is based; for example, in which form it is assumed that the world is subject to regularities and which other assumptions are included.
Third, the considerations in the preceding paragraph entail that humans and artificially created intelligence capture the world in fundamentally different ways: Humans capture the world first and foremost as being-in-the-world, they experience it consciously and in a specific interpretation. This experience is based on the specific configuration of their sensory systems and on their interpretations, which are grounded in cultures, for example. Artificially created intelligence is also based on particular specifications, and thus interpretations, both through the sensory systems with which it is equipped and through the assumptions that are given to it. However, the specifications and thus interpretations given to AI can be changed, while the ones for humans are relatively fixed (cf. Spelke, 2022). Furthermore, AI is subject to far fewer specifications and interpretations than humans are. While humans rely heavily on interpreted perceptions – for example, a car can be experienced visually only as a car and not as a cluster of lights – artificial systems can rely on significantly less strongly processed data (cf. Frith, 2007, pp. 40-50, 127-134). Humans are subject to many preconceptions that they cannot question or can only question with great difficulty, as they are very strongly characterised by their interpretation. A metaphor of Neurath (1932, p. 206) illustrates this problem: We are like sailors who have to rebuild their ship on the open sea without ever being able to dismantle it at a dock and rebuild it from scratch with the best components.
From a phenomenological perspective, AGI therefore has the advantage that it can be much closer to being, and can be much less influenced by interpretations than humans can.

2.7 Meaning & Understanding

The foregoing considerations raise the question of whether AI can be capable of assigning meaning to states of a world, i.e., to create interpretations, and if so, how

it must be designed to do so. In the following, meaning is defined as a function that is attributed to something in order to to achieve a specific goal (cf. Shanahan, 2005, p. 106; Yong, 2022, pp. 5f). The representation to which the function is attributed thus serves as a means for achieving a specific goal. For example, the function of a hammer is to drive nails into walls. Meaning is something that is attributed by an agent to something and does not exist independently of the agent. For instance, a hammer – or a stone – only becomes a hammer when this function is attributed to it. Nevertheless, the attributed function can be applied successfully only if the world in its entirety is such that the function enables the fulfilment of the goal. For example, something can only have the function of a hammer if it is hard enough to drive a nail into the wall, there are a matching wall and nail, and the subject is able to use the item accordingly. The successful fulfilment of an attributed function, a meaning, is therefore dependent on the world – yet, it is not a solicitation or an offer, but a possibility. The possibility can only be used, however, if the agent attributes it to the respective state of the world.

Closely related to meaning is understanding. Understanding is the ability to use something in such a way that it fulfils its meaning (cf. Preston, 1993, p. 44). For example, many people have an understanding of cars that allows them to use them as a means of transport by driving them from one place to another. Understanding is gradual and can be more or less pronounced in terms of both efficiency and effectiveness. For instance, some people can drive better than others and arrive at their destination faster and with less gasoline consumption. Similarly, some people can drive in conditions in which other people can no longer drive, such as in a snowstorm or in the desert. An agent therefore has a the greater understanding of something the more efficiently it can use it the more extensively. As such, understanding represents a skill that describes how well something can be used in a certain functionality, i.e., with regard to an attributed meaning.

In comparison, it can be said that meaning describes the function that is attributed to something, whereas understanding describes how something has to be used to fulfil this function. Understanding thus presupposes meaning: something can be understood only with regard to a specific meaning. For example, a car can be understood only as a means of transport – or as a status object or as an investment – if the respective function is known, as each function requires a different understanding. In the case of an investment, for instance, it is not a question of how the car can be steered with the aid of the steering wheel, but how which equipments contributes to the value of the car (cf. Safranski, 2014, pp. 144f).

The definition of meaning and understanding in a functional way implies that both are present when something is successfully used to achieve a specific goal. For example, for the water flea Daphnia, green light has the functional meaning of indicating food, and the water flea has an understanding of the light in that it uses it as an indicator of food. The water flea also uses water as a means of transport

and understands it such that it can move successfully in it. Unlike humans, for example, the water flea does not know what light or water is from a physical point of view. However, this is not necessary: In the past, people also used light as an indicator of food without knowing its physical properties. Equally, people used water for transport in the past without knowing its physical and chemical properties. Conversely, humans today have much greater knowledge of light and water, but it is still limited – hence there is only a difference in degree, not in kind.

It could be argued that meaning and understanding can only occur if mental states are present, which, for example, make it possible to experience them consciously. Searle (1980, pp. 417f), for instance, introduces the Chinese Room Argument[12] to argue that understanding can exist only if intentionality, which is like consciousness a mental state[13], is present. According to Searle (1980, pp. 421-424), mental states can be produced only by specific physiochemical structures that have particular causal powers. Such structures occur only in certain biological organisms: in humans, in primate species such as monkeys, and in domestic animals such as dogs. Formal models, on the other hand, do not have the biological structures required for the causal powers and are therefore unable to constitute mental states such as intentionality and, consequently, understanding. However, Searle does not explain why mental states can only originate from specific biological structures and how they originate from these structures. It therefore remains unknown why other structures that can perform formal operations cannot be capable of generating intentionality as well.

It is also unclear why intentionality is necessary for understanding and what additional properties or functions intentionality, or mental states in general, contribute. This in particular given that Searle (1980, pp. 422-424) takes a materialistic position[14] and thus does not require a separate quality from mental states that cannot

[12]The Chinese Room Argument is based on the following thought experiment: Searle, who does not understand Chinese, is locked in a room and given three batches of Chinese characters that have no meaning to him. In addition, Searle receives instructions in English that allow him to relate the elements of the different sets to each other to generate a fourth set, which he has to output. Unknown to Searle, the first set is called 'script', the second 'story', the third 'questions', and the fourth 'answers'. Based on this thought experiment, Searle argues that by following the script, he can answer the questions about the story correctly and therefore, from the outside, it appears that he understands the story. However, he does not understand the story as he does not understand the Chinese characters; instead, he only relates and manipulates these symbols according to the rules of the script.

[13]There are different views on how intentionality and consciousness are related. For instance, Searle (1992, pp. 93-100) and Gallagher and Zahavi (2020, p. 101) each argue in their own way for a close relation of intentionality and consciousness. Heidegger, on the other hand, rejects any identification of intentionality with consciousness or inner experience (cf. Haugeland, 2013, p. xii). For an overview of different conceptions of intentionality, see e.g. Gallagher and Zahavi (2020, pp. 96f) and Smith (2018).

[14]This assessment is discussed controversially, as Searle does not specify how mental states arise from physiochemical structures and what kind they are. For example, Haugeland (2000, p. 291) categorises Searle's approach as materialistic, whereas Smith (2018, sect. 6) denies this assessment.

arise from the material realm. In addition, it is not clear why the conscious experience of understanding and meaning, as it occurs in humans, should be a necessary prerequisite for them. As shown in the previous section, consciousness, in the form of being-in-the-world, enables the experience of a specific interpretation of a world, i.e., of already formed meaning and understanding. In contrast, the generation of meaning and understanding takes place before they are accessible to consciousness in the form of experiences, their generations seems thus prior to consciousness. It therefore seems appropriate to regard intelligence – the creation of skills and thus of understanding as well as the attribution of meaning to phenomena – and consciousness – the experience of particular interpretations of a world – as two separate aspects that are independent of each other (cf. Hempel, 1966, pp. 8f).
Accordingly, in Searle's argumentation, a distinction has to be made between understanding in the functional sense and the conscious experience of understanding: While the biological structures mentioned by Searle may be necessary for the occurrence of consciousness, understanding can occur independently of them. With regard to the comprehension and creation of intelligence, the questions of how consciousness arises and in which agents it occurs are irrelevant. To some extent, Searle (1980, p. 421) also advocates a functional perspective when he uses behavioural analyses to infer the existence of intentionality: He argues that the behaviour of some animals can be explained only by attributing intentionality to them. In this respect, Searle also advocates a functional interpretation of intentionality and understanding.
In the following, it is analysed how intelligence and, as part of it, the attribution of meaning and understanding, can be explained in a purely naturalistic way without requiring mental states such as intentionality or consciousness. Intelligence is the ability to create a skill in which a specific state of a world, i.e., a goal, is pursued in dependence on other states of the world. A skill hereby represents a function that leads to the fulfilment of certain states of a world. As such, a skill represents an algorithm: particular inputs, i.e., states of the world, lead to particular outputs, i.e., other states of the world like actions. A simple example of a skill is the water flea Daphnia: It constitutes a skill that reacts to green light in such a way that the goal of nutrient supply is fulfilled. A skill, i.e., the function it constitutes, is realised by an executing system. In the case of the water flea, the executing system is the physical body, which consists of various components arranged in a specific structure: The sensory system triggers neuronal activity in the presence of green light, which ultimately leads to swimming movements.
Since skills are functional, they can be realised in various ways and are not tied to a specific executing system. For example, an artificial neural network for digit recognition can be realised with electrons in a silicon-based computer chip, as well as with optical waves in a nanophotonic medium (Khoram et al., 2019). Nevertheless, a skill is only realised due to an executing system: the executing system is therefore necessary and constitutive. Accordingly, the executing system is material and skill

at once: The skill results from the properties of the executing system, such as its structure (cf. Hatfield, 1988, pp. 202-206).
The skill of the water flea Daphnia is functionally a very simple skill. Intelligence is functionally more complex but subject to the same considerations as it is also a skill. In functional terms, intelligence is an optimisation algorithm whose goal is to develop a skill that achieves a specific goal under specific circumstances. An example of an optimisation algorithm is a reinforcement learning algorithm.
The attribution of meaning results from the creation of a suitable skill. For example, a reinforcement learning algorithm can create a skill that is optimised based on the reward for fulfilling the goal of nutrient supply: If the skill created leads to movement towards green light, as this turns out to be beneficial for the achievement of the goal, green light takes on the meaning of serving as a means of nutrient supply. Equally, a skill implies an understanding of something if it is successfully used to achieve the goal. The skill above, for example, implies an understanding of green light if it can be successfully utilised to achieve the goal of nutrient supply.
Based on these considerations, it is possible to define information: Information consists of representations that are used functionally. For example, a representation that indicates the presence of green light becomes information in that it is used by the water flea to fulfil the goal of nutrient supply. From a functional point of view, a representation is information about the state of a world on which the representation is dependent (cf. Haugeland, 2000, pp. 300f). Like skills, information is functional and therefore also not bound to a specific executing system, a specific medium, but can be realised in different ways. As with skills, however, media are necessary and constitutive. Similarly, the ability of a medium to provide information results from and depends on its properties: A medium, for example an electron released during neuronal activity, can be information about a state of the world exactly then, if it represents it.
Gallagher and Zahavi (2020, pp. 121-123) argue that representations cannot serve as a basis for understanding: To know that a representation corresponds to the represented, one must first grasp the represented directly, i.e., non-representationally – but this is not possible from a representational point of view, since one can only grasp representations of something but never the represented itself as it is. However, as the preceding considerations here and in Section 2.4 show, representations are not depictions that have to be created on the basis of what is to be represented, but they are inherently dependent on that. In consequence, representations may be incomplete in the way that they reflect only partially the states they represent, but they are grounded in them and thus correspond to them; an assignment is therefore not necessary (cf. Beckmann et al., 2023, p. 402).
Overall, this naturalistic approach makes it possible to explain the realisation of intelligence without having to resort to controversially discussed and ambiguous concepts such as cognition, mind, thought, or intentionality. The approach also

dispenses with the need for consciousness, which is considered something that can co-occur with intelligence but is functionally independent of it. The approach advocated here does not take a position on how consciousness arises and in which agents it is present; whether, for example, insects such as the water flea Daphnia, the primate species and domestic animals mentioned by Searle – or certain forms of AI – exhibit consciousness and what its nature is. The approach also allows one to avoid several controversial assumptions, such as that a world offers solicitations or that agents are summoned by the world. Moreover, the approach makes it possible to solve the relevance problem: Meaning results from a representation taking on a particular function within a skill. The function is attributed by intelligence by drawing a relation between the represented state of the world to which the function is attributed and the state of the world to be achieved, i.e., the goal.

2.8 World Model & Reasoning

The entirety of all knowledge, i.e., all skills as well as all non-goal-orientated knowledge, such as knowledge about particular states of a world, is often referred to as world model in the field of AI.[15] Intelligence, i.e., the creation of skills, is thus about the expansion of a world model. Various methods are available for this purpose, which are analysed in the following.

World models can differentiate from each other with respect to their complexity, for example, with regard to the amount of knowledge they include, but also whether they take the temporal dimension into account. The Daphnia water flea represents a very simple world model in which the temporal dimension is not taken into account and which is mainly composed of simple, action-orientated knowledge, such as that it is helpful for the goal of nutrient supply to swim towards green light.[16] In comparison, humans have a very complex world model that takes into account the past and the future, and describes many states of the world in detail.[17] Similarly, in the world model of the water flea, few states of the world are attributed only few meanings; algae, for example, serve only as food. In human world models, on the other hand, many states of the world are attributed many meanings. For example, plants are used as food, medicine, wrapping material, decoration, poison, and combustion material.

[15] In philosophy, the terms (scientific) theory and background knowledge would be suitable for describing the entirety of all knowledge of an agent.

[16] To a certain extent, the temporal dimension is accounted for insofar as, for example, circadian rhythms influence the behaviour of the water flea. However, the temporal dimension is not incorporated in such a way that future states or future actions are considered.

[17] Despite its greater complexity, the human world model, like that of the water flea, is in principle action-orientated. This is illustrated by the environmental dependency syndrome, which can result from focal unilateral frontal lobe lesions and causes people to react directly and compulsively to environmental stimuli, for example, when they see a bed, they lie down (Lhermitte, 1986).

Knowledge that forms a world model can originate from three types of sources: Prior knowledge refers to all knowledge made available to an agent, for example, in the form of assumptions that serve as the basis for intelligence.[18] Perceived knowledge refers to all knowledge an agent gains through perception, for example, by vision. Derived knowledge refers to all knowledge an agent derives from other knowledge, for example, by inferential methods such as deduction and induction.

Intelligence is about the derivation of knowledge with the aim of determining actions that allow the fulfilment of given goals: Skills are created by deriving them from already present and possibly perceived knowledge, e.g., by observing new unknown states and expanding existing skills accordingly (cf. Pfister, 2025a, ch. 4-7). Intelligence can occur only through the derivation of knowledge: If skills were provided in the form of prior knowledge, they would not be created and therefore would not meet the definition of intelligence outlined in Section 2.2. Also, no new skills can be gained purely from perception, as perception has to be interpreted and set in relation to the goals to be achieved in order to become skills.

The derivation of new knowledge from existing knowledge is achieved by reasoning. Reasoning comprises various methods that make it possible to draw more or less reliable conclusions from existing knowledge. Among others, reasoning includes the three inference methods deduction, induction, and abduction. Deduction allows to derive certain conclusions, i.e., the truth of a conclusion necessarily follows from the truth of the premises. For example, if swans are birds and all birds lay eggs, then swans lay eggs. Induction allows generalisations, i.e., to make predictions about hitherto unknown states of a world, but is uncertain. For example, if all the swans one has seen are white, one can infer that all swans that exist are white. Abduction allows to infer from a known state of a world to another state of the world that implies the known one (ch. 4). For example, from wet grass one can infer that it has rained. Abduction is in general an uncertain conclusion, since there can also be other implying states of the world, for example, a lawn sprinkler.

In addition to that, abduction is the most powerful inference method as it allows the introduction of new, composed concepts: For example, one can infer abductively

[18] An overview of human prior knowledge, which is often referred to as core knowledge and which is provided among others by genes, can be found in Spelke (2022). For a possible implementation of core knowledge in the field of AI, see Lake et al. (2017, pp. 4, 9-12).

from the observation of apples falling from trees the concept of gravity.[19] Although intelligence relies only on existing knowledge and perceptions, it is thus possible, with the help of abduction, to extend world models and to create new representations (sect. 4.5; Thagard, 2012). Nevertheless, the creation of something new is limited insofar as everything new has to be based on something known; all newly formed representations originate from existing representations (cf. Locke, 1847, bk. II ch. I par. 24; Rosenthal, 2004, p. 193). For example, from the two existing representations of a red line and a green circle, it is possible to create a new representation of a red circle. However, it is not possible to create a representation of a new colour or shape without drawing on other existing representations.
It is unclear to what extent humans use the three inference methods deduction, induction and abduction and to what degree they use other, additional reasoning methods. For example, instead of induction, Bayes' theorem could also be applied (Okasha, 2001).[20] Humans also appear to possess the ability for causal reasoning as prior knowledge (Newman et al., 2008), although this could be derived inductively as well. Further research is therefore needed on which reasoning methods should serve as a basis for intelligence, particularly with regard to the development of AGI. The reasoning processes on which humans rely have proven to be advantageous in evolutionary terms and could therefore be viewed positively. Nevertheless, evolutionary development is a continuous process and is dependent on the environment and human limitations, such as the performance of the brain. Furthermore, it seems to be evolutionarily advantageous if humans are equipped with as many skills as possible right from the start; for example, it is easier to recognise causality if it is already known and does not have to be derived using intelligence. With respect to the development of AGI, however, it can be advantageous to provide it with only the most foundational methods possible as prior knowledge in order to minimise the number of potentially incorrect assumptions. Although, as shown in Section 2.2, it

[19] Concepts are – from the perspective of the approach presented in this article – synonymous with representations; the two terms differ primarily in that the term concept is common in the field of philosophy and psychology, whereas the term representation is primarily used in the field of AI. In Section 2.5, representations are defined as states that are dependent on one or more other states. This implies that representations can represent states of a world that do not themselves constitute representations, as well as states that themselves constitute representations. Representations therefore also include inferred representations. Some approaches consider as concepts only representations that do not just originate from direct perception but also include theoretical features (cf. Carey, 2000, pp. 4-8). This demarcation is ambiguous, however, and cannot be based on a qualitative difference: All sensory perceptions are already theoretical in nature due to the way they are perceived, as well as the way they are processed in the sensory systems (cf. sect. 2.4; Peirce, 1998, EP 2 p. 227). Another possible definition is based on the assumption that representations are only concepts if they are used to explain data but cannot be perceived themselves (Horst, 2005, pp. 14f). Here too, however, it is not possible to make a clear distinction; bacteria and electrons were originally purely theoretical concepts, but can now be perceived with the aid of microscopes.

[20] Although it is unclear exactly which methods are used by humans and animals, methods for recognising regularities are widely used, as a study of Skinner (1948) on pigeons illustrates.

is necessary to specify assumptions for intelligence due to the NFL theorems, these should be determined prudently and be as foundational as possible. In this respect, further research is also needed on whether there is one set of assumptions which represents an optimum for AGI – or whether it may be advantageous to create several instances of AGI with different sets of assumptions to achieve greater variety, and thereby more powerful forms of intelligence.

In addition to the aforementioned methods of reasoning, at least two further methods are required for processing representations, and which, for example, provide the foundation for the formation of new, derived representations: abstraction and classification. Abstraction describes the ability to select specific features of a representation and to disregard all other features. For example, from the representation of a green circle or a green tree, the abstraction green can be created. Abstraction, like abduction, thus allows the introduction of new representations. However, abstraction can only form new representations by removing features from existing representations. Abduction, on the other hand, can form new representations by combining different features from various representations and is therefore more powerful, as it can create composed representations. Abstraction occurs extensively in sensory systems in biological organisms, as it makes it possible to significantly reduce the number of representations in order to represent only relevant states of a world (cf. Yong, 2022, pp. 66f). Beyond this, abstraction is seen as an important method for creativity (Welling, 2007).

Classification is a method that is applied in two ways. First, classification can be based on abstraction: Features that have been abstracted can serve as a basis for a common classification of different representations. For example, the abstraction green allows all representations that contain this feature to be grouped together, e.g., green circles together with green trees. Second, classification takes place to form individual representations from the temporally continuous stream of perception. This often takes place in sensory systems and depends on their design. For instance, the configuration of neurons and their firing rate determine whether several flashes of light are classified into several separate or one combined representation. Classification is therefore one of the most elementary methods used to process representations.

Reasoning methods and other methods, such as abstraction and classification, for processing representations and deriving new knowledge hence represent methods that can be applied by intelligence to develop a world model – and thus skills. At least some of the methods have to be provided for intelligence and constitute assumptions on which it is based. The methods are hence a concretisation of the necessary assumptions discussed in Section 2.3, which have to be provided to intelligence due to the NFL theorems. The conclusions drawn via these methods are reliable only if the methods fit the respective world for which they are assumed. For example, induction can be successfully applied only if the world is such that

generalisation leads to success (cf. Hume, 2016, sect. IV).

2.9 Viability & Construction

The value of a world model is judged by its functional performance: The more extensively goals can be achieved with the help of skills that are part of the world model, the more useful the model is (Glasersfeld, 1996, pp. 116-128; Frith, 2007, pp. 136f). Glasersfeld (1996, pp. 14, 68f) uses the term viability to describe how successful an agent is in achieving its goals. Accordingly, the aim in developing a world model is to ensure that it corresponds to a world as far as possible, i.e., that the model is compatible with the world in the sense that the chosen actions lead to the fulfilment of the goals. In contrast, the world model is not meant to depict the truth, i.e., the world as it is (Glasersfeld, 1996, pp. 109-114).
Eliminating the need for truth and instead aligning a world model solely on viability offers several advantages. First, this allows complex issues to be represented in simplified forms, as long as they are functionally precise enough. For example, humans use simplified descriptions of how objects move, e.g., to predict the trajectory of a throw, which ignore many factors and are not true but functional (Lake et al., 2017, p. 10). Second, it is not necessary that truth has to be perceived. As shown in Section 2.4, humans only experience representations of states of the world through their sensory systems but not the states themselves (cf. Nietzsche, 1982, pp. 312f). Equally, as shown in Section 2.6, human conscious experience does not allow direct access to phenomena, i.e., truth, but only offers an interpretation, i.e., an experience based on a world model (cf. Nietzsche, 1982, pp. 317f). It is therefore unclear on what foundations a world model based on truth can be developed. The approach advocated here does not exclude the possibility of perceiving truth but does not require it either, which makes the approach less presuppositional.
Even though the approach does not aim to reflect the truth of a world, it nevertheless assumes a correspondence with the world: Although perceptions are indirect and only provide representations and not the world itself, the representations are dependent on the world and therefore correspond to it. For example, an opsin sends a particular neural signal, a representation, precisely when it perceives light. The neural signal is only a representation of the light and not the light itself, but it depends on it. In this way, world models are grounded in the world and correspond to it. The at least partial correspondence of the world model with the world is shown by its functional success: without correspondence, a world model could not be viable.
Applying reasoning methods to the perceived correspondences allows a multitude of different conclusions to be drawn. The reason for this is that some of the reasoning methods are uncertain and contingent and therefore allow only possible but uncertain conclusions, i.e., hypotheses. Not all hypotheses can be directly, e.g.,

empirically, assessed, and some hypotheses may imply the same perceivable correspondences. For example, the observation of apples falling from trees can be explained by the hypothesis that gravitational forces act on them. Alternatively, the observation can also be explained as an effect of spacetime curvature without the apples being exposed to forces. Both hypotheses represent viable conclusions, and without further observations and reasoning, neither can be proven to be superior, i.e., more viable, than the other (Glasersfeld, 1996, pp. 113f). Consequently, as long as the conclusions drawn are viable, any of the most different conclusions can be accepted, each being as valid as any other. Glasersfeld (1996, p. 118) describes this constructivist position, in which any representation can be constructed as long as it is functional and corresponds to a world, as follows: "What we ordinarily call reality is the domain of the relatively durable perceptual and conceptual structures which we manage to establish, use, and maintain in the flow of our actual experience."
Hypotheses can be evaluated using explanatory virtues to determine which of several competing hypotheses should be preferred. For example, hypotheses can be preferred that are simpler or make more comprehensive statements (Peirce, 1958, CP 6.447). However, the significance of explanatory virtues is unclear and it is not clear to what extent they enable an assessment of hypotheses (cf. sect. 4.2.3; Cabrera, 2017, sect. 3). Furthermore, it is unclear what their assessments are based on. On the one hand, it is conceivable that they are rooted in a statement about the nature of a world, such as that the world is simpler rather than more complex. This, however, is an assumption and it may not apply to the world in which the virtue is used; with the consequence that conclusions derived from it may also not apply to that world. On the other hand, virtues can be derived from existing assumptions – for example, that the preferability of a hypothesis is measured by its functionality. However, with this, virtues do not offer additional assessment opportunities beyond the existing ones.
The contingency in the processing of representations, i.e., the possibility of drawing not only one conclusion but a multitude of different ones, applies not only to inference methods but also to classification: Here, too, it is necessary to carry out the classification on the basis of specific assumptions, i.e., virtues, which determine, for example, how many classes should be created, or on the basis of which criteria elements should be classified as similar or different to each other.[21]
The representational character of intelligence, which is due to the indirectness of perception, constitutes its potency (cf. Thagard, 2012, pp. 400f): From existing representations, new representations can be formed that are grounded in a world but that do not necessarily correspond to it completely; in other words, it is possible to form representations that deviate from the world. This enables the formation of constructs and the realisation of planning, i.e., the creation of what-if scenarios and

[21] An illustration of this is provided by the Chinese Restaurant Process, which utilises the virtue simplicity to determine whether an element should be assigned to an existing class or whether a new class should be created for it (Tenenbaum et al., 2011, p. 1284).

the prediction of what future states of a world will be like without these states actually existing.
Overall, the considerations thus show that world models are not images of a world that are becoming increasingly detailed and depict ever more aspects of the world. Instead, world models are collections of contingent and uncertain conclusions that aim for the greatest possible correspondence with a world and the greatest possible viability, i.e., the possibility of achieving goals. World models may seem like truth from human conscious experience, since one is accustomed to them and since they correspond to the world, but nevertheless they are only constructs, as Nietzsche (1982, p. 314) metaphorically describes: So what is truth? A mobile legion of metaphors, metonymies, anthropomorphisms, in short a sum of human relations that have been poetically and rhetorically intensified, transferred, adorned, and which, after long use, seem firm, canonical and binding to a people: truths are illusions of which one has forgotten that they are such.

2.10 Agentness & Interrelation

Intelligence is applied to develop a world model that enables the fulfilment of goals as comprehensively as possible. In this way, the world model should have the greatest possible correspondence with the world and make it possible to identify actions that allow to influence the world in such a way that the goals are achieved as far as possible. By world is meant everything that is; in phenomenology this is referred to as the totality of phenomena. However, agents, including humans, animals and AGI, cannot access the world in which they are situated in its entirety and in a direct way, as has been shown in the discussion of phenomenology in Section 2.6. Instead, agents are faced with the challenge that they can grasp only corresponding representations of a world through perception, which usually concern only a small part of the world and can be distorted.
The correspondences usually appear to be in the form of a temporally continuous stream of perception.[22] Nevertheless, it is not clear to what degree time (and space) reflect a basic constitution of the world and to what extent it is only an interpretation, i.e., the experience of a world model (cf. Kant, 1968, pp. 78-80). The basis and starting point of all applications of intelligence is thus, as Heidegger pointed out, this subjective perspective of perception on the basis of which the world has to be functionally comprehended.
The assumption that there is a world which can be perceived often proves to be helpful from a functional point of view. The approach advocated here only assumes

[22]This applies at least to the world in which humans live and to many worlds created by humans, such as computer games. There are also worlds that do not involve a temporal aspect, such as some logical puzzles like Sudoku. Mathematics is in general also not based on temporal aspects, although it can be used to represent them.

the existence of such a world, but makes no further specific assumptions, such as that the world is material; it is only necessary that correlations can be perceived.[23] Equally, the division of a world into a self, in the functional sense, and an environment is often helpful from an agent's perspective, whereby the division is based on functional criteria (cf. Iriki et al., 1996). For example, it can be helpful to differentiate between one's own body, i.e., the executing system, and the rest of the world, since the own body forms a spatial and temporal unit, can be perceived differently, and manipulations to the body lead to different effects compared to the rest of the world (cf. Yong, 2022, pp. 325-328).

Irrespective of the functional advantages of dividing a world into a self and an environment, an agent is in general part of the world and as such is subject to close interactions with the rest of the world in many ways. For example, the possibilities of the agent's perception are determined by the world. The perception of light requires not only the presence of light, but also many other aspects of the world influence it. For instance, light propagates much better in air than in water, which is why visibility on land is much better than in water. MacIver et al. (2017) argue that as a result, the migration of animals from water to land about 400 million years ago led to a significant improvement in the eyes and, consequently, to more elaborate behaviour, as the more extensive perceptual possibilities allowed for more sophisticated planning. An agent is therefore not just something that perceives a world but is formed by it. Adaptations between agents and their environment take place in many respects, as they are functionally advantageous (cf. Yong, 2022, pp. 114f, 221-223, 228; Frith, 2007, p. 128).

One advantage of adaptations is that they make it possible to minimise the need to process representations in order to identify optimal actions. As an example, the sensory system for sound waves of female crickets is connected to their locomotor system in such a way that melodies produced by male crickets automatically lead to movements in the direction of their location, whereas all other sounds do not cause a reaction (Webb, 1993, pp. 1091-1093). A comprehensive analysis of all the sounds heard and filtering out melodies, as occurs to some extent in humans, can therefore be avoided. Another example is monkeys whose colour receptors are adapted to the colours of nutritious fruits, enabling them to recognise the fruits more easily and thus reducing the demands on perceptual analysis (Frith, 2007, p. 128). The adaptation of agents to their environments to achieve more efficient and effective goal fulfilment plays an important role in robotics (Hempel, 1966, p. 19). Brooks' robot Herbert, which was analysed in Section 2.5, was significantly more efficient than other models developed at that time, as it was strongly optimised to fulfil the objectives with minimal use of resources. For example, image analysis for object

[23] Consequently, it makes no difference whether an agent perceives the representations reflecting the correspondence directly from a world or by means of a computer, as is the case in the brain in a vat thought experiment of Putnam et al. (1981, ch. 1).

detection was avoided by using simpler but viable ultrasonic sensors instead.
Such adaptations of agents are particularly helpful in regard to executing skills, i.e., when particular actions have to be performed because of particular states of a world. Adaptions to support intelligence are much more difficult because states of a world have not yet been attributed a specific meaning, and it is therefore unknown which states of the world should be used in which way under what circumstances. Nevertheless, from a functional point of view, it seems advisable to design intelligent agents in such a way that they are able to manipulate a world as comprehensively as possible. On the one hand, this gives them more options for action and enables them to identify more favourable actions to achieve their goals. On the other hand, manipulating a world allows hypotheses to be tested and falsified, which allows world models to be developed with greater correspondence and thus greater viability. Shapiro (2019, pp. 80, 86f, 117) discusses the thesis of whether the nature of an agent's embodiment constrains or determines the concepts it can acquire, arguing that if the thesis "is correct, then human beings could not share thoughts with differently embodied aliens because they could not possess the same concepts". The nature of embodiment indeed influences the concepts that can be created, e.g., through the possibilities and limitations of perception, as well as the reasoning methods that can be applied. However, this does not imply that it is impossible for agents with different embodiments to have the same concepts. For example, a car may be perceived in different ways and the corresponding concepts may be created by different reasoning methods, but the created concepts may still represent the same states of the world and be functionally the same. Consequently, the development of different concepts due to different conditions and contingencies is possible but not inevitable. In addition, agents can synchronise their concepts through communication, as is common between humans; in this article, for example, through definitions and deliberations.

2.11 Conclusion

The aim of the article is to identify and analyse principles that have to be considered for the creation of AGI. Based on the analyses in the preceding sections, the following findings are drawn: The purpose of AGI is the fulfilment of given goals in a partially unknown world. To achieve these goals, AGI must develop skills, i.e., instructions for action that enable the fulfilment of the goals depending on states of the world. Novel skills for hitherto unknown conditions can be created by intelligence, which is based on the application of various reasoning methods such as deduction, induction and abduction, as well as other methods such as abstraction and classification. Due to the nature of perception, intelligence cannot grasp a world as it is but can only use representations that reflect the world indirectly and possibly incompletely and distorted. As representations correspond to the world, intelligence can draw

conclusions from them about the world using uncertain and contingent reasoning methods. This makes it possible to attribute functions to representations as to how they can be used to achieve goals; by doing so, representations are attributed meaning. The totality of all existing knowledge forms a world model, which contains, for example, all skills and which can be expanded with the help of reasoning methods and new perceptions. The value of a world model is functionally determined by its viability, i.e., its potential to fulfil the goals. Due to the uncertainty and contingency of the reasoning methods, many different possible viable conclusions can be drawn. As a consequence, the world model is constructivist, i.e., the conclusions drawn do not represent the world truthfully but only correspond to it. The methods of reasoning represent assumptions about the world; due to the NFL theorems, it is necessary to provide at least some assumptions as a basis for intelligence. However, intelligence is only successful if the assumptions apply to the world in which it is used, which is why they should be determined prudently. Overall, intelligence is considered an algorithm for an optimisation problem whose task is to find optimal actions to achieve particular goals in a partially unknown world. This interpretation relies on a naturalistic approach and does not require the assumption of mental features, such as consciousness, which are considered to be independent of intelligence. The performance of AGI is determined by how comprehensively it can perceive the world, how comprehensively it can manipulate the world, how comprehensively it can apply reasoning and other methods, and how efficient and consistent with the world the assumptions on which it is based are.

The considerations presented in this article also represent a constructivist-generated world model, a specific interpretation of all that is. From the author's point of view, based on conscious experiences, cultural influences, knowledge given at birth, and conclusions based on these, the considerations presented here appear to have the highest achievable correspondence with the world. Whether these considerations are viable, i.e., functional, and offer the possibility of creating AGI has yet to be determined. The considerations made here offer a new perspective on the development of AGI insofar as, in contrast to numerous other approaches, they focus away from the utilisation of knowledge towards the generation of knowledge by means of reasoning methods, in particular deduction, induction, and abduction. Abduction is a method that has so far received relatively less attention in the field of AI, but also in the field of philosophy; at the same time, it is the most powerful inference method, as it allows the generation of new, composed representations. Consequently, abduction, as well as other topics addressed in this article, requires a more detailed examination and further research.

The considerations developed in the article also allow for various considerations regarding generative AI approaches, which are currently gaining ground, particularly in the form of large language models and large multimodal models, and which are considered by many to be the closest to AGI currently available. While these models

deliver results that are considered impressive by many, they are based on an enormous amount of training data. They seem to be able to apply reasoning methods and solve unfamiliar problems, but only to a limited extent. From the perspective of the conception of intelligence developed here, these models are primarily, but not exclusively, based on skills rather than intelligence (ch. 3.4). The models also have the disadvantage that the knowledge provided to them does not represent very few fundamental assumptions about the world but an already highly processed and very specific interpretation of the world from a human perspective. As a result, the models are founded on representations similar to those of humans, which makes communication much easier, but the models cannot develop their own, possibly more viable representations.

Yet, this is precisely where the opportunity of AGI could lie, especially from a philosophical point of view, but also from a scientific point of view: AGI can receive much more raw and comprehensive representations of the world compared to humans and process them by other means, which, to draw on Neurath's metaphor, can allow it to build a new ship from scratch at a dock using better components. This new ship, an almost new interpretation of the world, could represent a comprehensive enrichment for humanity.

Chapter 3

An Analysis of Benchmarking Intelligence Based on the Results of OpenAI's o3 on ARC-AGI[1]

Recently, the generative AI model o3 from OpenAI achieved a high scoring of 87.5 % on ARC-AGI, a benchmark proposed to measure intelligence. This raises the question whether systems based on Large Language Models (LLMs), particularly o3, demonstrate intelligence and progress towards artificial general intelligence (AGI). Building on the conception of intelligence developed in the foregoing chapter, this chapter examines whether OpenAI's o3 exhibits intelligence and whether ARC-AGI is capable of benchmarking intelligence. An analysis of the ARC-AGI benchmark shows that its tasks represent a very specific type of problem that can be solved by massive trialling of combinations of predefined operations. This method is also applied by o3, achieving its high score through the extensive use of computing power. However, for most problems in the physical world and in the human domain, solutions cannot be tested in advance and predefined operations are not available. In consequence, massive trialling of predefined operations, as o3 does, cannot be a basis for AGI and ARC-AGI is not suitable as a benchmark for intelligence. To enable a comprehensive assessment of intelligence and of progress towards AGI, a new benchmark for intelligence is outlined that covers a much wider variety of unknown tasks to be solved and does not allow a high scoring without the application of intelligence.

[1] This chapter is based on an extract from an article published in collaboration with Hansueli Jud. The development of the argumentation and writing of the article was done by the author of this thesis, while the second author contributed through discussions and revisions (Pfister and Jud, 2025).

3.1 Introduction

The release of systems based on large language models (LLMs)[2], in particular OpenAI's ChatGPT in 2022, caused intense and ongoing debates about the extent of their intelligence. For example, Microsoft, one of the stakeholders in OpenAI, stated that the successor model "GPT-4 attains a form of general intelligence, indeed showing sparks of artificial general intelligence" (Bubeck et al., 2023, p. 92). Further statements that LLM-based systems represent artificial general intelligence (AGI) or at least major progress towards it have been made by OpenAI and other prominent AI companies, but also by AI experts, and in the media. At the same time, others take a more critical perspective on the performance of LLM-based systems, attributing their success not to intelligence, but to other factors such as the vast amounts of training data and the extensive computing resources used. The discussion has recently intensified again with the success of OpenAI's latest model o3 on the ARC-AGI benchmark, where it achieved 87.5 % on the semi-private test set; an achievement Chollet (2024) calls "a genuine breakthrough, marking a qualitative shift in AI capabilities".

ARC-AGI, originally called Abstraction and Reasoning Corpus (ARC), is designed as a benchmark for measuring general intelligence and was developed by Chollet (2019, pp. 46-58). In contrast to other benchmarks, ARC-AGI is not intended to measure the performance of an AI approach in a specific skill, but instead its ability to solve new, unknown tasks which it has not encountered before (cf. sect. 2.2). ARC-AGI consists of 1,000 unique tasks, of which 800 are publicly accessible and divided into 400 training tasks and 400 evaluation tasks. The remaining 200 tasks are divided into two private test sets. They are kept confidential to ensure that neither AI approaches nor their programmers can optimise for them in advance. One of the private test sets has been used as an undisclosed test set in various programming competitions since 2020, while the second one remains unused and confidential. In 2024, an additional semi-private test set with 100 newly created tasks was released to evaluate larger AI models that require API access and where the confidentiality of the test set can therefore not be guaranteed (ARC Prize, 2024).

Each task consists of a small number of example pairs and one or more test pairs, with each pair consisting of an input and an output grid. Each grid can have between one and thirty cells in width and height, with the two dimensions being independent of each other. Each cell can be in one of ten possible states, usually represented by colours for easier interpretation by humans. In each task, all inputs are manipulated according to a task-specific rule, which results in the corresponding outputs. For instance, a rule can be that all cells of a certain colour have to be changed to a different colour, or that the input grids have to be mirrored horizontally. All rules

[2]OpenAI's o3, but also its predecessors and comparable systems from other companies such as Google's Gemini and DeepSeek's R1 are based on LLMs at their core, but contain many additional modules that improve and extend their functions.

are based only on core knowledge, that is fundamental human beliefs such as the existence of objects or basic algebraic and geometric principles. To solve a task, the task-specific rule has to be first determined by analysing the example pairs and then applied to the test input(s) to generate the test output(s). A task is only considered solved if the submitted test output corresponds exactly to the correct solution in every single cell state; otherwise the task is considered failed. Each task is designed so that there is exactly one possible correct solution for each test output (Chollet, 2019, pp. 46-51).

Following the publication of the article introducing the Abstraction and Reasoning Corpus in 2019, a public competition was held on Kaggle in 2020, where the best approach achieved a 21 % success rate on the private test set (Chollet et al., 2020). In subsequent competitions in 2022 and 2023, the highest score reached was 30 % (Lab 42, 2023). In the 2024 Kaggle competition, with possible prizes totalling 725,000 USD, the winning approach achieved 53.5 % (Chollet et al., 2024a; Chollet et al., 2024b). Shortly thereafter, OpenAI's o3 achieved 87.5 % on the semi-private test set, which is intended to be similar in difficulty to the private test set. However, o3 was not subject to the computational restrictions imposed by the competitions; instead, its computational costs are estimated to be approximately USD 346,000 (ARC Prize, 2024).

Consequently, the question arises as to whether the success of o3 on the ARC-AGI benchmark, which was explicitly designed to test intelligence, is evidence that o3 exhibits intelligence – or whether ARC-AGI is only of limited suitability for measuring intelligence and other benchmarks are therefore needed to measure progress towards AGI. This chapter builds on the conception of intelligence introduced in section 2.2. Section 3.2 analyses the suitability of ARC-AGI as a benchmark for intelligence and for measuring progress towards AGI. This includes an analysis of the type of problem structure that ARC-AGI tasks represent, as well as the weaknesses that ARC-AGI possesses and the extent to which these can be overcome. Section 3.3 outlines a new benchmark for intelligence that is based on the definition of intelligence introduced in Section 2.2 and which is intended to enable a more comprehensive assessment of intelligence. Section 3.4 concludes with an evaluation of the performance of OpenAI's o3 on ARC-AGI.

3.2 Suitability of ARC-AGI as a Benchmark for AGI

Presenting o3's achievement on ARC-AGI, Chollet (2024) concludes: "ARC-AGI serves as a critical benchmark for detecting such breakthroughs, highlighting generalization power in a way that saturated or less demanding benchmarks cannot. However, it is important to note that ARC-AGI is not an acid test for AGI." Chollet

(2024) therefore proposes to develop a new version of ARC-AGI, a "next-gen, enduring AGI benchmark" in the same format. This poses the question to what extent ARC-AGI in its current form is suitable for measuring the capacity for broad generalisation, and to what extent and in what way it is possible to develop it further in the same format.

ARC-AGI is different compared to most other benchmarks used in the field of AI in that it is not designed to measure how good AI approaches are in a particular skill, but instead in their ability to generalise (Chollet, 2019, pp. 4, 53f). To accomplish this, the test set is kept secret, the tasks are designed to be diverse, and each task has only a few examples from which one has to generalise. By limiting the required knowledge to core knowledge, the emphasis is not on the application of existing knowledge, but on the ability to abstract and reason.[3] All these factors together place the focus on fulfilling the second interpretation of intelligence (sect. 2.2), i.e. the capacity to develop solutions for new, previously unknown tasks by means of generalisation. The minimalistic design of ARC-AGI tasks as simple, coloured grids, whose transformation can be described using core knowledge only, allow for easy development and testing of new AI approaches.

However, the minimalist and specific design of the ARC-AGI tasks also represents a very specific problem structure for the following two reasons: First, to solve an ARC-AGI task, it is necessary to determine the most simple transformation rule that describes the changes between the input and output example grids. The determined transformation rule then has to be applied to the test input(s) to generate the test output(s). Each transformation rule can be described by a combination of core knowledge. The entire core knowledge can be represented by a finite and small set of operations that determine certain properties of the grids or apply certain changes to them.[4] Although for each task a different transformation rule has to be determined and the large number of possible combinations of core knowledge allows a greater variety, the underlying problem structure is always the same: From the existing, small and finite set of potential core knowledge operations, those that together correctly describe the transformation must be selected and combined together. Second, ARC-AGI tasks represent a very specific problem structure as each task is required to have a single correct solution, i.e. there is exactly one correct output grid for each input grid. This makes it possible to test the correctness of possible transformation rules: A transformation rule is correct exactly then when it

[3] Chollet (2019, pp. 47-50) considers the core knowledge used to be explicitly described and complete. Yet it appears to be incomplete; for example, Boolean functions such as AND, OR, NOT are not mentioned but occur in several ARC-AGI tasks. In addition, it is unclear to what extent concepts that can be derived from described concepts are also considered valid. For example, from the included concept of addition, the concept of multiplication can be derived – a concept which is also used in ARC-AGI tasks. Equally, the concept of division and, with the help of this, the concept of prime numbers could be derived.

[4] An example of the implementation of core knowledge in the form of a finite and small set of operations is provided by Hodel (2024).

determines in every example pair for the input grid the correct corresponding output grid. Consequently, since the example pairs are given, it is possible to check whether a transformation rule is correct or not before submitting a solution by evaluating whether it generates for every example input the corresponding example output.
Both characteristics of the problem structure of ARC-AGI tasks allow the solution process to be considerably simplified in the following two regards: First, many problems require a solution process that can be described by a combination of exploration and exploitation (cf. Tromp, 2024). Exploration describes the process of representing a task in a form that allows a solution to be found; for example, to find the best route to a distant location, the task can be framed as a cost optimisation problem. Exploitation describes the process of finding the optimal solution within the representation of the task; for instance, by comparing the costs in money and time for different travel routes. While the exploitation of a problem representation can often be considered as an optimisation problem, exploration is considerably more challenging as it requires a suitable framing of the problem, i.e. the creation of a functional representation (cf. sect. 2.7, 2.8). This requires the identification of the relevant aspects that need to be considered as well as the creation of a model that represents the relationships between them (cf. sect. 4.2.3; Pfister, 2025a, sect. 5). The problem structure of ARC-AGI already implies a specific representation of the problem: For every task, the simplest possible transformation rule has to be found, which has to be composed of given core knowledge operations. Consequently, AI approaches can solve ARC-AGI by relying only on the exploitation of a given problem representation, without the need for exploration, i.e. the creation of a suitable problem representation, beforehand.
Second, the solution process can be considerably simplified in the following way: Since the correctness of a transformation rule can be tested on the example pairs, ARC-AGI allows, within the limited computational resources, for unrestricted trialling of possible transformation rules. While unrestricted trialling works well for ARC-AGI and other mathematical problems (cf. Trinh et al., 2024), such an approach does not work for many other types of problems: For many problems, especially in the physical world, but also in many other domains, one often has only one or at most a few attempts to check whether a solution is correct or not. For instance, pressing the wrong combination of buttons on a coffee machine will spoil the drink, and driving a car incorrectly can lead to a serious accident. There are ways to circumvent such problems, e.g. a robot, before grasping a cup of coffee, can simulate the grasping process in a virtual physical environment and thereby find a way to successfully hold the cup. However, this only works for domains that are sufficiently known so that all relevant aspects can be considered in the simulation – in other words, it only works for tasks whose conditions are already known, i.e. they must be realised as a skill. For tasks whose conditions are not known, this method does not work. Equally, it does not work for tasks that cannot be simulated for

other reasons; for example, because tasks are too complex, require too many computing resources, or actions must be performed faster than their simulations could be carried out.

In addition to the major weakness that the problem structure of ARC-AGI allows a much simpler solution finding process than many other problems, there are several other weaknesses that ARC-AGI possesses. For example, while the ARC-AGI benchmark limits the computing resources allowed for the solution finding process, it does not reflect the cost of training the approaches beforehand. Yet, training can significantly improve the rating of an approach; for instance, during the 2024 ARC-AGI competition, some participants trained their approaches on a large number of artificially generated ARC-AGI tasks. This is particularly of concern as training is a sign that an approach is not based on intelligence but on skills – whereas ARC-AGI intends to measure the former. Furthermore, although the test set is kept private, competition participants had the possibility to run their approaches several hundred times on it, allowing them to probe it and optimise their approach specifically for it. In summary, although ARC-AGI was created with the intention of being solvable only by broad generalisation and therefore by intelligence, it has several features that make it vulnerable to additional solution methods: The specific type of problem structure that ARC-AGI tasks represent is not only much easier to solve than many other types of problems, it also represents a very small subset of the huge diversity of possible problems, which supports the application of skill-based approaches. The possibility of massively trialling solutions allows approaches to test a large number of possible, low-quality solutions that may not be obtained by intelligence but, for example, by guided guessing. This, prior training, and probing of the test set allows ARC-AGI to be solved not by means of broad generalisation, but by skills-based approaches that are specifically optimised for ARC-AGI.

This raises the question of whether ARC-AGI can be improved to overcome its weaknesses while retaining the same format. Some of the issues can be overcome, for instance a new private test set can be used for which probing can be prohibited. However, the specific type of problem structure ARC-AGI represents is an inherent aspect of the current format and cannot be overcome by minor adjustments. Instead, addressing this aspect requires an entirely new type of benchmark that allows for other types of problems which represent a much greater diversity, require exploitation, and do not allow for massive trialling of possible solutions. Nevertheless, although ARC-AGI does not appear to be a sufficiently suitable intelligence benchmark for the future, it has to be concluded with Chollet (2024): "It's a research tool designed to focus attention on the most challenging unsolved problems in AI, a role it has fulfilled well over the past five years."

3.3 Towards a new Benchmark for Intelligence

The limitations of ARC-AGI lead to the question of how a benchmark can be designed that can be used to measure intelligence and progress towards AGI. A noticeable characteristic of the ARC-AGI benchmark is that it appears to be subject to Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure." This was evident in many places in the various ARC-AGI competitions. Instead of the approaches being developed towards AGI, they were developed to achieve the highest possible score on the test set: Core knowledge representations were optimised, ARC-AGI-specific skills were improved using artificially generated training data, and some participants submitted hundreds of approaches to probe and optimise for the test set. In order to avoid this, a future benchmark should have the greatest possible correspondence between the measure and the target, i.e. the development of (artificial) intelligence. To express it more directly: The best benchmark for intelligence is intelligence itself.
Consequently, an ideal benchmark should rate an agent as the more intelligent, the more efficiently it can achieve the more diverse goals in the more diverse worlds with the less knowledge (sect. 2.2). While human intelligence tests tend in this direction, they do so only to a very limited extent: They use time as a measure of efficiency, age as a measure of existing knowledge, and test various goals in various domains. Altogether, however, the tests are neither precise nor very diverse, and to a large extent they measure skills instead of intelligence. This is not least because human intelligence tests are designed to predict human performance in the human domain. In contrast, AGI, as a formal system, is neither bound to the human domain nor to the physical world. Instead, AGI can be given any type of goal which it has to fulfil in any type of world. These can be worlds that humans can access and understand, such as games, but also completely arbitrary worlds that can be simulated. For example, AGI can be situated in a simulation of a universe that is fundamentally different from ours. The universe could have more or fewer dimensions than ours, and fundamental aspects such as the laws of physics or the principle of causality could be altered. An AGI approach to be evaluated can be given any type of embodiment in such a universe – or none – as well as any type of goal to fulfil. The better and more efficiently it achieves the goals, the more it is rated intelligent.
A concrete example of the outlined benchmark could be as follows: All the approaches to be tested have to solve tasks in ten different worlds; nothing is known about the worlds in advance. One such world could be for example a simulation of Mars. In this world, the AI approach is embodied in a Mars robot, for which it must first develop an understanding. The approach must then fulfil the task of building an accommodation for astronauts, for which it must develop an understanding of the physical conditions on Mars. Another world could be, for example, the simulation of a gas planet in a four-dimensional universe in which the approach has to

produce a specific chemical element using an alien body. A further world could be simulated by a digital strategy game in which the AI approach has to win against human players. This requires the approach to develop an understanding of the world of the computer game and also a differentiated picture of agentness, which allows it to predict and anticipate the actions of the other players. A next world could be one in which the AI approach has no possibility of manipulating the world, e.g. through a body, but still has to make predictions about future states by analysing previous states of the world. Such a world could be as different as predicting the future development of a quantum world, or a habitat, or how well a newly launched product will sell based on available market data. In addition to these described worlds, the AI approaches are tested in several further worlds to ensure a greater diversity of the test set. An approach is never tested multiple times in the same world to prevent it from using previously acquired knowledge instead of intelligence. The approaches are evaluated according to how efficiently they fulfil the specified goals, for example how much time they need to do so and how comprehensively they fulfil them. The more efficiently an AI approach fulfils the more goals in the more worlds, the more intelligent it is rated. As the tasks illustrated here require intelligence at least on a human level, it is possible to start with the simulation of simpler worlds and goals. This makes it possible to identify AI approaches that only have a lower level of intelligence but are nevertheless more successful than others. As the level of intelligence of AI approaches increases, the worlds and the tasks to be performed in them can be made increasingly difficult.

The worlds tested in the benchmark may no longer be accessible or understandable for humans, as humans are limited by their capabilities and bound by their skills. Nevertheless, this does not pose an issue for AGI, on the contrary. For an AGI approach to be successful in such arbitrary worlds, a programmer must focus exclusively on the implementation of intelligence. Any implementation of world-specific skills, such as human core knowledge or a specific understanding of causality, would not only be pointless but even detrimental – the approach would be impaired by the skills if they are not feasible in the respective world (cf. sect. 2.3, 2.8). The benchmark is not perfect in that the types of worlds that can be generated still have a certain degree of conformity: they all need to be formalisable, executable with current computing limitations, and are limited by human imagination. Theoretically, this conformity allows the realisation of a skill other than intelligence that is tailored to these worlds: With enough training, an approach can be successful in these worlds to a limited extent, just as LLM-based systems are successful not through reasoning but through knowledge in some parts of the human domain. However, not only will there be a much greater diversity to manage with much less available knowledge, but the benchmark could also require that the training time of the approaches is taken into account and that the approaches have to be of a certain compactness.

Chollet (2019, pp. 20-24) argues that AGI should be benchmarked against human

intelligence, as intelligence would need to be tied to a precisely defined area of application and only those areas relevant to humans can be accessed. Yet this is not necessary, as the above assessment of the degree of intelligence shows: any goal can be specified in any world – only the degree of goal fulfilment has to be measurable. Instead, focussing on human areas of application harbours several risks: First, the development of AGI may focus on skills instead of intelligence, as has often happened in the history of AI. Second, it would limit AGI too much to problems from the human domain, although problems from other domains can also be of interest to humans. For example, AGI can discover the world of bacteria or of deep space, both domains in which human skills are likely to be of limited use, and make them comprehensible to humans. Lastly, there is a risk that human assumptions will be taken too much into account (cf. sect. 2.8). For instance, Chollet (2019, pp. 47-50) refers to core knowledge that represents very basic beliefs from a human perspective, but which are at the same time very specific and convey a very particular view of the world. The same applies, for example, to classical physics, which is regarded as a fundamental theory of our universe. However, it can only be applied to a limited range of physical phenomena and contradicts quantum physics; both are indications that the theory does not represent the true nature of the universe, but is a pragmatic model that fulfils human needs (cf. sect. 4.2.3). Providing classical physics as axioms to AGI would therefore limit its capabilities, as it would be constrained by these flawed beliefs.

In conclusion, to measure intelligence reliably, a benchmark should be created that provides AGI approaches randomly generated worlds that are as diverse as possible and whose only commonality is that each of it has some regularities. All AGI approaches are measured by how efficiently they can achieve various, as different as possible goals in these environments. To this end, the approaches must identify the often hidden regularities and utilise them to determine the best available measures to achieve the goals (cf. Pfister, 2025a). Many more details of the benchmark have to be specified and it has to be implemented in practice.[5] The benchmark outlined here should be universal, i.e. it should remain valid regardless of the approaches tested and their operating characteristics. Nonetheless, as with ARC-AGI, it is possible that adaptations to new developments will be necessary once the benchmark is applied, as benchmarking AGI involves the assessment of a moving target. This requires a continuous understanding of new approaches and their impact on the validity of benchmark tests in order to create a benchmark that addresses recognised shortcomings on the path towards AGI. Nevertheless, with the measure of intelligence so close to the target, the benchmark outlined here should require only minor adjustments and generally strongly support the development of new approaches that represent progress towards AGI.

[5] A possible environment for implementing an initial version of the benchmark could be a modified version of Genesis (2024), for example.

3.4 Conclusion

The analysis of ARC-AGI has shown that it cannot serve as a benchmark for intelligence and thus as a measure of progress towards AGI. Its simplicity, which makes it ideal for developing and testing new approaches, brings with it several weaknesses. Most importantly, the tasks have a very specific problem structure that allows the tasks to be solved exploiting a known problem representation without having to create one first, although this is often the more difficult part of solving problems. In addition, it enables a massive trialling of possible solutions, allowing a high score to be achieved by the massive generation of low quality solutions.

Based on these insights, it is possible to analyse o3's score of 87.5 % on ARC-AGI in more detail. To achieve the score, o3 incurred estimated computing costs of USD 346,000 – equivalent to USD 3,460 per task. A low-compute version of o3, achieving 75.7 % on the semi-private test set, incurred computing costs of USD 2,012 – equivalent to USD 20 per attempted task (Chollet, 2024). "The reason why solving a single ARC-AGI task can end up taking up tens of millions of tokens and cost thousands of dollars is because this search process has to explore an enormous number of paths through program space" (Chollet, 2024). Although this method can achieve a high score, given sufficient computing power, it cannot be regarded as very efficient.[6] Furthermore, this method is only suitable for a very specific type of problem, but not for most problems in the physical world or in the human domain, where massive testing of solutions in advance is not possible. The method also does not correspond well with the original intention of ARC-AGI: the development of new AI approaches that can reliably abstract and reason, and thus can determine the correct solution on the first or at least the first few attempts. While LLM-based systems appear to have some capacity for abstraction and reasoning – both processes considered fundamental to intelligence as shown in Section 2.8 – they do not appear to perform them reliably (Lewis and Mitchell, 2024; Qiu et al., 2023; Hong et al., 2024; Dziri et al., 2023; Jiang et al., 2024; Nezhurina et al., 2024). Instead, they seem to rely to a greater extent on memorisation, i.e. the application of skills (McCoy et al., 2023; Mirzadeh et al., 2024; Mondorf and Plank, 2024; Prabhakar et al., 2024; Yan et al., 2024; Wu et al., 2023).

Overall, o3's performance on ARC-AGI is not due to intelligence but due to the application of knowledge and computing resources that together enable an effective search in the given space of possible solutions.

This raises the question of how approaches can be developed that are centred more on intelligence. Building on the above, intelligence is not about how much data is processed, or how extensively it is processed, but about how it is processed. In other words, progress towards AGI requires a shift from datasets and computing resources

[6]The original ARC-AGI benchmark used in the competitions is subject to strict limitations in terms of computing power and would therefore not allow such a high score using this method. The result of o3 is only possible because the computing restrictions were waived for its testing.

towards the algorithm itself. It is hoped that the benchmark outlined in this article contributes to further research into AI approaches that focus on intelligence rather than skills, thereby supporting the development of AGI.

Chapter 4

Towards a Theory of Abduction Based on Conditionals[1]

Abduction is considered the most powerful, but also the most controversially discussed type of inference. Based on an analysis of Peirce's retroduction, Lipton's Inference to the Best Explanation and other theories, a new theory of abduction is proposed. It considers abduction not as intrinsically explanatory but as intrinsically conditional: for a given fact, abduction allows one to infer a fact that implies it. There are three types of abduction: Selective abduction selects an already known conditional whose consequent is the given fact and infers that its antecedent is true. Conditional-creative abduction creates a new conditional in which the given fact is the consequent and a defined fact becomes the antecedent that implies the given fact. Propositional-conditional-creative abduction assumes that the given fact is implied by a hitherto undefined fact and thus creates a new conditional with a new proposition as antecedent. The execution of abductive inferences is specified by theory-specific patterns. Each pattern consists of a set of rules for both generating and justifying abductive conclusions and covers the complete inference process. Consequently, abductive inferences can be formalised iff the whole pattern can be formalised. The empirical consistency of the proposed theory is demonstrated by a case study of Semmelweis' research on puerperal fever.

4.1 Introduction

Abduction is often described as an inference that allows one to infer a potential explanation for a given fact. However, there is no commonly accepted definition of abduction: It is the least theoretically understood type of inference and the "status

[1] This chapter was published as an article in Synthese (Pfister, 2022).

of abduction is very controversial. When dealing with abductive reasoning misinterpretations and equivocations are common" (Magnani, 2015, p. 313). One reason for this is that the term abduction is used by many quite different theories: Peirce (1958, 1998) introduced the term and developed two different concepts. Harman (1965) links Peirce's abduction with his own theory of Inference to the Best Explanation (IBE), which was significantly revised by Lipton (2004). IBE is often also called abduction (Campos, 2011), although many consider this to be highly misleading (cf. Park, 2015, pp. 228-234; McAuliffe, 2015). Moreover, ambiguity arises as some theories interpret abduction as a logical syllogism, while others view it primarily as a computational method or as a process of epistemic change (Beirlaen and Aliseda, 2014, p. 3749).

Regardless of the differences, many theories regard abduction as a cornerstone of scientific methodology (cf. Douven, 2017a, sect. 1.2). It is considered the most insecure but also the most insightful kind of inference since "all the ideas of science come to it by the way of Abduction" (Peirce, 1958, CP 5.145). It is the only kind of inference that allows the introduction of new kinds of concepts, which is also seen as an essential difference from inductive inferences (Campos, 2011, p. 428; Psillos, 2002, pp. 610f). For example, Psillos (2011, pp. 122, 144f) states that "no new ideas are generated by induction" since "[t]he extra content generated by induction is simply a generalisation of the content of the premises. Hence, with enumerative induction, although we may arguably gain knowledge of hitherto unobserved correlations between instances of the attributes involved, we cannot gain 'novel' knowledge, i.e., knowledge of entities and causes that operate behind the phenomena" (cf. Peirce, 1958, CP 5.145, 6.475, 7.202; Minnameier, 2004, pp. 78f). In contrast, at least some kinds of abductive inferences allow for the introduction of new types of concepts. For example, Schurz (2008, p. 201) employs the common distinction between selective abductions and creative abductions, whereby the former ones "choose the best candidate among a given multitude of possible explanations" and the latter ones "introduce new theoretical models or theories".

When examining inferences, it is common to distinguish between the context of discovery and the context of justification. The context of discovery concerns the generation of a new hypothesis, whereas the context of justification concerns its quality. Although the distinction is often attributed to Reichenbach, it can be found earlier in Popper, the Wiener Kreis, Husserl, Whewell, and Herschel; some trace it further back to Kant or even to Aristotle and Euclid (Hoyningen-Huene, 1987, pp. 502f). The distinction allows one to analyse the execution of inferences: The context of discovery examines how one creates a particular hypothesis. The context of justification examines conditions under which an inference is good, but does not provide guidance on how to generate specific hypotheses. Nevertheless, although the distinction is helpful, it is arbitrary: The discovery of a new hypothesis is already influenced by justificatory considerations; otherwise, it would be very unlikely to

generate a promising hypothesis by only a few trials (Peirce, 1958, CP 7.220).
Many controversies in the 20th century with respect to the philosophy of discovery revolved around the disagreement of whether the generation of hypotheses is part of the scientific process or not (Schickore, 2018, sect. 3). Some, e.g., Popper (1959, pp. 30-32) and Hempel (1966, p. 15), argue that, unlike the justification of hypotheses, their generation is completely illogical and therefore not part of the scientific process. In opposition, others developed different accounts to capture the generation of hypotheses. Some accounts see discovery as a logical process, whilst others claim that it is not logical but follows analysable patterns, is governed by a methodology, or is at least amenable to philosophical analysis (cf. Schickore, 2018, sect. 6-9; Paavola, 2006b, ch. 3). Consequently, some theories see abduction as a process of generating hypotheses, while others see it as a process of evaluation or as a combination of both (Beirlaen and Aliseda, 2014, p. 3734; Paavola, 2006a, p. 93). Still other theories leave the generation and selection of hypotheses open due to the numerous unanswered questions and focus on other aspects of abduction (cf. Woods, 2011, pp. 242f).
In conclusion, although the discussion of to what extent abductive inferences can be formalised is considered important (cf. Psillos, 2011, p. 148), there is so far no consensus. As Schurz (2016, p. 496) states, the major challenge is therefore to find out whether there are formally explicable rules and strategies that allow the execution of abductive inferences. This chapter aims to address this challenge. The aim is to lay the foundation for a theory of abduction that overcomes the limitations of current ones and covers both the context of discovery and the context of justification. If possible, the theory should allow to formalise the process of abduction, which would allow its application in the field of computer science and artificial intelligence as well as its practical validation.
In order to achieve this goal, the chapter presents an approach of abduction that is based not on explanations, but on conditionals. The chapter is divided into seven sections. Section 4.2 examines various important properties of abduction based on an analysis of Peirce's retroduction and Inference to the Best Explanation. Section 4.3 offers a discussion of conditionals and, in particular, inferentialism. Building on all this, a definition of abductive inferences founded on conditionals is given in Section 4.4. The different types of abductive inferences are discussed in Section 4.5, in which moreover the use of analogies in patterns is explored. Section 4.6 examines the conditions under which abductive inferences can be formalised, and finally a conclusion is drawn in Section 4.7.

4.2 Properties of Abductive Inferences

4.2.1 Introduction of New Concepts

In his later works, Peirce (1958, CP 5.189; 1998, EP2 p. 231) introduces his revised concept of abduction, often referred to as retroduction[2], for which he provides the following definition:

> The surprising fact, C, is observed;
> But, if A were true, C would be a matter of course.
> Hence, there is reason to suspect that A is true.

Peirce (1958, CP 5.188) regards abduction as an inference that allows new concepts to be introduced. This seems to contradict the definition above, since the concept A derived by the conclusion is already given by the second premise and is therefore not new. However, as Anderson (1987, p. 25) explicates, the premise is not to be understood in the sense that it actually already contains the new concept A but "in the sense that there is a logical relation between premises and conclusion". The definition specifies only the logical order, but not the temporal order. Consequently, the concept A can be newly introduced in both the premise and the conclusion at the same time (p. 35).
The introduction of the new concept A is achieved through a creative act, which Peirce (1998, EP2 p. 227) describes as follows: "The abductive suggestion comes to us like a flash. It is an act of insight, although of extremely fallible insight. It is true that the different elements of the hypothesis were in our minds before; but it is the idea of putting together what we had never before dreamed of putting together which flashes the new suggestion before our contemplation." The origin of all abductive insights lies in perception, which is the basis of all knowledge (Rosenthal, 2004, p. 193). Perception leads to perceptual judgements that are formed into abductive conjectures (Campos, 2011, p. 428). However, there is no clear distinction between perceptual judgments and abductive conjectures; rather, the "abductive inference shades into perceptual judgment without any sharp line of demarcation between them; or in other words our first premises, the perceptual judgments, are to be regarded as an extreme case of abductive inferences, from which they differ in being absolutely beyond criticism" (Peirce, 1998, EP2 p. 227).
Closely related to perception is imagination. Both have signs as semiotic outcomes and complement each other. As Campos (2011, p. 429) states: "When a perceptual judgment disrupts our expectations and presents us with a problem, the imagination works to form schemata or diagrams of the situation, searching for explanations. In

[2]Peirce uses the terms abduction, hypothesis, and retroduction. While the term hypothesis is often associated with his earlier concept of abduction, the term retroduction is used more often to name his later concept; yet, he uses all three terms in his later works as well (Paavola, 2006b, pp. 40f).

the case of abduction, explanatory hypotheses are signs – diagrams that rearrange the relations among facts so as to explain them. Sometimes new elements (explanatory facts) are introduced into the diagrammatic hypothesis to explain the perceived, unexpected facts. 'Diagrams' or explanatory schemata may include formalized theories, equations, statistical models, figures, representations of atomic or molecular structures, and so on. The abductive insight consists in associating or relating explanatory and perceived facts in a novel way." The capacity for abductive insight is an instinctive endowment of humans that enables them to find a correct hypothesis within a small number of guesses, despite the myriads of possible hypotheses (Peirce, 1958, CP 7.220).
In summary, the creative act leading to an insight which introduces the new concept A is an immanent part of abduction. For this reason, Peirce describes abduction as both an insight and a logical inference. He explicates "that abduction, although it is very little hampered by logical rules, nevertheless is logical inference, asserting its conclusion only problematically or conjecturally, [...] but nevertheless having a perfectly definite logical form".
Since Peirce's definition describes creative insights as instinctive, it does not allow for a fully formal account of abductive inferences (cf. Tschaepe, 2014, pp. 121-124). In comparison, Schurz (2016, p. 494) provides the following formal structure of abductive inferences:

> Premise 1: A (singular or general) fact E that is in need of explanation.
> 'Premise' 2: A background knowledge K, which implies for a hypothesis H that H is a possible and sufficiently plausible explanation for E.
>
> ---
>
> Abductive conjecture: H is true.

Similar to Peirce's account, the hypothesis H of the conclusion is already referred to in the second premise. In contrast, the background knowledge only supports the hypothesis H but does not necessarily contain it itself. Besides that, unlike Peirce, Schurz does not presuppose a creative act of insight. Instead, the background knowledge K can imply in a purely formal way that a hypothesis H is a possible and sufficiently plausible explanation for the fact E. Consequently, Schurz's account allows for fully formalised abductive inferences that introduce new concepts in the conclusion that are not part of the premises. Nevertheless, non-formalisable abductive inferences can also be represented: this through the background knowledge representing a non-formal process such as Peirce's intuitive creative act.

4.2.2 Surprisingness and Observability

Peirce requires the fact C to be surprising. The characteristic of surprise can trigger an abductive inference: Since surprising facts do not match our expectations, they

can lead to promising new insights (Paavola, 2004, p. 274). However, for the inference itself, the surprisingness of the fact does not matter, as it does not influence the generation or justification of the hypothesis. In addition, there are also many non-surprising circumstances in which abduction can be insightful, e.g., when results of a scientific experiment are to be further investigated. Therefore, even though surprisingness can be an indicator for promising investigations, its necessity should be dismissed.
Schurz (2008, p. 216; 2016, p.495) requires for all kinds of creative abduction that the facts are observable.[3] However, abductive reasoning is desirable and used for unobservable facts, e.g., for the structure of molecules or radiation. Schurz (2008, p. 206; 2016, p. 499) also requires for all kinds of hypothetical cause abduction that the inferred hypothesis is unobservable. Yet, abduction is used for observable causes as well; for instance, one concludes that some birds are flying away because a predator is approaching. It seems that the (non-)observability of a fact is relevant for the subsequent examination of a hypothesis, but not for the inference itself. Additionally, the meaning of the fact C should be understood in a broad sense. It could be not only a fact that is known to be true, but also, for example, a hypothesis of another inference or an assumption of a thought experiment.

4.2.3 Process of Abduction

A complete theory of abduction must provide a precise and complete description of how abductive inferences are performed. This is true for both the context of discovery, i.e., how a specific hypothesis is generated, and the context of justification, i.e., how the quality of a hypothesis is evaluated. In the following, the analysis focuses primarily on Peirce's retroduction and on Inference to the Best Explanation (IBE), which are considered the most popular theories of abduction.
IBE's basic idea is that "explanatory considerations contribute to making some hypotheses more credible, and others less so" (Douven, 2017a, sect. 4). Thus, given a multitude of abductive hypotheses, IBE allows one to determine which is the best hypothesis, i.e., the one most likely to be true. Different accounts of IBE suggest varying explanatory virtues that make a hypothesis preferable (cf. Cabrera, 2017, pp. 1248-1250). It is still under discussion, which explanatory virtues should be considered.
In addition, it is unclear why explanatory virtues are an indicator of truth (Cabrera, 2017, sect. 3). At least some of the suggested virtues, e.g., precision and scope, are non-confirmatory and only informational virtues: they do not indicate which hypotheses are true but rather which provide greater informational content and meet the goals of science (Cabrera, 2017, sect. 3.3, 5.1). Hence, some (Cabrera,

[3]Likewise, Peirce speaks of the fact C to be "observed". However, in his time, the term had a much broader meaning. Therefore, it is uncertain whether he requires the fact C to be tangible or not.

2017; Dawes, 2013; Jones, 2018) suggest that IBE is not about justification but about pursuit, that is, identifying hypotheses worthy of further investigation. This view is also supported by practice: Darwin's hypothesis of heredity, pangenesis, fulfilled explanatory virtues but was rejected by the biological community because of missing empirical evidence. Similarly, the chromosome theory offered overwhelming explanatory power, but could not gain acceptance until both the existence and the causal power of chromosomes were demonstrated in subsequent experiments (Novick and Scholl, 2020, sect. 3, 5). Furthermore, if one considers explanatory virtues not as an indicator of truth but of informational content, one can explain why scientists accept[4] contradictory hypotheses and theories. For example, quantum mechanics and general relativity are incompatible, but both have great explanatory power (and are empirically successful). Many scientists do not believe them to be true but accept both because they provide a solid basis for further reasoning (cf. Dawes, 2013, sect. 1.3).

Peirce proposes several virtues that abductive inferences should fulfil: They should be simple, natural, and plausible to us (Peirce, 1958, CP 6.447) and should cost us as little effort as possible (CP. 5600, 7.220). They should explain all relevant facts (CP 7.235), have a unifying power (CP 7.221, 7.410), be licenced by existing background beliefs (Psillos, 2011, p. 136) and their plausibility should be discriminated from their antecedent likelihoods (Peirce, 1958, CP 5.599). Finally, hypotheses should be experimentally testable by entailing deductive and inductive predictions (CP 7.220). Peirce argues that science is severely limited by economical constraints: "the process of verification [...] is so very costly in time, energy, and money" (CP 5.602). The suggested virtues allow one to determine which hypothesis can be tested most efficiently and should therefore be investigated further first (Peirce, 1958, CP 7.220, 5.602; McKaughan, 2008, pp. 452-458). As Peirce (1958, CP 1.120) states, "[t]he best hypothesis [...] is the one which can be the most readily refuted if it is false. This far outweighs the trifling merit of being likely". Thus, his proposed virtues are not about justification, but about pursuit worthiness.

To justify an inferred hypothesis, Peirce advocates determining by deduction necessary consequences that follow from it. Their truth can be tested experimentally and, by induction[5], it can be concluded that if the consequences of the hypothesis are true, then the hypothesis itself is true (Peirce, 1958, CP 7.203, 7.206). However, besides that, Peirce remains rather general and does not provide specific methods or concrete conditions under which a hypothesis is considered justified. One reason for this is that Peirce (1958, CP 7.679f, 5.173, 2.753; 1998, EP2 pp. 443f) considers

[4] A hypothesis or a theory is considered acceptable here in case it is worthy of commitment as a research program (Cabrera, 2017, pp. 1267-1270).

[5] An analysis of Peirce's understanding of induction, which includes three different types, is provided by Fann (2012, pp. 32f).

the human instinct to have an innate tendency "to conjecture rightly".[6] Thus, the justification is already provided by the human endowment and the correct hypothesis can be found within a few trials through experimentation. Overall, many regard Peirce's theory as one of discovery rather than justification (e.g. Minnameier, 2004; Campos, 2011; Douven, 2017a, Supplement: Peirce on Abduction).

As far as the context of discovery is concerned, Peirce's considerations are quite detailed (cf. sect. 4.2.1). Yet, since the "abductive suggestion comes to us like a flash [and] is an act of insight", our explanatory suggestions "are not subject to rational self-control" (Peirce, 1998, EP2 p. 227). Only once they have been created can we access them logically. Peirce thus describes the process of discovery in great detail, but he does not provide a method – indeed, he rejects its possibility – by which one can deliberately create abductive hypotheses. Instead, we must rely on our instinctual human endowment (Peirce, 1958, CP 7.220).

IBE is viewed primarily as a theory of justification, where candidate hypotheses are usually already given (cf. Douven, 2017a, Introduction; Lange, 2022, p. 87). Nevertheless, there are at least some approaches that address the context of discovery. For instance, Lipton (2004, pp. 59, 149-151) proposes IBE as a two-filter approach: The first filter generates a set of promising hypotheses by contrastive analysis and consideration of background knowledge. The second filter, based on explanatory virtues, selects then the best hypothesis among the generated ones. Lipton illustrates this approach with the research of Semmelweis, who investigated why cases of puerperal fever were much higher in one clinic of the Vienna maternity hospital than in the other.[7]

According to Lipton (2004, p. 83), the generation of new hypotheses begins with a contrastive analysis: For the fact to be explained, one needs a foil with a similar history, because "this sharply constrains the class of hypotheses that are worth testing". For example, Semmelweis was able to compare the conditions of the two clinics with each other, as well as with those of women who had street births on the way to the hospital (Semmelweis, 1861, pp. 2-4, 43-46; Carter, 1983, p. 49). As Lipton (2004, p. 149) notes, contrastive cases will never have just one difference, but many. To further reduce the number of possible hypotheses based on these differences, Lipton (2004, pp. 139, 149-151) suggests relying on background knowledge. It allows

[6] Peirce (1958, CP 5.602) states "that man has a certain Insight [sic], not strong enough to be oftener right than wrong, but strong enough not to be overwhelmingly more often wrong than right". This seems to contradict some of his other statements in which he argues that "proposals for hypotheses inundate us in an overwhelming flood" (Peirce, 1958, CP 5.602). The contradiction can be resolved by considering both aspects as two successive steps in the hypothesis generation process. When one experiences a surprising fact, there are an infinite number of possible explanations. But out of these myriads of possible hypotheses, the human instinct intuitively considers only a few promising ones, of which one becomes aware.

[7] Since Semmelweis' investigation is a popular case study within the philosophy of science, it is not presented here in its entirety. The original German text of Semmelweis (1861) is available online, and a translated excerpt of important passages in English is provided by Carter (1983).

considering already known explanations, determining the unificatory virtues of the hypotheses, and providing explanatory standards. For instance, Semmelweis (1861, pp. 4-10; Carter, 1983, p. 51; Scholl, 2013, pp. 67-72) rejected epidemic factors and focused on endemic ones, as only the latter could explain why only one but not both clinics had high mortality rates. Moreover, Semmelweis (1861, pp. 32f) rejected the hypothesis that puerperal fever could be caused by fear of death, as this was not compatible with his background knowledge: he could not imagine how a mental state could lead to the strong physical manifestations of puerperal fever.

But even if one can further narrow down the number of potentially interesting differences, e.g., to endemic factors, there is still an infinite number left that needs to be considered. Semmelweis (1861, pp. 4-39; Carter, 1983, p. 52) considered delivery positions, exposure to a priest giving the last rites, rough examinations, and many other differences. But despite his detailed investigation, still many more possible explanations would remain that fit well with the background knowledge: such as poisonous air from a nearby factory, inadequate cleaning of the place, or dangerous behaviour by non-examining personnel. Hence, taking background knowledge into account may increase the chances of finding important differences more quickly, but it does not solve the problem of multiple differences as Lipton (2004, p. 128) intends. Moreover, the method is highly dependent on the availability of suitable contrastive cases. Semmelweis was in the fortunate position of being able to compare two very similar clinics from the same hospital; had there been only one clinic, it would have been much more difficult to find a promising contrastive case. For other cases, e.g., the discovery of gravity or the explanation of heredity, it is not clear how to find suitable contrastive cases at all.

After many unsuccessful attempts, Semmelweis finally succeeded in identifying the cause of the increased rate of puerperal fever in one of the clinics: There, medical personnel regularly performed autopsies before examining women in labour. In doing so, they transferred "cadaverous particles"[8] that infected the women and caused the fever. However, Semmelweis did not reach the conclusion by comparing differences between the two clinics and identifying the performance of autopsies as a relevant one. Instead, one of his colleagues was pricked with a knife while performing an autopsy and developed all the symptoms of puerperal fever before eventually dying. Semmelweis (1861, pp. 52-55; Carter, 1983, p. 52) was certain that the cause for his death was the autopsy knife that contaminated him with cadaverous particles. By analogy, Semmelweis concluded that the particles were also transmitted to the women in labour, through the hands of the medical personnel.

Similarly, a while later there was another accumulation of cases of puerperal fever. From this, Semmelweis (1861, pp. 59f; Carter, 1983, p. 54) concluded that puerperal fever "is caused not only by cadaverous particles adhering to hands but also by ichor

[8] More accurately, puerperal fever is not caused by cadaveric matter but by bacteria living on it.

from living organisms". Again, the conclusion was reached by analogy and not by a contrastive analysis that revealed relevant differences.[9]

Lipton's two-filter approach suggests that once several potential explanations have been generated, one uses the second filter, based on the explanatory virtues, to determine the best, i.e., the actual explanation. Lipton (2004, pp. 89f) argues: "When Semmelweis inferred the cadaveric hypothesis, it was not simply that what turned out to be the likeliest hypothesis also seemed the best explanation: Semmelweis judged that the likeliest cause of most of the cases of childbed fever in his hospital was infection by cadaveric matter because this was the best explanation of his evidence." However, this description is not accurate: Semmelweis did not develop a range of possible explanations, evaluated their explanatory power, and chose the best one. Instead, Semmelweis developed and tested one hypothesis after another over a period of three years until he found one that could be experimentally verified. It was thus not an inference to the best explanation, but to the only one (Paavola, 2006a, p. 106).

Lipton (2004, pp. 90, 149) is mindful of this discrepancy and argues that Semmelweis was in a fortunate position, but typically several candidate explanations remain and then explanatory virtues come into play. Nevertheless, Lipton is also aware of the role of experimentation and the elimination of hypotheses until only one remains. The importance of experimentation is also evident in Semmelweis' case: Semmelweis, as well as others in the scientific community, did not accept his hypothesis until he could experimentally support it in clinical interventions and in several animal experiments (Semmelweis, 1861, pp. 55-58, 76-80; Scholl, 2013, pp. 72-75). Other practical examples, such as the discovery of AIDS (Bird, 2010, pp. 349f) or the heredity theory already mentioned (Novick and Scholl, 2020), provide further support for the preference for this type of justification: In both cases, explanations were accepted not by their explanatory virtues but by empirical verification and the elimination of all other hypotheses available.

In conclusion, both Peirce's retroduction and IBE fall short of providing a precise and complete description of how abductive inferences are performed. Peirce's retroduction does not concern the justification but only the generation of hypotheses, and although the discovery is described in great detail, it remains inaccessible as it is considered as an instinctual human endowment. IBE offers methods for both generating and justifying hypotheses, but they fall short from both a theoretical and a practical perspective.

[9]Precisely, Semmelweis found differences with regard to the cause of childbed fever. For example, he observed that in the clinic with the higher infection rates, neighbouring patients frequently fell ill together, while in the other clinic patients fell ill in a scattered manner. However, Semmelweis was not able to use this difference to find the cause; in fact, it led him away from the correct solution: From the scattered distribution Semmelweis (1861, pp. 47f; Carter, 1983, p. 50) concluded "that puerperal fever was not a contagious disease and that the disease was not spread from bed to bed by pathogens".

4.2.4 Explanatoriness

Peirce, Lipton, and many others state that the main purpose of abduction is to provide an explanation for a given fact. So far, however, there is no generally accepted theory of explanation. Proponents of IBE do not consider this as problematic: IBE and other abductive theories do not presuppose any particular explanatory theory, but are compatible with at least most of them (Lipton, 2004, p. 2; Cabrera, 2017, pp. 1250f). However, the underlying explanatory theory does significant conceptual and justificatory work; if it is not specified, the central element of IBE is missing (Cabrera, 2017, pp. 731f). For example, as long as the explanatory theory is not specified, it is not clear which hypotheses qualify as explanations and therefore, among which hypotheses the best explanation should be chosen (cf. Klärner, 2013, pp. 57-61).
In addition, explanatory theories influence the coverage of IBE: For example, Lipton (2004, pp. 30-33) theory of explanation allows only causal explanations, although non-causal explanations also exist, e.g., in mathematics, philosophy and physics. This not only makes it impossible to provide explanations for non-causal circumstances (cf. Klärner, 2013, pp. 202-204), but also calls into question the applicability of IBE in general: It may be that even if causal explanations are possible, the best explanation is a non-causal one. Thus, if the set of available explanations contains only causal explanations, the best explanation may not be considered, and another, wrong explanation may be chosen instead.
Moreover, many explanatory theories, such as the presently discussed counterfactual theory of explanation (Reutlinger, 2018, pp. 78-81), do not provide any explanatory virtues. Yet, these virtues are required by IBE to determine which is the best explanation amongst the possible ones. IBE furthermore requires that the explanatory virtues enable comparative evaluation and, if there are several, that they can be rated against each other (cf. Klärner, 2013, pp. 61-64, 117-121, 207-211).
To avoid the problem of not having a suitable explanatory theory, Cabrera (2020, pp. 744-746) suggests that IBE should not rely on a theory of explanation, but only on explanatory virtues themselves, since they do the intended justificatory work. Others question the claim that abduction is intrinsically explanatory at all, i.e., that abductive hypotheses have to be explanations. For instance, Park (2015, pp. 220-222) considers the requirement to be ill-founded and based not on theoretical motivations but only on practical ones, such as providing useful constraints.
Furthermore, not all types of abductively derived conclusions seem to be explanatory. Schurz (2008, pp. 230f) as well as Gabbay and Woods (2005, pp. 122f) remark that at least some kinds of abduction are implausible and purely instrumental, i.e., they provide true predictions but are unlikely to be true themselves. For instance, the action-at-a-distance equation "serves Newton's theory in a wholly instrumental sense. It allows gravitational theory to predict observations that it would not otherwise be able to predict" (Magnani et al., 2009, p.77). Such purely instrumental

abductions not only contradict IBE's pursuit of truth, they are also incapable of explanation, as they are false. Yet, instrumental abductions are of scientific interest because they provide otherwise unobtainable predictions. A similar kind of inference can be found in mathematics. In general, there, one reasons deductively from some given axioms to some target theorems. However, it is also possible to infer from given theorems to axioms (Easwaran, 2008, pp. 383-385; cf. Niiniluoto, 2018, ch. 2). As Baker (2020, sect. 2.2.2) notes, "the propositions of elementary arithmetic – '2+2=4', '7 is prime', etc. – are much more self-evident than the axioms of whatever logical or set-theoretic system one might come up with to ground them. [...] Deriving '2+2=4' from our set-theoretic axioms does not increase our confidence in the truth of '2+2=4', but the fact that we can derive this antecedently known fact (and not derive other propositions which we know to be false) does increase our confidence in the truth of the axioms".
The derivation of axioms from given theorems does not aim at explanatory results (Magnani et al., 2009, pp. 72, 122, cf. pp. 119-139). Rather, it should make it possible to discover suitable axioms for mathematics (Magnani et al., 2009, p. 72), to systematise uncontroversial facts, to prove further theorems (Easwaran, 2008, p. 383), and to discover new theorems (Schlimm, 2011, pp. 48f). Here, too, the conclusions are instrumental and do not necessarily lead to truth (Easwaran, 2008, pp. 384f). In addition, the relevance and applicability of truth in mathematics in general are still controversial (Baker, 2020, sect. 2.2.2; Easwaran, 2008, p. 384). Hence, an explanatory account does not seem to be able to capture the inference of axioms in mathematics. As a possible solution, Heron (2021) proposes an account to justify axioms that relies on theoretical virtues but not on explanations. In conclusion, it remains unclear why abductive inferences should be intrinsically explanatory. Instead, various kinds of abductively derived conclusions are instrumental, do not lead to truth, and neither should nor can explain the given fact. Thus, abductive inferences can provide explanations, and often they do, but they do not necessarily have to.

4.3 Conditionals as the Basis of Abduction

4.3.1 Special Properties of Conditionals

In consequence, it seems that abductions are not intrinsically explanatory but that for a given fact they allow one to infer another fact that implies it. Such an implication can be represented by a conditional of the form 'If A, [then] C'. The consequent C represents the given fact, and the antecedent A represents the to be inferred fact that implies the consequent. In many abductive cases, the implying fact A is taken to explain the implied fact C – but as shown above, while this is true in most cases, it is not true in all cases. The confusion arises because explanations are often ex-

pressed through conditionals, but not all conditionals express an explanation. In other words, being an explanation is not an intrinsic property of an abductive conclusion but a possible application for which it can be used. It therefore seems more promising to base abduction on conditionals. Conditionals allow one not only to infer explanations but all kinds of preceding facts. This includes non-explanatory facts such as instrumental models and axioms, which are common conclusions in science as well. Furthermore, conditionals have two special properties that lead to the potential for abductive reasoning:

First, conditionals are asymmetric: a conditional and its converse version, where the antecedent and the consequent are interchanged, are not logically equivalent ('If A, then C.' $\neq$ 'If C, then A.'). Only some logical operators have this property; in classical logic, material implication is the only asymmetric binary truth function.[10] The asymmetry of conditionals allows one to represent relations in which one proposition implies the other, but not vice versa. Such relations are common in science, where, for example, laws are represented by conditionals. Such relations are also common in reasoning and predictions to infer what follows from assumptions.

Second, conditionals allow one to infer from the truth of the antecedent to the truth of the consequent. Conditionals are not the only logical operator that allows one to infer from the truth of one proposition the truth state of the other. For example, it follows from the exclusive disjunction 'either p or q' and 'p' that 'not q'. Yet, the exclusive disjunction as well as the alternative denial let one infer from the truth of one proposition only the falsehood of the other. In contrast, the conditional and the logical biconditional allow one to infer from the truth of one proposition the truth of the other. The ability to infer the truth rather than the falsehood of a proposition is in general more informative, as science aims to find true rather than false statements.

Due to its asymmetry, a conditional only allows one to infer with certainty from the truth of the antecedent to the truth of the consequent, but not vice versa. The reverse inference from the truth of the consequent to the truth of the antecedent, called affirmation of the consequent (Godden and Zenker, 2015, pp. 88-103), is uncertain and is often considered a fallacy. This is because the consequent can be implied not only by the antecedent of the conditional, but also by another fact. Thus, for a high credibility of the conclusion, it must therefore be justified that the consequent is actually implied by the antecedent and not by something else (pp. 104-120). Abduction provides this justification by combining the two special properties of conditionals: It uses the valid entailment from the truth of the antecedent to the truth of the consequent to develop a justification that allows one to infer well-justified in the opposite direction – i.e., to infer uncertainly but plausibly from the

[10] The material nonimplication, the converse implication, and the converse nonimplication are also asymmetric binary truth functions, but they can be expressed as more complex versions of the material implication.

truth of the consequent to the truth of the antecedent.

4.3.2 Conditional Theory for Abduction

Material implication is a conditional theory widely used in logic, but it leads to counterintuitive results (Evans and Over, 2004, ch. 2, 3). Other conditional theories include mental model theory, suppositional theories, and inferentialism, of which especially the latter are currently under discussion (cf. Douven et al., 2018, pp. 51-53).

The suppositional theories are based on the Ramsey test (Ramsey, 1990, p. 155), according to which the acceptability of a conditional[11] can be determined as follows: One hypothetically assumes that the antecedent is true and adds it to one's stock of beliefs, makes minimal changes if necessary to maintain consistency, and finally assesses the acceptability of the consequent of the conditional. If the consequent is accepted, the conditional is also accepted; otherwise it is not. Suppositional theories differ in their details, e.g., with regard to the truth values of a conditional whose antecedent is false. For example, Stalnaker's (1968) possible worlds semantics regards such a conditional as true in case its consequent is true in the nearest world in which its antecedent is true. In contrast, Evans (2020, p. 62) argues that people always think about a conditional on the supposition of its antecedent, and hence cases with false antecedents are irrelevant.

Inferentialism is founded on the assumption that conditionals are used to express an inferential connection between the antecedent and the consequent.[12] A conditional is considered true iff its consequent follows argumentatively from its antecedent and possibly contextually relevant background knowledge (cf. Douven, 2015, pp. 35-43).[13] The inferential connection can be of various types and be based, for example, on a logical, heuristic, or causal relationship. Accordingly, the connection may consist of a series of deductive, inductive, or abductive inferential steps. A deductive connection is certain and based on logical necessities; an inductive connection is uncertain and based on statistical considerations; and an abductive connection is

[11] Unless otherwise stated, the chapter only refers to indicative conditionals, i.e., conditionals whose antecedents are in the indicative mood. Although some of the considerations also apply to subjunctive conditionals, they require their own analysis.

[12] Introductions to inferentialism are provided by Skovgaard-Olsen (2016) and Douven et al. (2018, pp. 52-54), and a general introduction to conditionals based on relevance connections is offered by Egré and Rott (2021, sect. 7).

[13] In addition, the antecedent must be deductively consistent with the background knowledge, else conditionals with a logically false antecedent could count as true (Douven, 2015, p. 38).

uncertain and based on explanatory considerations.[14]

Conditionals that have an inferential connection are called connected conditionals. In contrast, in unconnected conditionals, the antecedent and the consequent have no clear connection and are probabilistically independent of each other. Unconnected conditionals often seem strange or misleading, like: "If George Washington was the first president of the United States, then Paris is the capital of France." Nevertheless, most suppositional theories judge a conditional to be true in case both its antecedent and its consequent are true, regardless of whether it is a connected or an unconnected conditional (e.g. Evans and Over, 2004, ch. 9; Baratgin et al., 2013). Insofar as unconnected conditionals are considered strange or misleading, this is attributed to the violation of pragmatic requirements, i.e., requirements concerning the way speakers make meaningful utterances (Evans, 2020, pp. 64f; cf. Skovgaard-Olsen et al., 2016, p. 27). In contrast, inferentialism regards unconnected conditionals not only as a violation of pragmatic norms, but as genuinely defective. This, because they are not able to fulfil their function of expressing reason relations (Skovgaard-Olsen, 2016, sect. 2.2; Vidal and Baratgin, 2017, p. 778). Reason relations are necessary for reasoning, prediction, and argumentation: They allow one to infer from the antecedent to the consequent and to estimate which propositions increase or decrease the probability of other propositions.

Beyond explaining the strangeness of unconnected conditionals, inferentialism is also able to match intuition about the or-to-if principle and provides a solution to Gibbard's Riverboat argument (Krzyżanowska et al., 2014). Furthermore, it is capable of providing satisfying interpretations for complex cases that cannot be successfully interpreted by other conditional theories (Skovgaard-Olsen, 2016, pp. 575-577). Nevertheless, inferentialism is still under development and not all aspects have been clarified (Douven, 2017b, pp. 1150-1153). For example, since it is pluralistic and allows for different types of connection, it is not yet clear which connections are permissible and which properties they must fulfil. Furthermore, it is unresolved whether conditionals can only be either true or false, or whether they can also be neither true nor false, but void – which is how they are sometimes assessed in empirical studies (cf. Skovgaard-Olsen et al., 2017, p. 462).

Another unresolved issue is the determination of the probability of connected conditionals. One possibility is to use the conditional probability hypothesis $P(\text{if } A, \text{ then } C) = (P(C \mid A))$ as suggested by many suppositional theories (e.g. Evans and Over, 2004, ch. 9; Fugard et al., 2011; Evans, 2020). Alternatively, the probability can be determined by the strength of the inferential connection. The two evaluation

[14]The term abduction is understood here in a different sense than in the rest of the chapter. According to Mirabile and Douven (2020, p. 5), in an abductive connection, the antecedent is best explained by the consequent, which is therefore probably true. The definition builds on IBE and as such is subject to the same criticism described in Sections 4.2.3 and 4.2.4. The relationship between this notion of abduction and the one presented in the rest of the chapter is discussed in detail at the end of Section 4.4.

methods differ in the factors they take into account: The latter considers only the inherent inferential connection between the antecedent and the consequent; the former incorporates also other factors that influence the consequent. As an example, consider the conditional "If my neighbour throws a party, then I cannot sleep well at night." Given that the neighbour is only every other time so loud that one cannot sleep, the probability of the conditional is 0.5 according to both evaluation methods. Now, one additionally assumes that one cannot sleep well at night anyway due to insomnia. Then, based on the strength of the inference relation, the probability of the conditional is still 0.5, while according to the conditional probability hypothesis it becomes 1.

The conditional probability hypothesis thus alters the probability of uncertain conditionals in case the consequent is influenced by another, non-exclusive factor. Consequently, the probability of a conditional can change depending on other provided facts, although the inferential connection between its antecedent and its consequent remains the same. This seems incoherent with the purpose of conditionals to express a reason relation, since the probability reflects not only the relation itself but also unrelated factors. Therefore, evaluating the probability of a conditional based on the strength of the inferential connection seems preferable. Empirically, there is evidence both for inferentialism (Douven et al., 2018; Mirabile and Douven, 2020; Skovgaard-Olsen et al., 2019; Vidal and Baratgin, 2017) as well as for suppositional theories (Over et al., 2007; Fugard et al., 2011; Cruz and Oberauer, 2014; Baratgin et al., 2013). However, the ambiguous results can be explained by a variety of factors (Skovgaard-Olsen et al., 2016; Skovgaard-Olsen et al., 2019) and studies specifically comparing the two conditional theories provide support for inferentialism (Mirabile and Douven, 2020, p. 26; Skovgaard-Olsen et al., 2019; Krzyżanowska et al., 2021; Nickerson et al., 2019, pp. 61f; Krzyżanowska and Douven, 2018; Douven et al., 2022b).

In conclusion, inferentialism is able to provide a coherent understanding of conditionals in accordance with empirical results. Moreover, it accounts for the connection between the antecedent and the consequent – which can be used in abductive reasoning to develop a justification that the given fact, which constitutes the consequent, is plausibly implied by the antecedent and not by some other, unconnected fact. Hence, understanding conditionals by means of inferentialism provides a good basis for abductive reasoning.

4.4 Definition of Abduction

Based on the foregoing considerations, abduction is defined in this chapter as follows: For a given fact, an abductive inference infers a fact that implies it. The implication is represented by an inferential conditional, whereby the implying fact is the antecedent and the given fact is the consequent. There are several types of

abduction: Selective abduction allows one to infer an antecedent for a given fact by using a known conditional. Creative abduction allows one to infer an antecedent for a given fact by creating a new conditional. Creative abduction can be further divided into two types, depending on which kind of proposition is introduced as an antecedent: Conditional-creative abduction is based on a proposition that is already defined in the theory. Propositional-conditional-creative abduction introduces a new, so far undefined proposition.

The differentiation between the three types of abduction is important from a conceptual point of view because they allow one to add different types of new knowledge to an existing theory: Selective abduction relies on a known conditional and lets one infer only the truth of the antecedent, i.e., of a fact. Creative abduction, on the other hand, lets one infer not only the truth of an implying fact, but also of an inferential connection between the given fact and the implying one. A propositional-conditional-creative abduction moreover allows one to introduce a new proposition into a theory as an antecedent. A new proposition can be formed either by a new combination of existing propositions or by the introduction of a new term that is hitherto undefined. In both cases, the new proposition expresses a new concept and is therefore the most powerful kind of inference.[15]

Similarly, the differentiation between the three types is important for the execution of abductive inferences: Selective abduction uses a known conditional; thus, its implementation requires only a selection process to determine which conditional of the background knowledge to use for the inference. Conditional-creative abduction introduces a new conditional with a defined proposition as its antecedent; thus, a process is required to select a proposition of the theory and to create the conditional. Propositional-conditional-creative abduction introduces a new conditional with a new proposition; therefore, a process is required to create both a proposition and a conditional.

In summary, each type represents a different kind of inference, where both the conditional and the proposition are determined by either selective or creative processes. Nevertheless, the types do not instruct how the selective and creative processes are to be carried out: Selective abduction gives no guidance as to which available conditional should be chosen; and creative abduction does not specify which proposition to consider for the conditional to be created. Each type is neutral in terms of its implementation. Hence, different procedures can be used to select or create the proposition and the conditional. The procedures provide guidance on how to perform a specific abductive inference and are called patterns. A pattern consists of a set of rules for both generating and justifying an abductive conclusion and it

[15] In fact, not only abduction but also induction allows inferring a new proposition in the conclusion. However, induction allows one only to introduce as an antecedent a new proposition which is a generalised version of the propositions provided in the premises. In contrast, propositional-conditional-creative abduction allows creating a new proposition that is based not only on propositions from the premises but also from the background knowledge or on so far undefined terms.

covers the whole inference process. Justificatory rules are considered because they influence the generation process: they are intended to ensure a promising result, i.e., that the truth of the conclusion is as likely as possible.
Types and patterns are very distinct in their characteristics. There are three different types of abduction, each representing an inferential process with selective and creative components. Moreover, types are theory-independent, i.e., they do not presuppose any particular theory. In contrast, patterns are theory-dependent as their generative and justificatory rules are based on different assumptions, e.g., on the principle of causality. Furthermore, different methods can be used to perform the selective and creative processes, e.g., simple heuristics as well as complex statistical procedures. Consequently, there are an infinite number of patterns that rely on different theories and use different methods. As a result, the various patterns differ in their applicability, efficiency, and persuasiveness.
The differentiation between types and patterns has several advantages. It distinguishes between the conceptual power of types of inferences on the one hand and the generative and justificatory power of patterns on the other. Furthermore, it allows a clear distinction between selective and creative components of the inference process as well as a comparison of different patterns, e.g., of their underlying assumptions and their methods.
These considerations lead to the following formal structure of abductive inferences:

Premise 1: a fact F
Premise 2: a pattern P; i.e., a set of rules generating and justifying the conditional $A \rightarrow^{16} F$, with A being a fact that implies F
Premise 3: a background knowledge BK that is used by the pattern P

Conclusion: $(A \rightarrow F) \wedge A$

The conditional is only concluded in creative abduction. In selective abduction, the conditional is already known, and part of the background knowledge, i.e., the premises. In creative abduction, the truth of the conditional has to be concluded, since the justification of the truth of the antecedent relies on it. The conditional can be regarded either as an intermediate step to the conclusion of the antecedent or as a conclusion on its own. What is considered the main insight depends on the purpose of the inference; for instance, whether a cause or an inferential connection should be inferred.
The conclusion contains the conditional '$A \rightarrow F$'. In contrast, Douven (2015, p. 96) argues that in a so-called abductive conditional, the consequent best explains the antecedent, i.e., the abductive conditional has the form '$F \rightarrow A$'. The two conditionals are related in that the former is part of the conclusion, while the latter represents

[16] In this chapter, the sign $\rightarrow$ is used to express a conditional based on inferentialism.

the abductive inference as a whole. Accordingly, they express two different meanings and rely on two different inferential connections. Although the main purpose of abduction is to identify the fact that implies the given fact, both conditionals can provide additional insights. In case one is concerned with what one can infer from the truth of the given fact, the conditional representing the abductive inference as a whole is relevant. In case one is mainly concerned with what implies the given fact, the conditional stated in the conclusion of the abductive inference is of interest.
The inferential connection of the conditional '$F \rightarrow A$' is based on the abductive inference process. Therefore, the more credible the abductive inference, the higher the probability of the conditional being true. For example, the abductively inferred conditional "If Paula travels from Germany to Japan, then she travels by plane." is very likely because the abductive inference can be based on the strong argument that long distances are most often travelled by plane. On the other hand, the conditional "If the car does not start, then the battery is dead." is less credible because there are many likely alternatives, such as an empty tank or a blown fuse.

4.5 Types and Patterns of Abduction

4.5.1 Selective Abduction

Selective abduction is the best researched type of abduction (cf. Peirce, 1958, CP 2.636; Psillos, 2011, pp. 117-131). This is because it is rather simple: The inference starts with the given fact F. Then, a pattern selects from the background knowledge a conditional in which the fact F is the consequent, and the truth of its antecedent A is derived. The inference has the formal form:

$$\frac{\begin{array}{l} F \\ A \rightarrow F \end{array}}{A}$$

The credibility of the inference depends on many different aspects, e.g., the underlying formal system as well as the number of conditionals available that have F as a consequent. In case the background knowledge specified in a formal system contains every true statement and there is only one conditional that has the fact F as consequent, the inference is certain. In case there are several suitable conditionals available, the inference is uncertain and the pattern must provide a method to select the most likely one. Additional uncertainty arises if the formal system is incomplete or non-monotonic: then the fact F can also be realised by a fact for which the corresponding conditional is not listed in the background knowledge. This aspect illustrates the limitation of selective abduction: it can only infer antecedents that are already known to imply the given fact but not ones for which this is not known. To infer such, creative abduction is required.

Fully formalised patterns of selective abduction are provided in computer science, e.g., by Aliseda (2006), Flach and Kakas (2000), and are also discussed in psychology, e.g., by Thomas et al. (2008). An illustration of selective abduction can be found in Semmelweis' research on puerperal fever (cf. sect. 4.2.3): Semmelweis (1861, p. 38f; Carter, 1983, p. 47) examines several facts that are considered to have a possible influence on puerperal fever, e.g., hyperinosis, hydremia and plethora. However, since these known facts cannot explain why puerperal fever cases occur only in one clinic but not in the other, he dismisses them and suspects another, as yet unknown cause (Semmelweis, 1861, pp. 51f; Carter, 1983, p. 51).

4.5.2 Creative Abduction

Creative abduction infers that the given fact F is implied by a hitherto unrelated fact A. The implication is due to an unknown inferential connection between the two facts. Creative abduction therefore lets one infer not only the truth of the antecedent A, but also the truth of the conditional '$A \rightarrow F$' that expresses the inferential connection.

Schurz (2008, p. 218) argues that all creative abductions in science explain several mutually intercorrelated phenomena by inferring a new unobservable concept that is their common cause. Consequently, neither single nor unobservable facts can be explained nor observable causes inferred. However, these are not intrinsic limitations of creative abductive inferences, but result from the pattern used: Schurz's pattern uses statistical factor analysis and judges results by virtue of unification (Schurz, 2008, pp. 219-232). As a consequence, only causes that can explain several phenomena at once are considered worthwhile. However, also non-unifying creative abductions explaining only one fact can be scientifically insightful; for instance, in cases such as the appearance of a single fossil of an ancient fish at high altitude in the Andes or the brief dimming of a star. Schurz's creative abduction is also limited in that it allows only the introduction of new concepts, but not the use of already defined concepts as antecedent (Schurz 2008, pp. 216, 218; 2016, p.495). Nevertheless, creative abductions that infer already defined concepts can be insightful as well.

In contrast, the concept of creative abduction presented here overcomes these limitations by allowing for different patterns. Consequently, it can encompass both observable and unobservable facts as well as the inference of non-unifying and defined facts.

4.5.3 Conditional-creative Abduction

An abduction, in which a defined concept is concluded to imply the given fact F, is a conditional-creative abduction. It is selective regarding the implying fact and

creative concerning the inferred conditional that connects the implying fact and the given fact. It has the formal form:

$$\frac{\begin{array}{l}F\\[A]\end{array}}{(A \rightarrow F) \wedge A}$$

A is in square brackets in the premises to indicate that it must be a defined proposition, but its truth value may be unknown. The purpose of patterns of conditional-creative abduction is to determine which proposition available in the theory is most likely to be the antecedent of the given fact. A wide variety of methods and assumptions can be used for this. For example, patterns based on causal Bayes nets allow one to determine a structural link based on causal power by considering interventions and known mechanisms (Oaksford and Chater, 2020, pp. 121-125). Another pattern provides the search for spatio-temporal continuity: People have a strong tendency to assume a causal relationship between two events if they are no more than two seconds apart (Griffiths and Tenenbaum, 2009, pp. 662, 696). Other patterns are based on the search for similarities (Magid et al., 2015, p. 101) or by comparing the characteristics of the given fact and facts that can serve as possible antecedents (Magid et al., 2015, pp. 103-109). In general, theory-specific knowledge plays an important role in the selection of an appropriate proposition as antecedent: e.g. laws that explicate which types of proposition can imply which other types of propositions and thus the given fact. Hence, patterns used for conditional-creative abduction can rely on a large amount of background knowledge, which complicates their formulation.
An illustration of conditional-creative abduction is provided by Semmelweis: Having concluded that no known cause could account for the different rates of puerperal fever, Semmelweis considered facts that were known but not associated with puerperal fever so far. For instance, Semmelweis (1861, pp. 36, 51f; Carter, 1983, pp. 51f) considered the delivery position and the routes women had to take to their puerperium after giving birth. He obtained these facts by applying various generative patterns; e.g. looking at reasons for unwellness and illness in general, or comparing the two clinics and finding differences. Nevertheless, none of the possible reasons could be substantiated. Either they could not be justified during the inference process because they did not fit the background knowledge, or they could not be confirmed in subsequent experiments.

4.5.4 Propositional-conditional-creative Abduction

Propositional-conditional-creative abduction assumes that the given fact F is not implied by a fact already defined in the theory, but by a new, hitherto undefined one. It thus infers both the truth of a new fact and the truth of an inferential

connection between the new fact and the given fact. The inferential connection has to be inferred because it provides support for the truth of the implying fact. Propositional-conditional-creative abduction has the formal form:

$$\frac{F}{\overline{(A \rightarrow F) \wedge A}}$$

Schurz (2016, pp. 498-503) points out that there are cases where the given fact F is not simple but complex, i.e., consists of a plurality of facts. For example, the given fact can state that sugar, salt, sodium carbonate and copper sulphate are all soluble in water, insoluble in oil, have a higher melting point and conduct electricity. In addition, in some cases the multitude of facts subsumed in the given fact cannot be implied by a simple fact, but only by a complex one. For instance, using statistical factor analysis, the cultural characteristics of nations can be explained by the interplay of two main factors: the orientation between traditional-religious and secular-rational values on the one hand, and the orientation between survival and self-expression values on the other (pp. 506-508).[17] Neither factor alone would suffice to satisfactorily explain the cultural characteristics of a country. In some cases, only the inference of an antecedent that contains several facts leads to a satisfactory result.

Propositional-conditional-creative abduction consists of two steps: Once the number and relation of the facts of the antecedent have been determined, one must define them, i.e., introduce new propositions. A new proposition can be defined either by introducing a new term or by combining already defined propositions of the underlying theory in a new way. When defining the proposition more precisely, a newly introduced proposition may turn out to be an already defined one. It is possible to use separate subpatterns for determining the number of facts and for defining them. This is especially so since the definition of a new proposition often relies on other propositions from background knowledge and is therefore very theory-specific; whereas the inference of the number of possible facts in the antecedent is often based on more fundamental assumptions, e.g., statistical considerations. Furthermore, both steps can be performed independently of each other. For example, as Schurz (2016, p. 498) points out, the existence of the new proposition 'hydrophilic nature' was inferred long before the theory of atoms and molecules that allows it to be described.

Semmelweis' study of puerperal fever includes several illustrations of propositional-conditional-creative abductions. For instance, a commission suspected that the increased incidence of puerperal fever in one of the clinics was due to overly crude examinations by male students, especially foreigners (Semmelweis, 1861, pp. 48f;

[17] Glymour (2019) argues that the factor analysis used by Schurz is not suitable and suggests the use of a different pattern, which is based on other statistical methods and allows a more accurate abductive inference for this case.

Carter, 1983, p. 50). However, this hypothesis could not be verified in subsequent experiments. Another propositional-conditional-creative abduction finally led Semmelweis to the solution of the increased rates of puerperal fever. As mentioned above (cf. sect. 4.2.3), one of his colleagues, after being wounded with an autopsy knife, showed the same symptoms as those of puerperal fever and eventually died. Semmelweis ascribed his death to contamination with cadaverous particles in the course of the injury.
Based on this knowledge, Semmelweis inferred by analogy that the patients in the maternity ward also died from infection with cadaverous particles. However, in contrast to the case of his colleague, the infection was not transmitted by an autopsy knife, but by medical personnel who performed autopsies before examining the patients: Cadaverous particles remained on their hands, which were then absorbed by the genitals of the patients during the examination. In conclusion, Semmelweis inferred a new, hitherto undefined fact; the transmission of cadaveric particles via hands. This new fact is considered to have an inferential connection to the given fact, i.e., patients contracting puerperal fever, and is therefore its antecedent. Later, Semmelweis (1861, pp. 58-60; Carter, 1983, p. 54) performed two more inferences that illustrate propositional-conditional-creative abductions: First, a patient with uterine cancer was admitted and, subsequently, all patients in the room died. This led Semmelweis to infer that infectious matter can also be transmitted by ichor. Second, a patient was admitted with a healthy genital area but a discharging carious knee; again, most of the patients in the room subsequently died. From this, Semmelweis concluded that infectious matter can also be transmitted via air.

4.5.5 Analogical Patterns of Creative Abduction

Semmelweis' research shows that the use of analogies in abduction can lead to promising hypotheses. This chapter therefore explores in more detail how analogies can contribute to the generation and justification of hypotheses in patterns. Analogies are often given in the following form (Bartha, 2019, sect. 2.2; notation adapted):

P_1 is similar to P_k in certain (known) respects
P_k has some further feature Q_k

P_1 also has the feature Q_k, or some feature Q_1 similar to Q_k

This leads to the following formal representation:

P_1	given fact
$P_k \rightarrow Q_k$	with P_1 and P_k being similar in certain known respects
$P_1 \rightarrow Q_k$	conclusion 1: transfer of the same feature
$P_1 \rightarrow Q_1$	conclusion 2: transfer of a similar feature

In summary, an analogical inference transfers a characteristic, an inferential con-

nection with a consequent, from one proposition to another, similar proposition.[18] Depending on the nature of the analogy, the consequent can be altered and adapted to the similar proposition. Analogical conclusions are amplifying and uncertain because the inferential connection does not necessarily apply to the similar proposition as well. The legitimacy of analogical inferences rests on the assumption that similar conditions lead to similar results. As Mill (1974, p. 556) argues: "If $[P_1]$ resembled $[P_k]$ in all its ultimate properties, its possessing the attribute $[Q_k]$ would be a certainty, not a probability: and every resemblance which can be shown to exist between them, places it by so much the nearer to that point. If the resemblance be in an ultimate property, there will be resemblance in all the derivative properties dependent on that ultimate property, and of these $[Q_k]$ may be one."
Likewise, one can assume that similar results are based on similar conditions. Mill (1974, p. 556) continues: "If the resemblance be in a derivative property, there is reason to expect resemblance in the ultimate property on which it depends, and in the other derivative properties dependent on the same ultimate property." This assumption can be used to perform an analogical abduction: Given a particular fact, one searches for a fact, i.e., a proposition, which is similar and of which one knows the antecedent. One assumes that the antecedent is also that of the given fact – either in the form of the original proposition or in the form of a similar proposition adapted to the given fact. Formally, this can be expressed as follows:

Q_1	given fact
$P_k \rightarrow Q_k$	with Q_1 and Q_k being similar in certain known respects
$P_k \rightarrow Q_1$	conclusion 1: antecedent consists of the original proposition
$P_1 \rightarrow Q_1$	conclusion 2: antecedent consists of a modified proposition

In case the inferred antecedent contains the original or a defined similar proposition, it is a conditional-creative abduction. In case a similar, previously undefined proposition is inferred, it is a propositional-conditional-creative abduction.

4.5.6 Empirical Adequacy

Overall, the theory of abduction presented here provides a high degree of empirical adequacy with Semmelweis' research on puerperal fever. This does not necessarily mean that Semmelweis actually performed the processes of abduction described here – this is only an interpretation based on his writings, and there are many other interpretations of his research as well. In either case, Semmelweis' research provides an illustration of how the theory of abduction presented here could be successfully applied. It can represent all the inferences Semmelweis performed and their methods,

[18] For an assessment of similarity-based arguments in the context of inferentialism, see Douven et al. (2022a).

and it can explain in detail how the solution was finally reached through the use of analogy.

Furthermore, the abductive theory can explain the order in which Semmelweis executed the research process. First, he started from facts that were considered to be related to puerperal fever or diseases in general, for example, hyperinosis. When this was unsuccessful, he examined known facts, such as the delivery position and tried to establish an inferential connection to puerperal fever. When this also remained unsuccessful, he tried to identify new, hitherto undefined facts that imply puerperal fever. This order results from the fact that the selective and creative processes of abduction require different amounts of cognitive workload: Selective abduction requires only the selection of a known conditional in which the given proposition is the consequent, it is therefore the simplest type. Conditional-creative abduction uses a defined proposition, but there are usually many available, and an inferential connection must be created as well. Finally, propositional-conditional-creative abduction requires not only an inferential connection but also a new proposition to be created, which again requires additional cognitive effort.

With the abductive theory presented in this chapter, one can also explain why some inferences were performed together, but mostly each possible implying fact was inferred for itself. The first case, the inference of several possible causes at once, happened mostly at the beginning; this because selective abduction allows several available conditionals to be selected, compared with each other and evaluated together. In creative abduction, most possible causes were inferred individually, since each required its own process of generation and justification. The proposed abductive theory shows how the virtues of IBE, such as simplicity and coherence, can be used as guidance for the generative and justificatory processes in patterns. For example, simpler solutions are preferred because they are easier to generate; and more coherent solutions are preferred because a better fit with background knowledge reduces the likelihood of contradictions.

The contrastive inference approach proposed by Lipton (cf. 2004, sect. 2.3) can be carried out in the form of a pattern using, e.g., statistical factor analysis. However, the immanent problem of multiple differences becomes apparent here: The method is only successful in case the relevant data are taken into account. In Semmelweis' case, it would have been necessary to statistically compare the incidence of autopsies with the incidence of puerperal fever cases. But without knowing the connection, there was no reason to pay special attention to this small detail out of the myriad available. Therefore, this pattern is only successful if a large proportion of the data can be taken into account; otherwise, other patterns are preferable.

In the definition of abduction (cf. sect. 4.4), it was shown that abductive inferences not only allow one to conclude the fact A, but also, in the case of creative abduction, the conditional '$A \rightarrow F$'. In addition, the inference as a whole can be represented by the conditional '$F \rightarrow A$'. The different purposes of the three conclusions become

apparent in Semmelweis' case: His main interest was to determine A, the factor causing the high rates of puerperal fever in one of the two Vienna clinics. Besides that, the conditional '$A \rightarrow F$' was also of interest for him in several respects: First, he wanted to communicate it to other physicians so that they could avoid cases of puerperal fever in their own hospitals. Second, he used the conditional as a basis for further analogical abductive inferences to infer that cadaveric matter can be transmitted through ichor and the air as well. Finally, the conditional '$F \rightarrow A$' may be of interest in that if a case of puerperal fever appears, it can investigated whether it was caused by cadaveric matter. For instance, when Semmelweis (1861, pp. 81-85) heard of high rates of puerperal fever cases in the hospital at Pest, he suspected the transmission of cadaveric matter. His subsequent investigation revealed that the examination of women in labour was carried out by physicians who had performed operations before and thereby contaminated themselves thereby.

4.6 Formalisation of Abductive Inferences

Abductive theories vary widely in their understanding of the extent to which abductive inferences can be formalised, especially concerning the context of discovery (cf. sect. 4.1). Formalisation is understood here in the sense that it is possible to explicitly represent all information as well as all steps in which the information is processed. This means, a formalisable inference can be completely represented in a logical system and its implementation can be expressed in form of an algorithm that is Turing-computable. A formalisable theory of abduction has the advantage that it can be implemented in computer science and used for artificial intelligence.
The theory presented here defines abduction as an inference that allows one to infer for a given fact a fact that implies it. The implication is represented by a conditional which, following inferentialism, is considered true iff there is a connection from the antecedent to the consequent. Since inferentialism is pluralistic, the connection can be of different kinds, it can be deductive, inductive, or abductive. There is no unique criterion under which conditions a conditional connection is regarded as valid, and thus the conditional as true. Nevertheless, it is possible to provide rules to judge the validity of a conditional connection. For example, an inductive relation can be judged valid in case there are at least ten confirming and no falsifying instances. Such rules can be formally represented either as part of the axioms of a theory or as part of the context of justification of an abductive pattern. In summary, conditionals, which form the basis of abductive inferences, as well as their truth evaluation, can be formally represented.
Structurally, an abductive inference consists of the given fact, a pattern, and background knowledge. Since both the given fact and the background knowledge can be formally represented in the form of propositions, it follows that an abductive inference is formalisable iff its pattern is formalisable. A pattern is formalisable iff

every rule of the pattern, whether it concerns the generation or the justification, is formalisable and the pattern covers the complete inference process. Fully formalisable patterns exist for both selective abduction and creative abduction. For example, in selective abduction, the background knowledge is typically searched for all conditionals that contain the given fact F as a consequent. Subsequently, the available conditionals are ranked according to the joint probability of the antecedent and the strength of the conditional.[19] Finally, the antecedent of the highest ranked conditional is considered true (Aliseda, 2006; cf. Flach and Kakas, 2000). Patterns for creative abduction are more complex because they have to generate a new conditional and, depending on the type, a new proposition. Examples of patterns that allow specific kinds of creative abduction are Schurz's (2008, pp. 223-231) common cause abduction as well as BACON.4, which allows to search for lawful correlations in numerical data (Langley, 1987, ch. 4; cf. Jantzen, 2016, sect. 3.2f).
In conclusion, some abductive inferences can be formalised. However, this does not mean that all abductive inferences are formalisable: There are patterns, e.g., Peirce's intuitive creative act (cf. sect. 4.2.1), which are not formalisable and which therefore preclude the formalisation of an abductive inference. There seem to be several reasons why it is often claimed that abductions cannot be formalised: First, the underlying processes are often complex; therefore, it is difficult to explicate all rules of a pattern formally. Second, there is an infinite number of patterns because they are based on theory-specific knowledge, which makes them difficult to differentiate and capture. Third, the likelihood that the abductive conclusion is true is pattern-dependent, and many patterns yield a likelihood that is positive but not high enough to be considered feasible. Fourth, at least when real-world data are to be used as basis for abductive inferences, it is very difficult to formalise it, e.g., to determine the specific propositions – yet, this is crucial for successful inferences.

4.7 Conclusion

The goal of the chapter is to lay the foundation for a theory of abduction which is complete, i.e., covering both the context of generation and the context of justification, and formalisable, which allows its application in computer science and artificial intelligence. The theory proposed states that an abductive inference infers for a given fact a fact that implies it. By relying on conditionals, the theory stands in contrast to many other theories that consider explanations as one or even the cornerstone of abduction. Nevertheless, even though the theory does not consider abduction as intrinsically explanatory, it does not neglect the close relationship of abduction and explanation. Often abductive inferences can and do serve as explanations – but they do not have to.

[19] This approach is also empirically supported by Sebben and Ullrich (2021), who show that people tend to evaluate conditionals in this way.

Relying on conditionals rather than explanations as the basis for abduction offers several advantages. First, a theory of abduction based on conditionals allows not only the inference of explanations but of all kinds of preceding facts, which includes, for example, instrumental models and axioms. Second, when using conditionals, one can rely on two special properties of conditionals: they are asymmetric, and they allow one to infer the truth of the consequent from the truth of the antecedent. This inferential connection can be used to justify the conclusion in the opposite direction, i.e., to infer the truth of an antecedent from the truth of the consequent. This inference is uncertain, since the consequent may be implied by one of several known antecedents or even by an unknown one. Nevertheless, the inferential connections from the possible antecedents to the consequent can be used as a basis to generate and justify which antecedent actually implies the consequent. This justification is provided by patterns which can be based, for example, on probabilistic or analogical methods. Third, a theory of abduction based on conditionals does not require a theory of explanation. Since there is currently no generally accepted one, such a requirement would prevent the practical implementation and use of the abductive theory in computer science and artificial intelligence. Nevertheless, the theory presented here presupposes a theory of conditionals, which are also controversially discussed. This poses a challenge and requires further work; however, it is hoped that the open questions on conditionals – at least as far as abduction is concerned – can be resolved more easily than those on explanations.

The abductive theory presented in this chapter does not agree with IBE in many aspects, e.g., it is doubted that IBE's hypothesis generation is applicable and that explanatory virtues are sufficient to lead to the correct hypothesis. Nevertheless, IBE provides valuable insights. For example, empirical studies show that people actually assign extra value to the best explanation and thereby can achieve better results (Douven, 2020; Douven and Mirabile, 2018). Nonetheless, further research is required. For example, the studies only address the justification but not the generation of hypotheses, and the application is intrinsically context-sensitive (Douven, 2020, pp. 1, 11). Moreover, it is not clear by which explanatory virtues the quality of an explanation is to be judged – or whether non-explanatory considerations can play a role as well. Furthermore, it needs to be investigated whether a preference for the best hypothesis only occurs in abductive reasoning or also, e.g., in inductive reasoning. The first case would suggest that the preference is an intrinsic part of abduction, while the second case would suggest that it is a reasoning strategy based on economic reasons and independent of abduction. Another valuable aspect of IBE is its (explanatory) virtues, which can provide guidance as to which hypotheses are worth pursuing. Besides that, the theory discussed here incorporates components of many other theories; for example, Peirce's foundational understanding of abduction as well as methods of Schurz and others as patterns. Thus, although the approach presented here proposes a new understanding of abduction and aims to overcome

several limitations of current approaches, it also draws on them in many ways. It is hoped that the proposed theory will contribute to the ongoing discussion by providing an approach that is formalisable and computable. Additionally, it should allow for all kinds of abductive inferences to be covered while being sufficiently precise by enabling the use of specific patterns.
Many open questions remain that require further research. For example, more case studies need to be performed, and patterns as well as their formalisation and application need to be explored in more detail. Similarly, the combination of the presented theory of abduction with probability theories such as Bayesianism needs to be examined. Furthermore, the properties of complex antecedents and consequents, i.e., which consist of multiple facts, need to be investigated, as does the use of nested and counterfactual conditionals. Finally, especially for applications in computer science and artificial intelligence, a logic of abduction must be developed. The following considerations already show some possible characteristics of an abductive logic: Including probabilities, although not inherently required, allows the use of probability-based patterns as well as the determination of the likelihood of the conclusion. Non-monotonicity allows new statements to be added, e.g., experimental data that falsify previous abductive conclusions, which can lead to improved new conclusions. Other aspects, such as the derivation of additional assumptions and whether both a fact and its negation can imply a fact, are determined by the inferential conditional theory; these inferences are valid only if there is an inferential connection.

Chapter 5

The Role of Overdetermination and Alternative Implication in the Evaluation of Conditionals[1]

In this chapter, the suppositional account and different approaches of relevance conditionals are analysed on a specific type of conditional: Conditionals whose antecedent and consequent have a relevance connection, but where the acceptability of the antecedent has no influence on the acceptability of the consequent. Such conditionals occur in cases of multiple implication of a consequent, as in overdetermination. When evaluating such conditionals, the approaches examined lead to different and partly incoherent results. It is argued that approaches to conditionals should consider such conditionals acceptable, which is a challenge for e.g. approaches based on statistical measures. Furthermore, it is argued that the probability of a conditional should be evaluated only according to the strength of the relevance connection between the antecedent and the consequent, but not according to other relevance connections. It is shown that only two approaches correctly evaluate such conditionals, one of which, inferentialism, may provide a basis for a coherent theory of conditionals.

5.1 Introduction

Conditionals play an important role in everyday language use as well as in scientific reasoning, e.g., to describe conditions under which a fact is acceptable. There are many approaches to conditionals, but most lead to unsatisfactory results or have

[1] This chapter was submitted as an article for publication in Synthese, T.C.: Beyond Inferentialism (Pfister, 2024).

theoretical shortcomings. For example, the material implication fits well in first-order logic, but does not reflect how conditionals are used in everyday and scientific discussions (Skovgaard-Olsen et al., 2016, p. 27). As a result, a larger number of different approaches to conditionals have been developed, among which the suppositional account has become popular (cf. Evans and Over, 2004; Kaufmann et al., 2023). In addition, a larger number of relevance approaches are in development, which have been increasingly discussed lately (cf. Rott, 2025).
One of the most important differences between the suppositional account and relevance approaches concerns the connection between the antecedent and the consequent. As an example, consider the following two conditionals:

(1) If the sun shines, the solar farm produces a large amount of electricity.

(2) If food prices are high, the solar farm produces a large amount of electricity.

While (1) seems intuitively acceptable, (2) sounds odd according to proponents of relevance approaches. The reason is that there is no known relationship between the antecedent and the consequent of (2); hence, the acceptability of the consequent seems to be independent of the acceptability of the antecedent. However, in case both the antecedent and the consequent are acceptable, suppositional approaches consider not only (1), but also (2) to be acceptable. According to suppositional approaches, the strangeness of unconnected conditionals such as (2) is explained by pragmatic circumstances, e.g., by a violation of conversational implicatures (Over and Cruz, 2023). In contrast, relevance approaches regard unconnectedness in conditionals not only as a pragmatic issue but also as a genuine defect (Skovgaard-Olsen, 2016, pp. 563-570; Douven et al., 2023, sect. 1; Skovgaard-Olsen, 2020, pp. 201-203). Therefore, they consider a conditional acceptable only in case there is a connection between the antecedent and the consequent. Apart from this joint basis, relevance approaches differ widely in their details. For example, they define the connection between the antecedent and the consequent in different ways, e.g., statistically, inferentially, or causally. This can lead to divergent outcomes where a conditional is considered acceptable by one approach but not by another.
The aim of the chapter is not to advocate a particular approach to conditionals but to examine how a particular type of conditional is evaluated by different approaches: Conditionals whose antecedent and consequent have a relevance connection, but where the acceptability of the antecedent has no influence on the acceptability of the consequent. That is, learning whether the antecedent is accepted or not does not change the acceptance of the consequent. This happens, for example, in the case of overdetermination, where the consequent is implied not only by the antecedent in question but also by another antecedent. The chapter analyses and compares various approaches to conditionals and evaluates whether some approaches can cover these cases better than others. It is hoped that this allows one to identify approaches

that are more promising than others and whose further development may allow for a comprehensive and generally accepted theory of conditionals.
Unless otherwise stated, considerations are limited to standard conditionals[2] that are in indicative mood and that are simple, i.e., whose antecedent and consequent are not themselves conditionals. The various relevance approaches differ in whether they rely on truth, belief, probability, assertibility, or acceptability of conditionals. Insofar as conditionals are discussed in general or several approaches are dealt with at once, the term 'acceptability' is used to refer to the specific interpretations of the different approaches.
The chapter is structured as follows: Section 5.2 offers an overview of recently and widely discussed approaches to conditionals. Section 5.3 provides an analysis of the various approaches on conditionals whose consequents are implied by several mutually exclusive and exhaustive antecedents. Section 5.4 presents an analysis of the various approaches on conditionals whose consequents are implied by several non-exclusive antecedents. Section 5.5 discusses how the conditionals from sections 5.3 and 5.4 are ideally evaluated and compares this with the actual results. Section 5.6 examines the most promising approaches to conditionals in this respect in more detail for their general applicability.

5.2 Overview of Approaches to Conditionals

This section provides an overview of various approaches to conditionals, in particular of the suppositional account and of recent and widely discussed relevance approaches. The aim is not to provide a complete description of each approach, but to present their core aspects that are relevant for the evaluation of the conditionals discussed in the following sections.

5.2.1 Suppositional Account

The suppositional account has many different interpretations, but all are based on the Ramsey test (cf. Over and Cruz, 2017, pp. 438-442). The Ramsey test allows one to determine the acceptability of a conditional by hypothetically assuming the antecedent to be true: The antecedent is added to one's stock of beliefs, and when necessary, minimal changes are made to maintain consistency. Based on this, the

[2]Standard conditionals express some kind of conditional relation between the antecedent and the consequent (e.g. "(Only) if the phone rings, I answer it."). In contrast, non-standard conditionals rely on the same syntactic structure of "If ... then ...", but are homonymous in that they do not express a conditional relation, but something else. Examples are so-called biscuit conditionals (e.g. "If you're hungry, there are biscuits on the table."), even-if-conditionals (e.g. "(Even) if we leave now, we will be late.") and Dutchman conditionals (e.g. "If Harry passes the exam, I'm a Dutchman."). This chapter does not take a position on how non-standard conditionals should be interpreted; they are outside the scope of the inquiry (cf. Douven et al., 2023, pp. 206-209).

acceptability of the consequent is evaluated, and in case the consequent is accepted, the conditional is also accepted; otherwise, it is not. Probabilistic interpretations of the suppositional account generally follow the conditional probability hypothesis (cf. Over and Cruz, 2017, p. 439):[3]

$$P(A \rightarrow C) = P(C \mid A) \tag{CPH}$$

As mentioned in Section 5.1, suppositional approaches do not require any relevance connection between the antecedent and the consequent to consider a conditional acceptable, which distinguishes them from relevance approaches.

5.2.2 Douven, Elqayam and Krzyżanowska: Inferentialism

Douven et al. (2023) develop an approach to relevance conditionals called inferentialism. Building on the core idea that unconnected conditionals are genuinely defective, a conditional is required to obtain an inferential connection between the antecedent and the consequent (Douven et al., 2023, pp. 188f). In contrast to many other approaches, the inferential connection does not have to be of a specific type, such as necessarily deductive or causal, but can be of various types: It can be not only deductive, but also inductive or abductive, whereby abductive is understood in the sense that the consequent serves as an explanation for the antecedent.[4] In addition, it can be logical, statistical, causal, explanatory, metaphysical, epistemic, analogical, or a second-order functional property (Douven et al., 2023, pp. 188-190). A conditional is considered true in case there is a compelling argument from the antecedent and some contextually determined background knowledge to the consequent, where the antecedent is pivotal for this argument (i.e., without the antecedent the argument would not be compelling) (Douven et al., 2023, p. 190). In case there is a compelling argument from the antecedent and some contextually determined background knowledge to the negation of the consequent, the conditional is considered false; and in case there is no compelling argument, the conditional is considered indeterminate.

5.2.3 Rott: Difference-making Conditionals

Rott (2022a) introduces a non-probabilistic approach to relevance conditionals, called difference-making conditionals, which is based on belief-revision semantics. A conditional is accepted in case two conditions are fulfilled, which is called the Relevant Ramsey Test: First, the consequent is accepted in case the agent's belief

[3] In this chapter, the annotations in all formulae and citations are unified, with A for the antecedent and C for the consequent.

[4] Abductive conditionals are also often called diagnostic or evidential conditionals. Abductive conditionals must not be confused with conditionals inferred by an abductive inference (cf. sect. 4.2.3).

state is revised by the antecedent; and second, the consequent fails to be accepted in case the agent's belief state is revised by the antecedent's negation (Rott, 2022a, pp. 133, 139).[5]

Although Rott (2022a, p. 139) conceives the relevance connection not as a conjunction of two object-language sentences such as $(A > C) \wedge \neg(\neg A > C)$[6] but as an intrinsically contrastive connective, it does not have to be defined in terms of belief-revision semantics. Instead, it can also be used in standard conditional logics such as System P (cf. Rott, 2025, p. 152) to determine the truth, acceptability, or assertability of conditionals (Rott, 2022a, p. 152).

5.2.4 Crupi and Iacona: Evidential Interpretation

Crupi and Iacona advocate an account called evidential interpretation. It is based on Chrysippus' idea that a conditional holds whenever the denial of its consequent is incompatible with its antecedent: In case the antecedent is true, the consequent cannot easily be false; and in case the consequent is false, the antecedent cannot easily be true (Crupi and Iacona, 2022a, pp. 2900f). This idea can be spelt out in a modal approach (Crupi and Iacona, 2022a; Raidl et al., 2022) and in a probabilistic approach (Crupi and Iacona, 2022b; Crupi and Iacona, 2021).[7]

In the modal approach, a conditional is considered true in case two requirements are fulfilled: (i) in the closest world in which the antecedent is true, the consequent must not be false, and (ii) in the closest world in which the consequent is false, the antecedent must not be true. While the first requirement expresses the commonly known Ramsey test, the second requirement is intended to capture the idea that the consequent holds in virtue of the antecedent (Crupi and Iacona, 2023, p. 121). In case an antecedent is always false or a consequent is always true, the conditional is considered true (Crupi and Iacona, 2022a, p. 2902).

In the probabilistic approach, the acceptability of a conditional $A \rightarrow C$ is equal to the degree of incompatibility $A \uparrow C$ between the antecedent and the negation of the consequent (Crupi and Iacona, 2023, p. 122):

$$A \uparrow C = 1 - \frac{P(A \wedge \neg C)}{P(A) * P(\neg C)} \tag{DI}$$

in case that $P(A \wedge \neg C) \leq P(A) * P(\neg C)$. In the case of $P(A) = 0$ or $P(C) = 1$,

[5] Rott (2022a, pp. 133, 149) also proposes a slightly different alternative, called the Dependent Ramsey Test. It differs from the Relevant Ramsey Test by the second condition, which requires that the consequent is rejected (i.e. its negation is accepted) in case the belief state is revised by the negation of the antecedent.

[6] The character '>' has the meaning 'If A, then plainly C.' (Rott, 2022a, p. 139).

[7] Rott (2022b, pp. 13f) shows that both approaches do not result in the same logic and that only the modal, but not the probabilistic, approach validates disjunctive rationality ($((A_1 \vee A_2 \rightarrow C) \wedge (\neg A_1 \rightarrow C) \vdash (\neg A_2 \rightarrow C))$. However, these differences are not important for the analyses in this chapter.

the degree of incompatibility is 1, and in all other cases, it is 0.

5.2.5 Skovgaard-Olsen: Statistical Relevance

Skovgaard-Olsen (2020, p. 206) emphasises the role of conditionals as arguments in reasoning and therefore considers unconnected conditionals as semantically defective (Skovgaard-Olsen, 2020, pp. 201-203). The relevance of conditionals can be measured by the measure of difference:

$$\Delta P = P(C \mid A) - P(C \mid \neg A) \tag{MD}$$

whereby $\Delta P > 0$ indicates positive relevance, $\Delta P < 0$ negative relevance, and $\Delta P = 0$ irrelevance (Skovgaard-Olsen et al., 2016, pp. 27f).
Empirically, the evaluation of conditionals can be described by the default and penalty hypothesis: By default, people assume that the antecedent and consequent are positively connected and therefore directly evaluate the acceptability of a conditional by $Acc(A \rightarrow C) = P(C \mid A)$ (Skovgaard-Olsen et al., 2016, p. 28). However, once the assumption of a positive connection is refuted, $Acc(A \rightarrow C)$ is considered to be 0. Besides theoretical considerations on the question of whether $P(C \mid A)$ should be a measure of the probability or the acceptability of a conditional (cf. Skovgaard-Olsen, 2016, p. 558), there are also mixed empirical results. For example, the evaluation of P(if A, then C) and Acc(if A, then C) may differ depending on the type of inferential relation of the conditional, as a comparison with the results of Douven and Verbrugge (2010) indicates (Skovgaard-Olsen et al., 2016, p. 34). In addition, experiments show a clear dissociation in the evaluation of truth, probability, and acceptability (Skovgaard-Olsen et al., 2017, p. 474).[8]

5.2.6 Van Rooij and Schulz: Causal Relative Difference

van Rooij and Schulz (2019, pp. 58f) argue that the assertibility of a conditional can be determined by the measure of relative difference: A conditional is assertible iff

$$\Delta^* P_A^C = \frac{P(C \mid A) - P(C)}{P(\neg A \wedge \neg C)} \tag{MRD}$$

is high. Alternatively, it is suggested that $\Delta^* P_A^C$ does not need to be high but that $\Delta^* P_A^C >> \Delta^* P_a^C$, whereby a stands for all (or the disjunction of all) relevant alternative antecedents (van Rooij and Schulz, 2019, p. 59). Compared to Skovgaard Olsen's measure of difference ΔP, the measure of relative difference $\Delta^* P_A^C$ allows for the consideration of two additional intuitions: First, with increasing $P(C \mid \neg A)$

[8] However, there are also contradictory empirical results, see Douven et al. (2023, p. 189).

the required difference between $P(C \mid A)$ and $P(C \mid \neg A)$ decreases. Second, the value $P(C \mid A)$ is more important than the value of $P(C \mid \neg A)$.
The measure of relative difference represents an asymmetric correlation that is due to a causal relationship between the antecedent and the consequent. This understanding allows for the evaluation of the assertibility of conditionals that express a causal relationship, such as

(3) If it rains, the street is wet.

It also permits the evaluation of diagnostic conditionals (van Rooij and Schulz, 2019, pp. 65-69). In such, one infers from the assertability of a cause to the assertability of its effect, e.g., as in

(4) If the street is wet, it rains.

Furthermore, van Rooij and Schulz (2019, p. 69) consider conditionals to be assertible in case both the antecedent and the consequent are caused by a common cause. An example is the conditional

(5) If the barometer falls, there is a storm.

where both propositions are caused by low air pressure. In addition, conditionals are considered assertible in case the antecedent and the consequent have a deductive or semantic relationship or can be metaphysically grounded.

5.2.7 Günther: Causality

Günther (2022) proposes a conditional approach based on causal models, allowing for both causal and evidential conditionals. Conditionals are believed by an agent to be true in case they are true in the most plausible world(s). A world is the more plausible the more it corresponds to the agent's beliefs about which facts are true and, subordinately, the more the world corresponds to the agent's causal beliefs (Günther, 2022, p. 616).[9]
While causal conditionals represent causal relations in which the antecedent causes the consequent, evidential[10] conditionals represent causal relations in which the antecedent is caused by the consequent, such as e.g. in (4). In addition, the representation allows the evaluation of backtracking conditionals, where the non-occurrence of an effect indicates that some of its causes are not present (Günther, 2022, p. 622). For example, in case Tom is seen leaving an interview dissatisfied, one can conclude

[9]The account does not require absolute certainty, but only relative certainty, i.e., the agent only has to be "most certain" about the state of a fact. This is the case when the agent is at least quite certain about the state of the fact and is not more certain about any other state of the fact (Günther, 2022, p. 624).
[10]Evidential conditionals are often also called diagnostic or abductive conditionals.

(6) If Tom had left the interview smiling, the interview would have gone well.

In contrast, conditionals whose antecedent and consequent are based on a common cause, such as (5), are not considered true.

5.2.8 Berto and Özgün: Topicality

Berto and Özgün (2021, p. 3708) present an approach to relevance conditionals in which conditionals are considered acceptable in case the antecedent and the consequent are about the same topic. More precisely, the topic of the consequent has to be fully included in the topic contextually determined by its antecedent. The consequent can either be about the same topic as the antecedent or of a topic of some relevant background assumptions, which are determined by the antecedent and the context. For example,

(7) If we keep burning fossil fuel at this pace, the polar ice will melt.

is considered an acceptable conditional. Even though the antecedent and the consequent do not share the same topic, they are connected by topics of background assumptions, such as "emission of CO_2" and "raising global temperature". "The criterion of relevance [...] aims at giving a catch-all condition, covering relevance of any kind, whether inferential or not" (Berto and Özgün, 2021, p. 3702). In case the antecedent and the consequent are topically connected, the acceptability of a conditional is equal to the conditional probability $P(C \mid A)$. In case they are not topically connected, the acceptability of the conditional is 0.

5.3 Evaluation of Conditionals with Several Mutually Exclusive Antecedents

In this section, conditionals are to be analysed whose consequent is implied not only by one but by several antecedents. Moreover, the antecedents are together exhaustive, i.e., no other antecedent implies the consequent. While the next section examines cases where the antecedents are non-exclusive, this section considers mutually exclusive antecedents. The simplest case of mutually exclusive antecedents $A_1 ... A_n$ occurs when both a fact A and its negation $\neg A$ imply a consequent C. As an example, consider a case in which Alice expresses

(8) If the weather will be good on the weekend, I will go to the mountains.

Alice states the conditional because she likes to hike and plans to hike in the mountains with Bob on the weekend. Since Alice normally does not go to the mountains, the conditional is considered acceptable by all approaches to conditionals presented

in the previous section. This is because the antecedent and the consequent are causally connected, and the consequent is only acceptable in case the antecedent is accepted.
Now suppose Alice is also looking for plans in case the weather will be bad on the weekend. Carol suggests that they go to a spa in the mountains, since the spa is unusually empty on bad weather days. Alice agrees and therefore expresses

(9) If the weather will be not good on the weekend, I will go to the mountains.

In case Alice only states (9) but not (8), (9) is also considered acceptable by all approaches to conditionals mentioned in the previous section. However, in case both conditionals are stated together, the evaluation of the conditionals differs among the various approaches, as shown next. For simplicity, the two conditionals are expressed with conditional variables, whereby A stands for "the weather will be good on the weekend" and C for "I will go to the mountains".

(8') $A \rightarrow C$

(9') $\neg A \rightarrow C$

At first sight, this constellation seems similar to an example from Stalnaker (1968, p. 42f), which is about the evaluation of the conditional

(10) If the Chinese enter the Vietnam conflict, the U.S. will use nuclear weapons.

Stalnaker argues, in case one believes that the use of nuclear weapons by the U.S. is inevitable, e.g., due to arrogance of power or domestic causes, one believes

(11) If the Chinese enter the Vietnam conflict, the U.S. will use nuclear weapons, and if the Chinese do not enter the Vietnam conflict, the U.S. will use nuclear weapons.

This belief seems to be very similar to believing (8) and (9) together. However, there is an essential difference: While in (8) and (9) both A and $\neg A$ imply the consequent, in (11) neither A nor $\neg A$ imply the consequent, but it is implied by another fact.[11] In the following, it is examined how the individual approaches evaluate the two conditionals (8) and (9) when Alice expresses both together; i.e., in case the weather will be good, she will go to the mountains to hike, and in case the weather will be bad, she will go to the mountains to visit the spa.

[11] Stalnaker (1968, 43) uses the example to argue against approaches that require some sort of logical or causal connection between the antecedent and the consequent. He claims that the example refutes such approaches, because in case the use of nuclear weapons is inevitable, one considers (10) "[c]learly [...] to be true" despite the absence of a connection. However, as shown above, there are reasons not to consider conditionals like (10) to be clearly true, since they have no relevance connection.

The suppositional account evaluates the probability of a conditional based on the formula $P(A \rightarrow C) = P(C \mid A)$. Since the consequent is certain for the occurrence of each antecedent, both (8) and (9) have a conditional probability of $P = 1$ and are therefore considered acceptable.
Douven, Krzyżanowska and Elqayam's inferentialism requires an inferential connection between the antecedent and the consequent. Such a connection is present in both conditionals, since both are based on strong causal relations. Consequently, both conditionals are evaluated as true.
Rott's approach to difference-making conditionals accepts a conditional in case the following two requirements are met: In case the antecedent is accepted, the consequent is accepted, and in case the negation of the antecedent is accepted, the consequent is not accepted. Thus, to accept (8'), it must be true that $A \rightarrow C$ and that $\neg A \rightarrow \neg C$; whereas to accept (9'), it must be true that $\neg A \rightarrow C$ and that $A \rightarrow \neg C$. Since these two sets of statements contradict each other, the acceptance of (8') and (9') together has to be negated. Moreover, Rott (2022a, pp. 145-148) considers Aristotle's second thesis (AST) to be valid:

$$\neg((A \rightarrow C) \wedge (\neg A \rightarrow C)) \tag{AST}$$

AST allows one to conclude from the truth of (8') that (9') is false, and likewise from the truth of (9') that (8') is false. Consequently, it is not possible for (8) and (9) to be considered true at the same time, which also speaks for their non-acceptance. In general, AST seems intuitively appealing, as an example from Crupi and Iacona (2023, p. 122) illustrates: "If the presence of white smoke is a reason for believing that a new pope has been elected, it is hard to see how the absence of white smoke can also be a reason for believing that a new pope has been elected." In this example, however, AST is convincing because the example expresses a case in which the consequent has only one relevance connection. But, as shown above, there are also cases in which the consequent has not only one but several relevance connections, i.e., it can be implied in several ways. Therefore, it seems that AST cannot be accepted as a generally valid rule.
Crupi and Iacona's evidential interpretation requires that the consequent cannot easily be false in case the antecedent is true, and that the antecedent cannot easily be true in case the consequent is false. Both conditions are fulfilled for (8) and (9) and therefore both are considered true. This is also underlined by the statement that conditionals are true in case the consequent is necessary, which is here the case (Crupi and Iacona, 2022a, p. 2913). Similar to Rott, Crupi and Iacona (2022a, p. 2913) consider AST appealing, but prefer a restricted version called Restricted Aristotle's Second Thesis (RAST):

$$\Diamond\neg C \models \neg((A \rightarrow C) \wedge (\neg A \rightarrow C)) \tag{RAST}$$

RAST differs from AST in that an additional requirement must be fulfilled: Only in case the consequent is not necessarily true, it cannot be true that both an antecedent and the negation of the antecedent imply the same consequent. Since in the case of (8) and (9) the consequent is necessarily true – as the antecedents are exhaustive – RAST, unlike AST, does not apply and thus plays no role in their evaluation.

Skovgaard-Olsen's statistical relevance approach considers conditionals to be acceptable in case $\Delta P = P(C \mid A) - P(C \mid \neg A)$ is positive. This allows for two different cases: In the first case, both conditionals have the same probability[12] of the consequent being acceptable in case the antecedent is accepted. Then, $P(C \mid A)$ and $P(C \mid \neg A)$ have the same value, which leads to both $\Delta P = 0$. Consequently, both conditionals are considered irrelevant and thus unacceptable. In the second case, both antecedents have different probabilities[13] of the consequent being acceptable in case the antecedent is accepted. In that case, ΔP evaluates the more probable conditional as positively relevant and therefore acceptable. The less probable conditional is evaluated by ΔP as negatively relevant and therefore unacceptable.

Van Rooij and Schulz's approach to causal relative difference evaluates a conditional assertible in case it satisfies the measure of relative difference $\Delta^* P_A^C$. Although it defines the measure differently from Skovgaard-Olsen's ΔP, the result is the same: In case both (8') and (9') are given and have the same probability, both $\Delta^* P_A^C = 0$ and they are considered not assertible. In case both conditionals have different probabilities, the more probable one has a positive $\Delta^* P_A^C$ value and is considered assertible, whereas the less probable one has a negative $\Delta^* P_A^C$ value and is considered not assertible. van Rooij and Schulz (2019, pp. 60-63) consider $\Delta^* P_A^C$ to be an accurate indicator of a causal relationship between the antecedent and the consequent. In this example case, however, this is not true, neither for both conditionals in case they have the same probability, nor for the less probable conditional in case they have different probabilities: Even though $\Delta^* P_A^C$ being not high indicates that there is no causal relation, there is one between the antecedent and the consequent in both conditionals.

Günther's causality approach considers conditionals to be believed as true in case they correspond most to the facts and the causal model believed by an agent. In the case of (8) and (9), both conditionals correspond to the facts, and in both the antecedent is a causal reason for the consequent. That the antecedents of the two conditionals are contradictory is not a problem with respect to the requirement that the most plausible world needs to correspond with the agent's belief about which facts are true. This, because the agent has no belief about which of the two mutually exclusive antecedents is true, i.e., what the weather will be like on the weekend. Thus, according to Günther's approach, the two conditionals together are

[12] Respectively the difference between the two probabilities is smaller than some significance factor ϵ.

[13] Respectively the difference between the two probabilities is larger than some significance factor ϵ.

believed to be true.
Berto and Özgün's topicality approach requires that the antecedent and the consequent are about the same topic or are topically connected by some background assumptions. Although the requirement is imprecise, it can be assumed that it is fulfilled for both (8) and (9) – in both cases, the antecedent and the consequent are connected by some background knowledge of Alice wanting to enjoy activities with her friends. Consequently, both conditionals are considered acceptable.
Overall, it becomes apparent that the various approaches evaluate conditionals whose consequent is fulfilled by several mutually exclusive and exhaustive antecedents differently. While five approaches consider them acceptable, three do not.

5.4 Evaluation of Conditionals with Several Non-exclusive Antecedents

As the previous section, this section analyses the evaluation of conditionals by the approaches presented in Section 5.2. The consequent of the conditionals is again implied not by only one, but by several antecedents. Unlike in the previous section, however, the antecedents are not mutually exclusive but non-exclusive, i.e., several of them can occur simultaneously. Consequently, they need not be exhaustive and there may be other, unknown antecedents to the same consequent. Consider the following example (cf. sect. 4.2.3): David has a neighbour who often throws parties that are so loud that David feels disturbed at night. More specifically, David cannot sleep well on four out of five nights in which the neighbour has a party. Therefore, David states

(12) If my neighbour throws a party, I cannot sleep well at night.

As such, the conditional is rated acceptable by all approaches to conditionals presented in Section 5.2: The antecedent and the consequent are causally related, and the consequent is only acceptable in case the antecedent is accepted.
Suppose David next learns that a new bar is moving in directly below his flat. He also learns that the bar will play very loud music and that the sound insulation of the house is very poor. Therefore, he states

(13) If the bar under my flat is open, I cannot sleep well at night.

and he is certain of it. In case (13) is to be evaluated without (12), it is considered acceptable by all approaches to conditionals, as it meets all requirements. For simplicity, the two conditionals are expressed with conditional variables, whereby A_1 stands for "my neighbour throws a party", A_2 for "the bar under my flat is open", and C for "I cannot sleep well at night".

(12') $A_1 \rightarrow C$

(13') $A_2 \rightarrow C$

In the following, it is examined how the two conditionals are evaluated in case both A_1 and A_2 are given, as well as their relevance connections to the consequent C.
The suppositional account evaluates (12) and (13) by $P(A \rightarrow C) = P(C \mid A)$. Since the consequent is certainly fulfilled by A_2 (and in four out of five cases additionally by A_1), the consequent is certain, i.e., $P(C) = 1$. Thus, both (12) and (13) are assigned $P = 1$ as well and are considered acceptable.
Douven, Elqayam and Krzyżanowska's inferentialism evaluates both conditionals as true, since in both conditionals there exists an inferential connection between the antecedent and the consequent. Inferentialism determines the probability of a conditional by the inference heuristic: the probability that a conditional is true is "the likelihood that we can make a compelling case for the consequent, starting from the antecedent plus background knowledge" (Douven et al., 2023, p. 200). This heuristic is shown to be empirically more accurate than the thesis of the suppositional account that probability ratings express conditional probability ratings, i.e., that $(A \rightarrow C) = P(C \mid A)$ (Douven et al., 2022b). Based on the inference heuristic, (12) is assigned a probability of $P = 0.8$, since four times out of five David does not sleep well at night when his neighbour throws a party. (13) is assigned a probability of $P = 1$ because it is certain that David cannot sleep well in case the bar is open.
Rott's approach to difference-making conditionals accepts a conditional in case two conditions are met: First, the consequent is accepted in case the agent's belief state is revised by the antecedent; and second, the consequent fails to be accepted in case the agent's belief state is revised by the antecedent's negation. For (12), the first, but not the second, condition is satisfied: The consequent is accepted due to its implication by A_2, regardless of whether the antecedent is believed to be true or false. Consequently, (12) is not considered acceptable. For (13), the first condition is always fulfilled and the second in the case that A_1 does not imply C, which occurs 20 % of the time. Since Rott offers a purely qualitative framework and does not propose any probabilistic version, a probabilistic interpretation can only be based on own assumptions. In case one follows the simplest interpretation – the acceptability of a conditional is equal to the probability that both conditions are fulfilled – then the acceptability of (13) would be 0.2.[14] Rott (2022b, p. 17) explicitly discusses a case where two different antecedents both imply the same consequent. In case only one of the antecedents is fulfilled, the corresponding conditional is considered acceptable, since the antecedent makes a difference to the outcome. In case both antecedents are fulfilled, each alone makes no difference. However, Rott considers the corresponding conditionals to be "rather unassertable than unacceptable". It is

[14] Alternatively, for example, one could consider a conditional acceptable to the degree of $P(C \mid A)$ in case both conditions are satisfied, which would lead to an acceptability of 1.

not entirely clear how this assessment relates to the above evaluation results, but since unassertability is relatively closer to unacceptability than to acceptability, the results seem to be confirmed.

Crupi and Iacona's evidential interpretation offers not only a modal but also a probabilistic version (cf. sect. 5.2.4). The acceptability of (12) is determined by the degree of incompatibility (DI), since $P(A \wedge \neg C) \leq P(A) * P(\neg C)$, which leads to $Acc(12) = 1$. For (13), $P(C) = 1$ and therefore $Acc(13) = 1$.

Skovgaard-Olsen's statistical relevance approach evaluates the acceptability of conditionals by default by $Acc(A \rightarrow C) = P(C \mid A)$. Since the consequent is always fulfilled by A_2, both $P(C \mid A_1)$ and $P(C \mid A_2)$ are 1. Therefore, by default, $Acc(12) = 1$ and $Acc(13) = 1$. However, conditionals are only considered acceptable in case they also have a positive ΔP value, which is measured by the measure of difference (MD). Since the consequent is always fulfilled by A_2 but only in four out of five cases by A_1, $\Delta P(12) = 0$ and $\Delta P(13) = 0.2$. Hence, only (13) but not (12) is considered acceptable since only A_2 but not A_1 increases the probability of the consequent being true.

Van Rooij and Schulz's approach to causal relative difference evaluates the assertibility of a conditional by the measure of relative difference MRD. Although the approach relies on probabilities, van Rooij and Schulz (2019, pp. 58, 63) state that the assertibility of a conditional itself cannot be indicated by degree: A conditional is either assertible – iff $\Delta^* P_A^C$ is high – or not assertible. Independent of that, in case both A_1 and A_2 are taken to be true, the measure of relative difference leads to an invalid result, since one would need to divide by 0; an alternative method of calculation is not given for such cases. Nevertheless, van Rooij and Schulz discuss the occurrence of alternative causes, concluding for pragmatic reasons that alternative causes are complete causal explanations for the consequent and are therefore considered incompatible with each other. Based on these findings and considering that the main idea of the approach is that conditionals must be causally relevant, at least (12), and arguably also (13), is considered non-assertible.

Günther's causality approach does not provide a probabilistic interpretation. Nevertheless, it offers some indications on how an evaluation could be made. In general, a conditional is believed to be true in case it is true in the most plausible world(s). Taking A_1 and A_2 as given, the most plausible world is the one in which both the antecedents and the consequent are true. Whether (12) and (13) are believed thus depends on whether their causal relationships are believed. Their belief can be affirmed not only because the consequent could otherwise not be true, but also because both conditionals are based on a strong causal connection. Due to the certain causal relationship in (13), it appears appropriate to set $Bel(13) = 1$. For (12), where the causal link is less strong and the antecedent implies the consequent only in four out of five cases, it seems appropriate to assign $Bel(12) = 0.8$.

Berto and Özgün's topicality approach considers a conditional acceptable to the

degree of the conditional probability $P(C \mid A)$ in case the antecedent and the consequent are topically connected; otherwise, the conditional is unacceptable. The requirement of being topically connected is fulfilled by (12) as well as by (13). Since the consequent is always fulfilled by at least A_2, $Acc(12) = 1$ and $Acc(13) = 1$.
The summary in Table 5.1 shows that the approaches evaluate conditionals whose consequent is fulfilled by several non-exclusive antecedents quite diversely.

Table 5.1: Evaluation of conditionals in the case of the consequent being implied by several non-exclusive antecedents

Approach	(12') $A_1 \rightarrow C$	(13') $A_2 \rightarrow C$
Suppositional account	1	1
Douven et al.: inferentialism	0.8	1
Rott: difference-making conditionals	0	* 0.2
Crupi & Iacona: evidential interpretation	1	1
Skovgaard-Olsen: statistical relevance	0	1
Van Rooij & Schulz: causal relative difference	0	0
Günther: causality	* 0.8	1
Berto & Özgün: topicality	1	1

* the value is based on an own interpretation, since the approach itself does not provide a probabilistic interpretation.

5.5 Interpretation of the Evaluation Results

In the last two sections, it was shown that the suppositional account and the discussed relevance approaches evaluate certain types of conditionals quite differently. While Section 5.3 concerns conditionals whose consequent is implied by several mutually exclusive and exhaustive antecedents, Section 5.4 concerns conditionals whose consequent is implied by several non-exclusive antecedents. Both types of conditionals share one important aspect: the antecedent and the consequent of the conditionals have a relevance connection, but the acceptability of the antecedent has no influence on the acceptability of the consequent. However, the two types of conditionals differ on the reason for the absence of the influence: In the case of mutually exclusive antecedents, the consequent is implied either way, whereas, in the case of non-exclusive conditionals, the consequent is implied anyway. More precisely, in the first case, the consequent C is implied not only by the antecedent A_1 but also by other antecedents $A_2 \ldots A_n$ that are mutually exclusive, exhaustive together with A_1, and have all the same or a higher probability of implying the consequent C as A_1.[15] In the second case, the consequent C is implied not only by the antecedent A_1 but also by other antecedents $A_2 \ldots A_n$ that are non-exclusive and whose combined

[15] In the case of a non-probabilistic interpretation, all conditionals $(A_{2\ldots n} \rightarrow C)$ are considered acceptable.

probability of implying the consequent C is 1[16].[17] It could be argued that both cases are purely theoretical without practical relevance and therefore do not need to be covered by approaches to conditionals. However, not only are the above cases realistic – both Alice's and David's situations can occur in everyday life – but also the following examples show that such cases are common and therefore approaches to conditionals must be able to handle them.
In the case of mutually exclusive antecedents, imagine a discussion (in mid-2024) about the war between Russia and Ukraine in which the following two statements are uttered:

(14) If Russia loses the Russia-Ukraine war, there will be a new Cold War.

(15) If Russia wins the Russia-Ukraine war, there will be a new Cold War.

Both conditionals can be well justified: For instance, it can be reasoned that in case Russia loses the war, a new nationalistic Russian government is likely to come to power and increase its hostility towards Western countries; and in case Russia wins the war, Western countries will tighten their sanctions and try to isolate Russia to prevent it from invading another country. Both conditionals can be stated separately, but also together – both scenarios seem possible and plausible and as such acceptable. This applies regardless of how likely one considers each of the two antecedents to occur. Even in case one considers it much more likely that Russia will lose than win the war, or conversely, both conditionals themselves remain plausible. In case another scenario with a different outcome is also conceivable, e.g.,

(16) If Russia and Ukraine sign a peace treaty, there will be no new Cold War.

and it is assigned a probability which is greater than 0, (14) and (15) are considered acceptable by most relevance approaches, and the contradictory evaluation results above would not occur.[18] However, at least at the time of writing in mid-2024, a peace treaty seems very unlikely, and the crucial point is not whether there could be other war outcomes in this particular case, but that there are realistic situations in which all potential scenarios are equally likely to imply the same outcome.
Similarly, there are many situations in which occurs the case of several non-exclusive antecedents that all imply the same consequent. For example, one buys a plant in a nursery, whereupon the gardener, based on his experience that many customers fulfil one or both of the antecedents, says

(17) If the plant is placed in direct sun, it will die.

[16]Respectively larger than 1 minus some significance factor ϵ.
[17]In the case of a non-probabilistic interpretation, there is at least one conditional $(A_{2\ldots n} \rightarrow C)$ considered acceptable.
[18]Except that AST remains a problem, since both A and $\neg A$ from (14) and (15) still can lead to the same consequent.

(18) If the plant is not watered regularly, it will die.

Again, it seems to be an everyday situation, and it seems appropriate to accept each conditional separately as well as both together.
Conditionals, which have a relevance connection, but where the acceptability of the antecedent has no influence on the acceptability of the consequent, can also not be expressed as concessive conditionals, i.e., as "even if" conditionals. As an example, for Alice's case, consider conditionals (8) and (9) in their concessive form

(19) Even if the weather will be good on the weekend, I will go to the mountains.

(20) Even if the weather will not be good on the weekend, I will go to the mountains.

Although both conditionals can be acceptable in certain circumstances, in Alice's situation, they do not express the underlying reasons: Alice will not go to the mountains although the weather will be good (or bad), but because the weather will be good (or bad). In both cases, each conditional is based on a positive relevance connection in which the antecedent provides a reason for the consequent. Consequently, expressing such cases through concessive conditionals is not a solution. The examples in this section already indicate that the two types of conditionals in question are not only common but also seem acceptable. This is because the conditionals fulfil the basic idea of relevance approaches: A conditional is considered acceptable in case there is a supportive relevance connection between the antecedent and the consequent. In the following, additional deliberations are made to determine whether such conditionals should be considered acceptable – as some of the approaches to conditionals claim – or unacceptable – as some other of the approaches claim.
Among the approaches that consider such conditionals unacceptable are those that use statistical measures such as Skovgaard-Olsen's measure of difference MD ΔP and Van Rooij and Schulz's measure of relative difference MRD $\Delta^* P_A^C$.[19] Both approaches are based on the idea that a relevance connection implies positive statistical relevance. However, as shown above, this is not true for the types of conditionals discussed in this chapter, raising the question of which of the two aspects is more important. Although Skovgaard-Olsen does not explicitly address their relation, statistical relevance seems to be a means to measure the more fundamental relevance connection. For example, Skovgaard-Olsen (2020, pp. 201-203) argues that the relevance connection of conditionals plays a central role in argumentation and reasoning and makes it possible, for instance, to express arguments. Similarly, van Rooij and Schulz (2022, p. 366) argue that the semantic analysis of a conditional suffices and that the relevance measure turns out to be a pragmatic and cancellable implicature.
Consequently, statistical measures can be considered as a helpful but not completely reliable indicator of the existence of a relevance connection: A positive value of ΔP

[19] An analysis of various measures of evidential support can be found in Rott (2025, pp. 171-187).

or $\Delta^* P_A^C$ can be a sufficient, but not necessary indicator of a relevance connection.[20] On this basis, it seems that relevance approaches that rely on statistical measures consider the two types of conditionals in question to be unacceptable not because the conditionals are genuinely unacceptable, but because the statistical measure is incapable of correctly capturing the relevance connection. As such, it seems more appropriate to consider the conditionals acceptable rather than unacceptable.

A further possibility to determine the acceptability of the conditionals is offered by coherence. Conditionals with mutually exclusive antecedents, such as (8) and (9), are individually considered acceptable because they obtain a relevance connection. A relevance connection is between the antecedent and the consequent and exists independently of other possible relevance connections. Consequently, in case a relevance connection is accepted when it is the only one present, it should also be accepted when others are present. This is especially true as, since the antecedents are mutually exclusive, only one of the relevance connections implies the consequent. Not accepting a relevance connection just because the consequent can also be implied in the absence of the antecedent by another antecedent that has the same or a higher probability of implying the consequent seems incoherent.

Similarly, incoherence occurs in the following way in case the conditionals in question, such as (14) and (15), are not accepted together: In case an additional conditional not leading to the same consequent is accepted, such as (16), (14) and (15) would be suddenly considered acceptable again by all approaches to conditionals. Yet, it is not clear why their acceptability should depend on the acceptability of an additional conditional.

Additionally, incoherence would also occur in another way, in case conditionals such as (8) and (9) are accepted alone but not both together: (8) would be acceptable for Bob, but not for Alice, and (9) would be acceptable for Carol, but again not for Alice. However, since the same relevance connection applies to Alice and Bob respectively Alice and Carol, it seems incoherent that the conditional is accepted once and once not. This applies equally to non-exclusive conditionals such as (12) and (13): Imagine David lives together with Eve. Unlike David, Eve can sleep well when music is played; hence (13) does not apply to her. However, like David, Eve feels stronlgy disturbed by voices from the neighbour's party; hence (12) does apply to her. This again would lead to an incoherence in case (12) and (13) are accepted alone but not together: Then, David considers (12) as unacceptable, whereas Eve considers it acceptable – although for both applies the same relevance connection.

Not accepting the conditionals also leads to another kind of incoherence: In case none of the conditionals gets accepted, none of them would consequently imply the consequent and hence the consequent would be considered as unacceptable.

[20] Whether a positive value is always a sufficient indicator or whether there are cases in which a conditional is to be considered as unacceptable despite a positive value has to be investigated separately and depends on additional theoretical assumptions.

However, the consequent becomes a fact and should therefore be as such accepted – for example, Alice will go to the mountains and David cannot sleep well at night. Overall, all of the considerations above indicate that conditionals whose antecedent and consequent have a relevance connection and where the acceptability of the antecedent has no influence on the acceptability of the consequent should be considered acceptable. Hence, (8) and (9), and also (12) and (13), should be considered acceptable, both individually and together. A question that arises here is how probabilistic evaluations should be, for example, in the case of (12), where the antecedent leads to the consequent in only four out of five cases. It is recommended to follow the probability of implication and assign the same probability to the conditional. This reflects how often the relevance connection actually leads to the implication of the consequent in case the antecedent is given. Hence, for example, $P(12) = 0.8$ and $P(13) = 1$.
Table 5.2 compares which of the approaches to conditionals examined in this chapter determine the correct evaluation based on these results and which do not.

Table 5.2: Evaluation of conditionals with relevance connections and for which the acceptability of an antecedent has no influence on the acceptability of its consequent

Approach	*with mutually exclusive antecedents*	*with non-exclusive antecedents*
Suppositional account	correct	incorrect
Douven et al.: inferentialism	correct	correct
Rott: difference-making conditionals	incorrect	incorrect
Crupi & Iacona: evidential interpretation	correct	incorrect
Skovgaard-Olsen: statistical relevance	incorrect	incorrect
Van Rooij & Schulz: causal relative difference	incorrect	incorrect
Günther: causality	correct	correct
Berto & Özgün: topicality	correct	incorrect

Only two approaches to conditionals, Douven, Elqayam & Krzyżanowska's inferentialism and Günther's causality approach, correctly evaluate the two types of conditionals. All other approaches fail in at least one of the cases.

5.6 Examination of Promising Approaches to Conditionals

In the following, the two approaches that lead to the correct evaluation, Douven, Elqayam & Krzyżanowska's inferentialism and Günther's causality approach, are examined in more detail to determine their general suitability for capturing the nature and evaluation of conditionals.

Douven, Elqayam & Krzyżanowska's inferentialism's main idea and conceptual outline is described in Section 5.2.2. There are several aspects that are salient and require closer examination.

First, the inferential connection can not only be deductive, inductive, or abductive, but also be logical, statistical, causal, explanatory, metaphysical, epistemic, analogical, or a second-order functional property (Douven et al., 2023, p. 191). Not only is this understanding very broad, but some of the concepts, such as abductive and explanatory connections, are not well-defined (cf. sect. 4.2.4, 4.7). Consequently, the evaluation of conditionals and especially of the argumentative strength of the connection between antecedent and consequent are difficult to assess.

Second, inferentialism, at least at present, offers no logic that can be used to evaluate conditionals. Douven et al. (2023, p. 19) point out that inferentialism is still under development and that a logic may be developed at a later stage. Moreover, it may be that the principles that people follow in regard to conditionals cannot be expressed through logic – but nevertheless, inferentialism can help to better understand the role of conditionals (Douven et al., 2023, ch. 3.1). While both arguments are convincing, a logic would still be desirable, as it would support the formalisation of conditionals, which would be beneficial for scientific reasoning and artificial intelligence. Douven et al. (2023, p. 204) argue that there are already two other relevance approaches with logics that appear promising, in particular Crupi & Iacona's evidential interpretation and Berto & Özgün's topicality approach. However, as shown above, both approaches incorrectly evaluate the two types of conditionals in question, which not only shows that they are inappropriate in this respect, but also that they are different from inferentialism.

In addition, both approaches also face other problems. For example, Crupi & Iacona provide a logic for a modal interpretation as well as a logic for a probabilistic interpretation (cf. sect. 5.2.4). Rott (2022b, p. 13) shows not only that the two logics are not identical, but also that the satisfaction of contraposition, the main idea on which the approach is built, supports the relevance connection only to a limited extent (Rott, 2022b, pp. 6-11; Rott, 2023).

Berto & Özgün require that the antecedent and the consequent are about the same topic or are connected by the topics of background assumptions (cf. sect. 5.2.8). Even though Berto and Özgün (2021, pp. 3606-3608) elaborate on the notion of topicality, it remains unclear how exactly to evaluate whether the antecedent and the consequent are topically connected or not. Based on the specifications provided, the requirement of topicality as an indicator of a relevance connection may be too permissive. For example, consider:

(21) If Alice likes sweets, Bob likes sweets.

The requirement of topicality seems to be fulfilled in the conditional – Alice and Bob are topically connected through their friendship, and in both cases, it is about

liking sweets. However, assuming that their preferences for sweets are independent of each other and did not play a role in their friendship, there does not seem to be a relevance connection in that the antecedent influences the consequent in any way. Therefore, the notion of topicality seems to be either under-defined or too permissive and is not a suitable indicator of relevant connections.
Third, the fact that inferentialism allows for inductive and statistical inference connections can be problematic. Since the concept of induction is not precisely defined, it may be too permissive and allow for assigning a relevance connection to unconnected conditionals. As an example, consider the conditional

(22) If mankind uses electricity, Antarctica is covered in snow that year.

of which both the antecedent and the consequent have been true for many years. Since there are many positive occurrences and not a single negative one, an inductive or statistical argument is well supported, and consequently, the conditional can be considered acceptable. However, there is no relevance connection between the antecedent and the consequent such that the antecedent influences the consequent in any way.[21] It is therefore questionable whether a purely inductive or statistical connection is sufficient or whether this allows for the same criticism that the suppositional account faces (cf. sect. 5.1).
Fourth, unlike most other approaches to conditionals, inferentialism does not consider the closure Modus Ponens

$$A, A \to C \vdash C \tag{MP}$$

to be valid. Douven et al. (2023, ch. 2.2) argue that MP should be invalid because in everyday practice we tend to rely much more on compelling but inconclusive, i.e., non-truth-preserving, arguments than on deductively valid ones. As an example, Douven et al. (2023, p. 189) provide the conditional

(23) If John lives in Chelsea, he is rich.

which is compelling – as most people in Chelsea are rich – but not truth-preserving – as not all people in Chelsea are rich. Since it could be that John is one of the few people who live in Chelsea but are not rich, MP must be considered invalid according to Douven et al. (2023, p. 190). However, it seems that the inconclusiveness is not due to MP but to the inductive argument on which the conditional is based. Since the inductive inference is only true for most but not all cases, its argumentative strength is less than one.
Thus, in (23), the uncertainty in inferring from the truth of the antecedent to the truth of the consequent does not arise from MP itself, but from its non-maximum argumentative strength. For comparison, the deductive conditional

[21] In fact, there may be a weak relevance connection due to climate change, but this would be a negative one.

(24) If 2*x = 10, then x = 5.

has an argumentative strength of 1 and is truth-preserving. Consequently, it seems advisable to accept MP as a valid conclusion and instead consider the argumentative strength of a conditional for its uncertainty. In case the argumentative strength is less than 1, the inference from the truth of the antecedent and the truth of the conditional to the truth of the consequent may be false – but not because MP is invalid, but because the argument is; for example, one of the premises may not be true in this specific instance. This also fits well with Douven et al. (2023, p. 200)'s inference heuristic, which states that the probability that a conditional is true is "the likelihood that we can make a compelling case for the consequent, starting from the antecedent plus background knowledge".

Moreover, this understanding also fits well with the previously discussed aspect of inferentialism, the problem that inductive or statistical connections can be too permissive. Understanding it in this way not only allows MP to be considered valid, but also strengthens the inductive relationship to the point where unrelated correlations are no longer sufficient for a condition to be considered true. Specifically, inductive arguments could be understood as those that have the same form as deductive arguments but are inconclusive for some reason, e.g., because there are exceptions or possible preventions. For example, (23) can be supported by an argument whose premises state that owning a home in Chelsea is expensive and that only rich people can afford expensive housing. Nevertheless, exceptions are possible; e.g., one can live with a friend or has only recently become poor. Consequently, a conditional has a deductive relevance connection if there is a compelling and conclusive argument from the antecedent to the consequent. In case the argument is compelling but inconclusive, the conditional has an inductive relevance connection. In case there is no compelling argument, but only an unrelated correlation, as in (22), a conditional has no relevance context and is not considered acceptable despite its inductive or statistical generalisability.

Overall, none of the four aspects examined opposes inferentialism in its entirety, and it seems that they can be at least partially resolved. Nonetheless, they pose a challenge to inferentialism and must be addressed in case inferentialism is to be used to evaluate the truth of various types of conditionals. This is especially true for the exact specification of the different types of relevance connections – what types there are, how they are exactly defined, and how they can be formalised. This being the case, it has to be agreed with Douven et al. (2023, ch. 3.1) that inferentialism is still under development, and it is to be hoped that the open questions can be solved soon.

Günther's causality approach's main idea is described in Section 5.2.7. One aspect that requires a more thorough consideration is the acceptance of indicative conditionals compared to subjunctive conditionals. For this, Günther (2022, p. 620) provides an example in which one supposes that on a Sunday night one approaches

a small town that has exactly two snackbars. Seeing a person eating a hamburger shortly before entering the town, Günther argues that one has a good reason to accept

(25) If snackbar A is closed, then snackbar C is open.

After entering town, one sees that snackbar A is in fact open. Günther (2022, pp. 620-622) shows that under these circumstances

(26) If snackbar A were closed, then snackbar C would be open.

is not accepted by the approach and argues that this is desired for the following reason: Indicative conditionals such as (25) are understood epistemically and show how one revises one's belief on learning the antecedent. In contrast, subjunctive conditionals such as (26) tell how the world would be in case the antecedent were true. From this, Günther (2022, p. 620) concludes that (26) must be rejected because there is no causal connection between the antecedent and the consequent.
While Günther's reason is correct in itself – the antecedent and the consequent are not causally connected – his conclusion not to accept (26) seems problematic for the following reason: Seeing a person eating a hamburger when entering the town allows one to conclude

(27) Snackbar A is open or snackbar C is open (or both).

Learning later that snackbar A is open does not object to accepting (27) from now on; in fact, it supports it further. However, in case one accepts (27), one also has to accept (26), since (27) provides a relevance connection for (26). More precisely, (27) provides a deductive connection for (26): from (27) $A \vee C$ and (26)'s antecedent $\neg A$ necessarily follows (26)'s consequent C. As a result, (26) should be considered acceptable in the example.
The fact that the conditional is considered unacceptable by Günther's causality approach shows that the approach is too limited in that it can only analyse causal and evidential conditionals, but not non-causal conditionals such as deductive ones. Equally, it does not allow the evaluation of other types of relevance connections, e.g., inductive ones like (23), mathematical ones like (24) or analogical ones like

(28) If Jim's son likes ice skating, he will like ice hockey.

Apart from the fact that the approach can only evaluate causal but not all types of conditionals, the other types are not simply classified as unevaluable but as false; hence, it is not clear when the limits of the approach are exceeded.[22]

[22] This is because in the absence of a causal connection, the approach cannot distinguish whether there is no relevance connection at all or a non-causal, e.g., deductive, one.

Both aspects – the limitation to causally connected conditionals and the impossibility of distinguishing between evaluable and unevaluable conditionals – pose serious challenges to Günther's causality approach. While other challenges appear to be solvable, such as considering uncertainty, at least for the moment, it is not foreseeable how these two main challenges can be solved.

5.7 Conclusion

The chapter shows that most relevance approaches as well as the suppositional account fail to correctly evaluate conditionals which have a relevance connection but where the acceptability of the antecedent has no influence on the acceptability of the consequent. This applies to cases of mutually exclusive, exhaustive antecedents, cases of non-exclusive antecedents, or both. Among others, the evaluation of approaches to conditionals on these cases shows that approaches relying on statistical measures such as ΔP to determine whether a relevance connection exists fail. This is because statistical measures do not measure the strength of the relevance connection ($P(A \models C)$), but only the influence the acceptance of the antecedent has on the acceptance of the consequent ($P(C \mid A)$). Furthermore, it is shown that the relevance connection should be evaluated independently of the presence or absence of other relevance connections. This is because a relevance connection exists independently of others and, in contrast to the acceptance of the consequent, is not influenced by other relevance connections. Besides that, incoherences would arise in case relevance connections are not evaluated independently of others.
Only two approaches, Douven, Elqayam & Krzyżanowska's inferentialism and Günther's causality approach, can correctly capture the two types of conditionals analysed in this chapter. An examination of both approaches in detail shows that the causality approach is too restrictive due to its exclusive focus on causal relationships and cannot successfully evaluate all types of conditional relevance connections, at least at present. Inferentialism, in contrast, is very permissive and requires further specification, especially regarding how the different types of relevance connections can be defined and evaluated, or even formalised. Nevertheless, inferentialism constitutes a promising approach, and its further development could form the basis for a coherent theory of conditionals that meets the expectations for more complex cases. It is hoped that this chapter contributes to this development and points out directions that may be more promising than others.

Chapter 6

Conclusion

The aim of this thesis is to analyse the nature as well as the measurement of intelligence, the inference method abduction, and approaches to conditionals to develop an approach towards how abduction as a powerful reasoning method can serve as a basis for artificial intelligence in order to come closer to the goal of AGI. In the individual chapters, the following insights are gained:
Chapter 2 is concerned with the examination of intelligence to determine its nature and to identify fundamental principles that have to be considered in the creation of AGI. To fulfil goals, AGI must develop skills, i.e., instructions for action that enable the fulfilment of the goals depending on states of the world. It is shown that intelligence is the ability to generate novel skills, which makes it possible to fulfil predefined goals under previously unknown conditions. Novel skills can be created by the application of various reasoning methods such as deduction, induction and abduction, as well as other methods such as abstraction and classification. Due to the nature of perception, intelligence cannot grasp the world as it is, but can only utilise representations that reflect the world indirectly and possibly incompletely and distorted. As representations correspond to the world, intelligence can draw conclusions from them about the world by applying uncertain and contingent reasoning methods. This makes it possible to ascribe functions to representations as to how they can be used to achieve goals; by doing so, representations are attributed meaning. The totality of all existing knowledge forms a world model, which contains, for example, all skills and which can be expanded with the help of reasoning methods and new perceptions. The value of a world model is functionally determined by its its viability, i.e., its potential to fulfil the goals. Because of the uncertainty and contingency of the reasoning methods, many different possible viable conclusions can be drawn. As a consequence, the world model is constructivist, i.e., the conclusions drawn do not represent the world truthfully but only correspond to it. The methods of reasoning represent assumptions about the world; due to the No Free Lunch theorems, it is necessary to provide at least some assumptions as axioms for intelligence. However, intelligence is only successful if the assumptions apply to the world

in which it is used, which is why they should be determined prudently. Overall, intelligence is considered as an algorithm for an optimisation problem whose task is to find optimal actions to fulfil particular goals in a partially unknown world. This interpretation relies on a naturalistic approach and does not require the assumption of mental features, such as consciousness, which are considered to be independent of intelligence. The performance of AGI is determined by how comprehensively it can perceive the world, how comprehensively it can manipulate the world, how comprehensively it can apply reasoning and other methods, and how efficient and consistent with the world the assumptions on which it is based are.

For the development of AGI, it is necessary to be able to measure the degree of intelligence that an AI approach exhibits. This in particular as OpenAI recently presented the generative AI model o3, which achieved a solution rate of 87.5 % on the ARC-AGI benchmark, a benchmark developed specifically to measure intelligence, raising the question of whether AGI has already been achieved. Chapter 3 therefore examines the measurement of intelligence, drawing on the conception of intelligence developed in the foregoing chapter. Hereby it is shown that although the ARC-AGI benchmark is not designed to measure skills but the creation of skills, i.e. intelligence, it possesses several weaknesses. It is possible to solve its tasks by exploiting a given problem space without having to comprehend the problem beforehand – however, this is the much more difficult part of solving problems and must also be covered by AI systems, as the phenomenological analysis in Section 2.6 shows. In addition, ARC-AGI allows in theory an unlimited number of possible solutions to be tried out due to the structure of the tasks. OpenAI's o3 model exploits both weaknesses by generating a large number of solutions and trialling them until it finds the correct solution. Since this method is not only computationally very intensive, but also works for only a few problems – in general, one has only one or a few attempts to solve a problem – a new benchmark is outlined. This is based on the conception of intelligence developed in this thesis and involves testing AI approaches in virtual, unknown worlds in which they have to fulfil a variety of tasks. Equally, the goals to be achieved are unknown in advance. As a result, AI approaches can only achieve their goals by utilising intelligence to develop a model of the world. Thereby, an AI approach is the more intelligent, the more efficiently it can fulfil the more goals in the more worlds. All other attempts, such as providing the AI approaches with knowledge, i.e. skills, in advance, or trialling a large number of solutions, would not lead to success.

One of the most powerful reasoning methods that can be used for intelligence is abduction. That is because abduction makes it possible to introduce new concepts that are composed of several others and thus constitute something novel. For this reason, Chapter 4 examines abduction in detail with the aim of laying the foundation for a theory of abduction that is complete, i.e. covers both the generation and justification of hypotheses, and is formalisable, enabling its application in artificial

intelligence. Based on an analysis of Peirce's retroduction, Lipton's Inference to the Best Explanation and other theories, a new theory of abduction is proposed. It considers abduction not as intrinsically explanatory but as intrinsically conditional: for a given fact, abduction allows one to infer a fact that implies it. There are three types of abduction: Selective abduction selects an already known conditional whose consequent is the given fact and infers that its antecedent is true. Conditional-creative abduction creates a new conditional in which the given fact is the consequent and a fact defined by the theory is the antecedent that implies the given fact. Propositional-conditional-creative abduction assumes that the given fact is implied by a hitherto undefined fact and thus creates a new conditional with a new proposition as antecedent. The execution of abductive inferences is specified by theory-specific patterns. Each pattern consists of a set of rules for both generating and justifying abductive conclusions and covers the complete inference process. In consequence, abductive inferences can be formalised iff the whole pattern can be formalised. The empirical consistency of the proposed theory is demonstrated by a case study of Semmelweis' research on puerperal fever.

The proposed theory of abduction is based on conditionals and requires therefore a theory of conditionals for its formalisation. However, although there are many different approaches to conditionals, they are all controversial and have various weaknesses and limitations. To identify a suitable theory of conditionals, Chapter 5 analyses the suppositional account and different approaches to relevance conditionals on a particular type of conditional: Conditionals whose antecedent and consequent have a relevance connection, but where the acceptability of the antecedent has no influence on the acceptability of the consequent. Such conditionals occur in cases of multiple implication of a consequent, as in overdetermination. When evaluating the conditionals, the approaches examined lead to different and partly incoherent results. This applies to cases of mutually exclusive, exhaustive antecedents, cases of non-exclusive antecedents, or both. Among others, the evaluation on these cases shows that approaches of conditionals relying on statistical measures such as ΔP to determine whether a relevance connection exists fail. This is because statistical measures do not measure the strength of the relevance connection ($P(A \models C)$), but only the influence the acceptance of the antecedent has on the acceptance of the consequent ($P(C \mid A)$). Furthermore, it is shown that the relevance connection should be evaluated independently of the presence or absence of other relevance connections. This is because a relevance connection exists independently of others and, in contrast to the acceptance of the consequent, is not influenced by other relevance connections. Besides that, incoherences would arise in case relevance connections are not evaluated independently of others. Only two approaches, Douven, Elqayam & Krzyżanowska's inferentialism and Günther's causality approach, can correctly capture the two types of conditionals analysed in the chapter. An examination of both approaches in detail shows that the causality approach is too restrictive due to

its exclusive focus on causal relationships and cannot successfully evaluate all types of conditional relevance connections, at least at present. Inferentialism, in contrast, is very permissive and requires further specification, especially regarding how the different types of relevance connections can be defined and evaluated, as well as formalised. Nevertheless, inferentialism constitutes a promising approach, and its further development could form the basis for a coherent theory of conditionals that meets the requirements for more complex cases.

Overall, the thesis shows that there are many open questions that require further research on intelligence as well as on abduction and on conditionals. The aim of this thesis is to take a step in this direction and to introduce a possible approach to the application of abduction that can lead to the development of new, more efficient methods of automated data processing. Abduction is a powerful inference method which allows the introduction of novel, composed concepts and if it is possible to develop systems of automated data processing that can apply abduction in a targeted manner, it can lead to a significant increase in their performance. The thesis therefore intends to lay a foundation for further research, including the development and formalisation of a theory of abduction as well as its implementation in AI approaches as a further step towards AGI.

Chapter 7

Summary of the Thesis in German

Die automatisierte Verarbeitung von Daten, oft als Künstliche Intelligenz (KI) bezeichnet, hat die Entwicklung menschlicher Gesellschaften in den letzten Jahrzehnten wie kaum ein anderes Thema beeinflusst. Mit dem Aufkommen elektronischer Computer in den 1940er Jahren und der Entwicklung von Transistoren in den 1950er Jahren wurde es möglich, die zuvor entwickelten theoretischen Überlegungen, insbesondere im Bereich der Logik, einer breiteren Anwendung zuzuführen. Die automatisierte Datenverarbeitung führte zu zahlreichen bekannten Erfolgen wie dem Knacken der Enigma-Verschlüsselung durch Alan Turing, dem Sieg von Deep Blue gegen den damals amtierenden Weltmeister Garris Kasparov im Schach und dem Sieg von AlphaGo gegen Lee Sedol im Go. Aufgrund der enormen Fortschritte bei der Entwicklung von Rechenkapazitäten und der massiven Sammlung von Daten wird die automatisierte Datenverarbeitung heute in allen Bereichen der Gesellschaft eingesetzt und bietet zahlreiche Vor- und Nachteile. Dazu hat auch die Entwicklung neuer Ansätze zur Datenverarbeitung, sprich von Algorithmen, wesentlich beigetragen. Dazu gehören insbesondere Deep-Learning-Ansätze, die auf stark vereinfachten Konzepten von Neuronen beruhen und künstliche neuronale Netze aus einer großen Anzahl solcher künstlicher Neuronen bilden. Zu den neuesten Deep-Learning-Ansätzen gehört die generative KI, die die Grundlage für Transformer-basierte Modelle wie die GPT-Serie von OpenAI oder die Gemini-Serie von Google bildet.
Trotz all dieser Fortschritte erfüllen die bestehenden Methoden der automatisierten Datenverarbeitung nicht die Erwartungen vieler Vertreter aus dem Bereich der Künstlichen Intelligenz: Ein System der automatisierten Datenverarbeitung, oft als Künstliche Allgemeine Intelligenz (Artificial General Intelligence (AGI)) bezeichnet, das mindestens so gut wie ein durchschnittlicher Mensch in der menschlichen Welt zurechtkommt und die Vielzahl der dort anfallenden Aufgaben erledigen kann. Obwohl eine große Anzahl von übermenschlichen Erfolgen bei spezifischen Aufgaben erzielt wurde, gibt es kein System, das in der Lage ist, ein so breites Spektrum an Aufgaben lösen zu können wie Menschen. Die vorliegende Arbeit will sich dieser Diskrepanz aus einer philosophischen Perspektive nähern, indem sie Erkenntnisse

und Methoden der Philosophie nutzt, um zu analysieren, was Intelligenz ist und welche grundlegenden Prinzipien bei der Schaffung Künstlicher Allgemeiner Intelligenz zu berücksichtigen sind. Die Philosophie scheint für die Analyse besonders geeignet, nicht nur, weil die Logik, die die Grundlage der automatisierten Datenverarbeitung bildet, eine große Bedeutung sowie ihre Wurzeln in der Philosophie hat, sondern auch, weil sich die Philosophie seit langem intensiv mit der Analyse des Denkens und den Methoden des Erkenntnisgewinns beschäftigt.
In Kapitel 2 wird das Wesen von Intelligenz und ihr zugrundeliegende Prinzipien, welche für die Erschaffung von Allgemeiner Künstlicher Intelligenz von Bedeutung sind, untersucht. Kapitel 3 befasst sich damit, wie Intelligenz gemessen werden kann, untersucht die aktuellen Erfolge von OpenAI's o3 Modell auf dem ARC-AGI Benchmark und skizziert ein neues Benchmark für Intelligenz. Eines der Ergebnisse bei der Untersuchung von Intelligenz ist, dass Abduktion, eine Schlussfolgerungsmethode zur Ableitung neuen Wissens, erheblich zur Leistungsfähigkeit von Intelligenz beitragen kann: Durch Abduktion lassen sich neue Konzepte einführen, die aus mehreren anderen Konzepten zusammengesetzt sind und somit etwas Neues darstellen. Beispiele sind das Konzept der Schwerkraft und Modelle über die Beschaffenheit von Atomen, welche allesamt leistungsfähige Vorhersagen ermöglichen. Die Einführung neuer, zusammengesetzter Konzepte ist ein Alleinstellungsmerkmal der Abduktion im Vergleich zu den beiden anderen Schlussfolgerungsmethoden Deduktion und Induktion und macht sie zur mächtigsten Schlussfolgerungsmethode. Gleichzeitig ist die Abduktion eine der am wenigsten erforschten Schlussfolgerungsmethoden, sowohl im Bereich der Philosophie als auch im Bereich der Künstlichen Intelligenz. Aus diesem Grund wird in Kapitel 4 die Abduktion und ihre Formalisierbarkeit im Detail untersucht und gezeigt, dass sie inhärent auf Konditionalen basiert. Da es bis heute keine allgemein anerkannte Theorie zu Konditionalen gibt, werden in Kapitel 5 mehrere Theorien zu Konditionalen an konkreten Beispielen evaluiert, um festzustellen, welche der Theorien eine Grundlage für die Abduktion bilden können. Insgesamt besteht das Ziel dieser Arbeit darin, einen Ansatz für die Entwicklung neuer, leistungsfähigerer Verfahren Künstlicher Intelligenz aufzuzeigen, welche im Gegensatz zu bisherigen Modellen nicht nur zu zuverlässigem deduktiven und induktiven, sondern auch abduktiven Schließen fähig sind.
Hinsichtlich der Struktur des Werkes sind zwei Aspekte zu berücksichtigen. Zum einen bauen die Kapitel aufeinander auf, die späteren Kapitel behandeln jeweils einen bestimmten Aspekt ihnen vorausgehender Kapitel ausführlicher. Während sich Kapitel 2 mit Intelligenz im Allgemeinen befasst, wird in Kapitel 3 ein Aspekt von Intelligenz untersucht, nämlich wie diese gemessen werden kann und wie ein Benchmark hierfür aussehen kann. In Kapitel 4 wird Abduktion im Detail untersucht, welche für die Leistungsfähigkeit von Intelligenz wesentlich ist und welche gleichermaßen eine Komponente von dieser darstellt. Ebenso befasst sich Kapitel 5, in dem verschiedene Theorien zu Konditionalen analysiert werden, im Detail mit

einem spezifischen, aber grundlegenden Aspekt der Abduktion, welche im vorangehenden Kapitel erörtert wird.
Zweitens sind die Veröffentlichungen, auf denen die Kapitel beruhen, im Laufe mehrerer Jahre entstanden und nicht in der hier dargestellten Reihenfolge. Stattdessen wurde zuerst die Publikation über Abduktion (Kapitel 4) veröffentlicht und anschließend die Publikation über Konditionale (Kapitel 5). Danach kam es zur Veröffentlichung der Publikation über das Messen von Intelligenz (Kapitel 3) sowie der Publikation über das Verständnis von Intelligenz (Kapitel 2). Infolgedessen sind die Kapitel miteinander kompatibel, aber es gibt Unterschiede in den Einzelheiten. So entwickelt das Kapitel über das Verständnis von Intelligenz ein konstruktivistisches Weltverständnis, während das Kapitel über Abduktion (noch) ein wahrheitsbasiertes Weltverständnis vertritt. Dies stellt jedoch keine Inkompatibilität dar, da der Ansatz der Abduktion auch mit einem konstruktivistischen Weltverständnis angewendet werden kann.
Basierend auf den Untersuchungen werden in den einzelnen Kapiteln die folgenden Erkenntnisse gewonnen:
Kapitel 2 befasst sich mit der Untersuchung von Intelligenz, um ihre Beschaffenheit zu bestimmen und um grundlegende Prinzipien zu identifizieren, die bei der Schaffung von Allgemeiner Künstlicher Intelligenz zu berücksichtigen sind. Um vorgegebene Ziele erfüllen zu können, muss Allgemeine Künstliche Intelligenz Fähigkeiten entwickeln können, das heißt Handlungsoptionen bestimmen, die die Erfüllung der Ziele in Abhängigkeit von Zuständen der Welt ermöglichen. Es wird gezeigt, dass Intelligenz die Fähigkeit ist, neuartige Fähigkeiten zu erzeugen, die es ermöglichen, vorgegebene Ziele unter bisher unbekannten Bedingungen zu erfüllen. Neuartige Fähigkeiten für bisher unbekannte Bedingungen können durch die Anwendung verschiedener Schlussmethoden wie Deduktion, Induktion und Abduktion sowie anderer Methoden wie Abstraktion und Klassifikation erzeugt werden. Aufgrund der Natur der Wahrnehmung kann Intelligenz die Welt nicht so erfassen, wie sie an sich ist, sondern lediglich auf Repräsentationen zurückgreifen, welche die Welt indirekt und möglicherweise unvollständig und verzerrt wiedergeben. Da Repräsentationen dennoch mit der Welt korrespondieren, kann Intelligenz aus ihnen Schlussfolgerungen über die Welt ziehen, indem sie unsichere und kontingente Schlussmethoden anwendet. Dadurch ist es möglich, Repräsentationen Funktionen zuzuschreiben, wie die Repräsentationen zur Erreichung von Zielen eingesetzt werden können; hierdurch wird den Repräsentationen Bedeutung zugeschrieben. Die Gesamtheit des vorhandenen Wissens bildet ein Weltmodell, das zum Beispiel alle Fähigkeiten enthält und das mit Hilfe von Schlussmethoden und neu erworbenem Wissen, beispielsweise durch Wahrnehmung, erweitert werden kann. Der Wert eines Weltmodells wird funktional anhand seiner Viabilität, das heißt seinem Potenzial zur Erfüllung der Ziele, bestimmt. Aufgrund der Unsicherheit und der Kontingenz vieler Schlussmethoden können zahlreiche verschiedene Schlussfolgerungen gezogen werden, die sowohl

möglich als auch viabel sind. Infolgedessen sind Weltmodelle konstruktivistisch, d. h. die gezogenen Schlussfolgerungen bilden die Welt nicht wahrheitsgetreu ab, sondern korrespondieren lediglich mit ihr. Die Methoden des Schlussfolgerns stellen Annahmen über die Welt dar. Aufgrund der No Free Lunch-Theoreme ist es notwendig, Annahmen zu treffen, da nur aufgrund dieser Intelligenz Ziele mit überdurchschnittlichem Erfolg erfüllen kann. Intelligenz ist gleichwohl nur dann erfolgreich, wenn die Annahmen auf die Welt, in der sie angewendet wird, zutreffen, weshalb die Annahmen mit Sorgfalt getroffen werden sollten. Insgesamt wird Intelligenz als ein Algorithmus für ein Optimierungsproblem betrachtet, dessen Aufgabe es ist, in einer teilweise unbekannten Welt optimale Handlungen zur Erfüllung bestimmter Ziele zu finden. Diese Interpretation beruht auf einem naturalistischen Ansatz und erfordert nicht die Annahme von mentalen Merkmalen wie Intentionalität oder Bewusstsein, die als unabhängig von Intelligenz betrachtet werden. Die Leistung einer Allgemeinen Künstlichen Intelligenz wird daran bemessen, wie umfassend sie die Welt wahrnehmen kann, wie umfassend sie die Welt manipulieren kann, wie umfassend sie Schlussmethoden und andere Methoden anwenden kann und wie mächtig und mit der Welt übereinstimmend die Annahmen sind, auf denen sie basiert.

Für die Entwicklung von Allgemeiner Künstlicher Intelligenz ist es notwendig, den Grad der Intelligenz von einem KI-Verfahren messen zu können. Dies insbesondere, da kürzlich von OpenAI das generative KI-Modell o3 vorgestellt wurde, welches auf dem ARC-AGI Benchmark, einem Test, der dezidiert zur Messung von Intelligenz entwickelt wurde, einen Lösungsrate von 87.5 % erzielt hat und sich somit die Frage stellt, ob Allgemeine Künstliche Intelligenz bereits erreicht wurde. Kapitel 3 befasst sich deshalb mit der Messung von Intelligenz; dies unter Berücksichtigung der im vorangehenden Kapitel entwickelten Konzeption von Intelligenz. Dabei zeigt sich, dass das ARC-AGI Benchmark zwar darauf ausgerichtet ist, nicht bestimmte Fähigkeiten sondern die Entwicklung von neuen Fähigkeiten, sprich Intelligenz, zu messen, jedoch über mehrere Schwächen verfügt. So können die Aufgaben gelöst werden indem ein vorgegebener Problemraum abgesucht wird, ohne dass das Problem zuvor erfasst werden musste – dies ist jedoch der deutlich schwierigere Teil in der Lösung von Problemen und muss von KI-Systemen ebenfalls abgedeckt werden, wie auch die phänomenologische Analyse in der Sektion 2.6 zeigt. Darüber hinaus erlaubt ARC-AGI aufgrund der Struktur der Aufgaben, mögliche Lösungen theoretisch unlimitiert auszuprobieren. OpenAI's o3 Modell nutzt beide Schwächen in Kombination aus, indem es eine große Anzahl an Lösungen generiert und diese ausprobiert, bis es die korrekte Lösung findet. Da dieses Verfahren nicht nur sehr rechenintensiv ist, sondern auch für nur wenige Probleme funktioniert – Menschen haben im Alltag üblicherweise nur einen oder ein paar Versuche, um eine Aufgabe zu lösen – wird ein neuer Benchmark skizziert. Dieser basiert auf der in dieser Arbeit entwickelten Konzeption von Intelligenz und sieht das Testen von KI-Verfahren in

virtuellen Welten vor, in welcher diese verschiedene Aufgaben erfüllen müssen. Jede virtuelle Welt ist künstlich generiert und hat nur wenige Gemeinsamkeiten mit den anderen Welten; zudem ist nichts über die Welten im Vorhinein bekannt. Ebenso sind die zu erreichenden Ziele im Vorfeld unbekannt. Aufgrund dessen können KI-Verfahren die Ziele nur erreichen, indem sie Intelligenz nutzen, um ein Modell der Welt zu entwickeln, das hierfür geeignete Handlungsoptionen bestimmt. Dabei ist ein KI-Verfahren umso intelligenter, je effizienter es je mehr Ziele in je mehr Welten erfüllen kann. Alle anderen Vorgehensweisen, wie den KI-Modellen Wissen, sprich Fähigkeiten, vorzugeben oder eine große Anzahl an Lösungen zu testen, würden nicht zum Erfolg führen.
Eine der leistungsfähigsten Schlussmethoden, die für Intelligenz genutzt werden kann, ist Abduktion. Der Grund dafür ist, dass Abduktion es ermöglicht, neue Konzepte einzuführen, die aus mehreren anderen Konzepten zusammengesetzt sind und somit etwas Neuartiges darstellen. Aus diesem Grund wird in Kapitel 4 die Abduktion eingehend untersucht, um die Grundlage für eine Theorie der Abduktion zu schaffen, die vollständig ist, sprich sowohl den Generierungs- als auch den Begründungskontext abdeckt, und die formalisierbar ist, was ihre Anwendung in der Künstlichen Intelligenz ermöglicht. Auf der Grundlage einer Analyse verschiedener bestheneder Theorien der Abduktion wie der Retroduktion von Peirce und dem Schluss auf die beste Erklärung (Inference to the Best Explanation (IBE)) von Lipton wird eine neue Theorie der Abduktion vorgeschlagen. Sie betrachtet Abduktion nicht als intrinsisch erklärend, sondern als intrinsisch konditional: Für einen gegebenen Sachverhalt kann durch Abduktion auf einen Sachverhalt geschlossen werden, der diesen impliziert. Es gibt drei Arten von Abduktion: Die selektive Abduktion wählt ein bereits bekanntes Konditional aus, dessen Konsequenz der gegebene Sachverhalt ist, und folgert daraus, dass die Antezedens des Konditionals wahr ist. Die konditional-kreative Abduktion erzeugt ein neues Konditional, bei dem der gegebene Sachverhalt die Konsequenz und ein in der vorhandenen Theorie bereits definierter Sachverhalt die Antezedenz ist, welche den gegebenen Sachverhalt impliziert. Die propositional-konditional-kreative Abduktion geht davon aus, dass der gegebene Sachverhalt durch einen bisher nicht definierten Sachverhalt impliziert wird und bildet somit einen neuen Konditional mit einem neuen Sachverhalt als Antezedens. Die Ausführung der abduktiven Schlüsse wird durch theorie-spezifische Schemata bestimmt. Jedes Schema besteht aus einem Satz von Regeln zur Erzeugung sowie zur Begründung abduktiver Schlussfolgerungen und deckt den gesamten Prozess ab. Entsprechend können abduktive Schlussfolgerungen genau dann formalisiert werden, wenn das gesamte Schema formalisiert werden kann. Die empirische Konsistenz der vorgeschlagenen Theorie wird anhand einer Fallstudie über Semmelweis' Forschung zum Kindbettfieber demonstriert.
Die vorgeschlagene Theorie der Abduktion basiert auf Konditionalen, weshalb für die Formalisierung von Abduktion eine Theorie zu Konditionalen erforderlich ist.

Wenngleich es viele verschiedene Ansätze zu Konditionalen gibt, sind alle umstritten und haben verschiedene Schwächen und Einschränkungen. Infolgedessen werden in Kapitel 5 der suppositionelle Ansatz und verschiedene Relevanz-Ansätze für Konditionale anhand eines bestimmten Typs von Konditionalen analysiert: Konditionale, deren Antezedens und Konsequenz eine Relevanzverbindung haben, bei denen aber die Akzeptanz der Antezedens keinen Einfluss auf die Akzeptanz der Konsequenz hat. Solche Konditionale treten bei mehrfacher Implikation einer Konsequenz auf, zum Beispiel bei deren Überdetermination. Bei der Bewertung solcher Konditionale führen die untersuchten Ansätze zu unterschiedlichen und teilweise inkohärenten Ergebnissen. Dies entweder in Fällen sich gegenseitig ausschließender und ihrer Gesamtheit vollständiger Antezedenzien oder in Fällen sich nicht ausschließender Antezedenzien, oder in beiden Fällen. Die Bewertung von Ansätzen zu Konditionalen für diese Fälle zeigt unter anderem, dass Ansätze, die sich auf statistische Maße wie ΔP stützen, um zu bestimmen, ob ein Relevanzzusammenhang besteht, nicht zum korrekten Ergebnis kommen. Dies liegt daran, dass statistische Maße nicht die Stärke des Relevanzzusammenhangs ($P(A \models C)$) messen, sondern nur den Einfluss, den die Akzeptanz der Antezedens auf die Akzeptanz der Konsequenz hat ($P(C \mid A)$). Außerdem wird gezeigt, dass der Relevanzzusammenhang unabhängig vom Vorhandensein oder Fehlen anderer Relevanzzusammenhänge bewertet werden sollte. Dies liegt daran, dass ein Relevanzzusammenhang unabhängig von anderen existiert und im Gegensatz zur Akzeptanz der Konsequenz nicht von anderen Relevanzzusammenhängen beeinflusst wird. Außerdem würden Inkohärenzen entstehen, wenn Relevanzzusammenhänge nicht unabhängig von anderen bewertet werden. Nur zwei Ansätze, der Inferentialismus von Douven, Elqayam & Krzyżanowska und der Kausalitätsansatz von Günther, können die beiden in diesem Kapitel untersuchten Arten von Konditionalen korrekt bewerten. Eine detaillierte Betrachtung beider Ansätze zeigt, dass der Kausalitätsansatz aufgrund seiner ausschließlichen Fokussierung auf kausale Beziehungen zu restriktiv ist und zumindest derzeit nicht alle Arten von konditionalen Relevanzzusammenhängen erfolgreich bewerten kann. Der Inferentialismus hingegen ist sehr permissiv und bedarf weiterer Spezifizierung, insbesondere hinsichtlich der Frage, wie die verschiedenen Arten von Relevanzzusammenhängen definiert und bewertet sowie formalisiert werden können. Dennoch stellt der Inferentialismus einen vielversprechenden Ansatz dar, dessen Weiterentwicklung die Grundlage für eine kohärente Theorie zu Konditionalen bilden könnte, welche auch in der Lage ist, eine Vielzahl von komplexeren Fällen korrekt zu bewerten.
Insgesamt zeigt die Arbeit, dass es viele offene Fragen gibt, die weiterer Forschung bedürfen – sowohl zum Thema Intelligenz als auch zu Abduktion und Konditionalen. Dies betrifft beispielsweise die Spezifizierung und Formalisierung einer Theorie von Konditionalen ebenso wie die Formalisierung verschiedener abduktiver Schemata. Das Ziel der Arbeit ist es, Beiträge, die als Grundlage für diese weiteren Forschungen dienen können, bereitzustellen. Ebenso ist das Ziel einen Ansatz aufzuzeigen,

wie die Entwicklung neuer, leistungsfähigerer Methoden der automatisierten Datenverarbeitung, welche auf der Abduktion beruhen, angestrebt werden kann. Da die Abduktion eine mächtige Schlussfolgerungsmethode darstellt, welche die Einführung neuartiger, zusammengesetzter Konzepte ermöglicht, kann ihre erfolgreiche Implementierung dazu führen, dass KI-Ansätze deutlich bessere Modelle der Welt entwickeln können. Dies kann einen wichtigen Schritt für die Entwicklung von Allgemeiner Künstlicher Intelligenz darstellen.

Chapter 8

Bibliography

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3.

Aliseda, A. (2006). *Abductive reasoning*, volume 330. Springer.

Anderson, D. R. (1987). *Creativity and the Philosophy of CS Peirce*, volume 27. Springer Science & Business Media.

ARC Prize (2024). Arc prize 2024. `https://arcprize.org/2024-results`.

Baker, A. (2020). Non-deductive methods in mathematics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition.

Baker, C. V., Modrell, M. S., and Gillis, J. A. (2013). The evolution and development of vertebrate lateral line electroreceptors. *Journal of Experimental Biology*, 216(13):2515–2522.

Baratgin, J., Over, D. E., and Politzer, G. (2013). Uncertainty and the de finetti tables. *Thinking & Reasoning*, 19(3-4):308–328.

Bartha, P. (2019). Analogy and analogical reasoning. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2019 edition.

Beckmann, P., Köstner, G., and Hipólito, I. (2023). An alternative to cognitivism: computational phenomenology for deep learning. *Minds and Machines*, 33(3):397–427.

Beirlaen, M. and Aliseda, A. (2014). A conditional logic for abduction. *Synthese*, 191:3733–3758.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. (2023). The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv:2309.12288*.

Berto, F. and Özgün, A. (2021). Indicative conditionals: Probabilities and relevance. *Philosophical Studies*, 178(11):3697–3730.

Bird, A. (2010). Eliminative abduction: Examples from medicine. *Studies in History and Philosophy of Science Part A*, 41(4):345–352.

Brooks, R. A. (1991a). Intelligence without reason. *MIT AI Laboratory: AI Memos*.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv:2303.12712*.

Cabrera, F. (2017). Can there be a bayesian explanationism? on the prospects of a productive partnership. *Synthese*, 194(4):1245–1272.

Cabrera, F. (2020). Does ibe require a 'model'of explanation? *The British Journal for the Philosophy of Science*.

Campos, D. G. (2011). On the distinction between peirce's abduction and lipton's inference to the best explanation. *Synthese*, 180:419–442.

Carey, S. (2000). The origin of concepts. *Journal of Cognition and Development*, 1(1):37–41.

Carter, C. (1983). *The etiology, concept, and prophylaxis of childbed fever (Excerpt)*. Number 2 in Social Medicine. Univ of Wisconsin Press.

Chollet, F. (2019). On the measure of intelligence. *arXiv:1911.01547*.

Chollet, F. (2024). Openai o3 breakthrough high score on arc-agi-pub. `https://arcprize.org/blog/oai-o3-pub-breakthrough`.

Chollet, F., Knoop, M., Kamradt, G., and Landers, B. (2024a). Arc prize 2024: Technical report. *arXiv:2412.04604*.

Chollet, F., Knoop, M., Landers, B., Kamradt, G., Jud, H., Reade, W., and Howard, A. (2024b). Arc prize 2024. `https://kaggle.com/competitions/arc-prize-2024`. Kaggle.

Chollet, F., Tong, K., Reade, W., and Elliott, J. (2020). Abstraction and reasoning challenge. `https://kaggle.com/competitions/abstraction-and-reasoning-challenge`. Kaggle.

Crupi, V. and Iacona, A. (2021). Probability, evidential support, and the logic of conditionals. *Argumenta*, 6:211–222.

Crupi, V. and Iacona, A. (2022a). The evidential conditional. *Erkenntnis*, 87(6):2897–2921.

Crupi, V. and Iacona, A. (2022b). Three ways of being non-material. *Studia Logica*, pages 1–47.

Crupi, V. and Iacona, A. (2023). Outline of a theory of reasons. *The Philosophical Quarterly*, 73(1):117–142.

Cruz, N. and Oberauer, K. (2014). Comparing the meanings of "if" and "all". *Memory & cognition*, 42:1345–1356.

Cummings, M. L. and Bauchwitz, B. (2024). Unreliable pedestrian detection and driver alerting in intelligent vehicles. *IEEE Transactions on Intelligent Vehicles*.

Dawes, G. W. (2013). Belief is not the issue: A defence of inference to the best explanation. *Ratio*, 26(1):62–78.

Dohare, S., Hernandez-Garcia, J. F., Lan, Q., Rahman, P., Mahmood, A. R., and Sutton, R. S. (2024). Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774.

Douven, I. (2015). *The epistemology of indicative conditionals: Formal and empirical approaches*. Cambridge University Press.

Douven, I. (2017a). Abduction. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2017 edition.

Douven, I. (2017b). How to account for the oddness of missing-link conditionals. *Synthese*, 194:1541–1554.

Douven, I. (2020). The ecological rationality of explanatory reasoning. *Studies in History and Philosophy of Science Part A*, 79:1–14.

Douven, I., Elqayam, S., Gärdenfors, P., and Mirabile, P. (2022a). Conceptual spaces and the strength of similarity-based arguments. *Cognition*, 218:104951.

Douven, I., Elqayam, S., and Krzyżanowska, K. (2023). Inferentialism: A manifesto. In *Conditionals: Logic, Linguistics and Psychology*, pages 175–221. Springer.

Douven, I., Elqayam, S., and Mirabile, P. (2022b). Inference strength predicts the probability of conditionals better than conditional probability does. *Journal of Memory and Language*, 123:104302.

Douven, I., Elqayam, S., Singmann, H., and van Wijnbergen-Huitink, J. (2018). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive psychology*, 101:50–81.

Douven, I. and Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(11):1792–1813.

Douven, I. and Verbrugge, S. (2010). The adams family. *Cognition*, 117(3):302–318.

Dreyfus, H. L. (2002). Intelligence without representation–merleau-ponty's critique of mental representation the relevance of phenomenology to scientific explanation. *Phenomenology and the cognitive sciences*, 1(4):367–383.

Dreyfus, H. L. (2007). Why heideggerian ai failed and how fixing it would require making it more heideggerian. *Philosophical psychology*, 20(2):247–268.

Dreyfus, H. L. and Dreyfus, S. E. (1986). *Mind over machine*. Simon and Schuster.

Dusenbery, D. B. (1992). *Sensory ecology: how organisms acquire and respond to information*. WH Freeman New York.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Le Bras, R., Hwang, J. D., et al. (2023). Faith and fate: Limits of transformers on compositionality (2023). *arXiv:2305.18654*.

Easwaran, K. (2008). The role of axioms in mathematics. *Erkenntnis*, 68:381–391.

Egré, P. and Rott, H. (2021). The logic of conditionals. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2021 edition.

Evans, J. S. B. (2020). The suppositional conditional is not (just) the probability conditional. In *Logic and Uncertainty in the Human Mind*, pages 57–70. Routledge.

Evans, J. S. B. and Over, D. E. (2004). *If: Supposition, pragmatics, and dual processes*. Oxford University Press, USA.

Fann, K. T. (2012). *Peirce's theory of abduction*. Springer Science & Business Media.

Flach, P. A. and Kakas, A. C. (2000). *Abduction and Induction: Essays on their relation and integration*, volume 18. Springer Science & Business Media.

Frith, C. (2007). *Making up the mind: How the brain creates our mental world*. John Wiley & Sons.

Fugard, A. J., Pfeifer, N., Mayerhofer, B., and Kleiter, G. D. (2011). How people interpret conditionals: shifts toward the conditional event. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3):635–648.

Gabbay, D. and Woods, J. (2005). *The Reach of Abduction: Insight and Trial (A Practical Logic of Cognitive Systems, Vol. 2)*. Elsevier.

Gallagher, S. and Zahavi, D. (2020). *The phenomenological mind*. Routledge.

Genesis (2024). Genesis: A universal and generative physics engine for robotics and beyond. https://github.com/Genesis-Embodied-AI/Genesis.

Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology press.

Glasersfeld, E. v. (1996). *Radical constructivism: A way of knowing and learning*. Routledge.

Glymour, C. (2019). Creative abduction, factor analysis, and the causes of liberal democracy. *Kriterion–Journal of Philosophy*, 33(1):1–22.

Godden, D. and Zenker, F. (2015). Denying antecedents and affirming consequents: The state of the art. *Informal Logic*, 35(1).

Griffiths, T. L. and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, 116(4):661–716.

Günther, M. (2022). Causal and evidential conditionals. *Minds and Machines*, 32(4):613–626.

Harman, G. H. (1965). The inference to the best explanation. *The philosophical review*, 74(1):88–95.

Hatfield, G. (1988). Representation and content in some (actual) theories of perception. *Studies in History and Philosophy of Science Part A*, 19(2):175–214.

Haugeland, J. (2000). *Having thought: Essays in the metaphysics of mind*. Harvard University Press.

Haugeland, J. (2013). *Dasein Disclosed: John Haugeland's Heidegger*. Harvard University Press.

Heidegger, M. (1967). *Sein und Zeit*. Max Niemeyer Verlag Tübingen.

Heidegger, M. (1994). *Phänomenologische Interpretationen zu Aristoteles*. Gesamtausgabe / Martin Heidegger. Abt. II, Vorlesungen. Vittorio Klostermann.

Heidegger, M. (1997). *Der Satz vom Grund*. Gesamtausgabe / Martin Heidegger. Abt. I, Veröffentlichte schriften 1910-1976. Vittorio Klostermann.

Heidegger, M. (1999). *Metaphysik und Nihilismus*. Gesamtausgabe / Martin Heidegger. Abt. III, Unveröffentlichte Abhandlungen. Vittorio Klostermann.

Heidegger, M. (2000). *Über den Humanismus*. Vittorio Klostermann.

Heidegger, M. (2001). *Sein und Wahrheit*. Gesamtausgabe / Martin Heidegger. Abt. II, Vorlesungenn. Vittorio Klostermann.

Heidegger, M. (2012). *Der Ursprung des Kunstwerkes*. Vittorio Klostermann.

Hempel, C. (1966). *Philosophy of Natural Science*. Number Bd. 491,S. 1966 in Foundations of philosophy series. Prentice-Hall.

Hernández-Orallo, J. (2017). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48:397–447.

Heron, J. (2021). Set-theoretic justification and the theoretical virtues. *Synthese*, 199(1):1245–1267.

Hodel, M. (2024). Domain specific language for the abstraction and reasoning corpus (arc-dsl). `https://github.com/michaelhodel/arc-dsl/tree/main`.

Hoffstaetter, L. J., Bagriantsev, S. N., and Gracheva, E. O. (2018). Trps et al.: a molecular toolkit for thermosensory adaptations. *Pflügers Archiv-European Journal of Physiology*, 470:745–759.

Hong, P., Ghosal, D., Majumder, N., Aditya, S., Mihalcea, R., and Poria, S. (2024). Evaluating llms' mathematical and coding competency through ontology-guided interventions. *arXiv:2401.09395*.

Horst, S. (2005). Phenomenology and psychophysics. *Phenomenology and the cognitive sciences*, 4:1–21.

Hoyningen-Huene, P. (1987). Context of discovery and context of justification. *Studies in History and Philosophy of Science Part A*, 18(4):501–515.

Hume, D. (2016). An enquiry concerning human understanding. In *Seven masterpieces of philosophy*, pages 183–276. Routledge.

Husserl, E. (1984). *Logische Untersuchungen. Zweiter Band, erster Teil. Edited by U. Panzer.* Springer.

Iriki, A., Tanaka, M., and Iwamura, Y. (1996). Coding of modified body schema during tool use by macaque postcentral neurones. *Neuroreport*, 7(14):2325–2330.

Jantzen, B. C. (2016). Discovery without a 'logic'would be a miracle. *Synthese*, 193(10):3209–3238.

Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W. J., Taylor, C. J., and Roth, D. (2024). A peek into token bias: Large language models are not yet genuine reasoners. *arXiv:2406.11050*.

Jones, N. (2018). Inference to the more robust explanation. *The British Journal for the Philosophy of Science*, 69(1):75–102.

Kant, I. (1968). *Kritik der reinen Vernunft*, volume 3 of *Werke [in zwölf Bänden. Edited by Weischedel, Wilhelm*. Suhrkamp.

Kaufmann, S., Over, D. E., and Sharma, G. (2023). *Conditionals: Logic, Linguistics and Psychology*. Springer Nature.

Kelber, A., Vorobyev, M., and Osorio, D. (2003). Animal colour vision–behavioural tests and physiological concepts. *Biological Reviews*, 78(1):81–118.

Keller, A., Zhuang, H., Chi, Q., Vosshall, L. B., and Matsunami, H. (2007). Genetic variation in a human odorant receptor alters odour perception. *Nature*, 449(7161):468–472.

Khoram, E., Chen, A., Liu, D., Ying, L., Wang, Q., Yuan, M., and Yu, Z. (2019). Nanophotonic media for artificial neural inference. *Photonics Research*, 7(8):823–827.

Klärner, H. (2013). *Der Schluß auf die beste Erklärung*. Walter de Gruyter.

Krzyżanowska, K., Collins, P. J., and Hahn, U. (2021). True clauses and false connections. *Journal of Memory and Language*, 121:104252.

Krzyżanowska, K. and Douven, I. (2018). Missing-link conditionals: pragmatically infelicitous or semantically defective? *Intercultural Pragmatics*, 15(2):191–211.

Krzyżanowska, K., Wenmackers, S., and Douven, I. (2014). Rethinking gibbard's riverboat argument. *Studia Logica*, 102:771–792.

Lab 42 (2023). Arcathon 2023. `https://lab42.global/past-challenges/2023-arcathon`.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Lange, M. (2022). Putting explanation back into "inference to the best explanation". *Noûs*, 56(1):84–109.

Langley, P. (1987). *Scientific discovery: Computational explorations of the creative processes*. MIT press.

Legg, S., Hutter, M., et al. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157:17.

Lewis, M. and Mitchell, M. (2024). Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv:2402.08955*.

Lhermitte, F. (1986). Human autonomy and the frontal lobes. part ii: patient behavior in complex and social situations: the "environmental dependency syndrome". *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 19(4):335–343.

Lipton, P. (2004). *Inference to the Best Explanation. Second edition*. International library of philosophy and scientific method. Routledge/Taylor and Francis Group.

Locke, J. (1847). *An essay concerning human understanding*. Kay & Troutman.

Ma, Y., Tsao, D., and Shum, H.-Y. (2022). On the principles of parsimony and self-consistency for the emergence of intelligence. *Frontiers of Information Technology & Electronic Engineering*, 23(9):1298–1323.

MacIver, M. A., Schmitz, L., Mugan, U., Murphey, T. D., and Mobley, C. D. (2017). Massive increase in visual range preceded the origin of terrestrial vertebrates. *Proceedings of the National Academy of Sciences*, 114(12):E2375–E2384.

Magid, R. W., Sheskin, M., and Schulz, L. E. (2015). Imagination and the generation of new ideas. *Cognitive Development*, 34:99–110.

Magnani, L. (2015). The eco-cognitive model of abduction: $\acute{a}\pi\alpha\gamma\omega\gamma\acute{\eta}$ now: Naturalizing the logic of abduction. *Journal of Applied Logic*, 13(3):285–315.

Magnani, L. et al. (2009). *Abductive cognition: The epistemological and eco-cognitive dimensions of hypothetical reasoning*, volume 3. Springer.

McAuliffe, W. H. (2015). How did abduction get confused with inference to the best explanation? *Transactions of the Charles S. Peirce Society*, 51(3):300–319.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., and Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv:2309.13638*.

McKaughan, D. J. (2008). From ugly duckling to swan: Cs peirce, abduction, and the pursuit of scientific theories. *Transactions of the Charles S. Peirce Society: A Quarterly Journal in American Philosophy*, 44(3):446–468.

Merleau-Ponty, M. (2012). *Phenomenology of Perception*. Routledge.

Mill, J. S. (1974). *A system of logic*. Collected works of John Stuart Mill, vols 7 & 8. Routledge.

Minnameier, G. (2004). Peirce-suit of truth–why inference to the best explanation and abduction ought not to be confused. *Erkenntnis*, 60(1):75–105.

Mirabile, P. and Douven, I. (2020). Abductive conditionals as a test case for inferentialism. *Cognition*, 200:104232.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv:2410.05229*.

Mitchell, M. (2021). Why ai is harder than we think. *arXiv:2104.12871*.

Mondorf, P. and Plank, B. (2024). Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *arXiv:2404.01869*.

Neurath, O. (1932). Protokollsätze. *Erkenntnis*, 3:204–214.

Newman, G. E., Choi, H., Wynn, K., and Scholl, B. J. (2008). The origins of causal perception: Evidence from postdictive processing in infancy. *Cognitive psychology*, 57(3):262–291.

Nezhurina, M., Cipolina-Kun, L., Cherti, M., and Jitsev, J. (2024). Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv:2406.02061*.

Nickerson, R. S., Barch, D. H., and Butler, S. F. (2019). Evaluating conditional arguments with uncertain premises. *Thinking & Reasoning*, 25(1):48–71.

Nietzsche, F. (1982). *Über Wahrheit und Lüge im außermoralischen Sinne.* Friedrich Nietzsche Werke in drei Bänden, Band 3, Herausgegeben von Karl Schlechta. Carl Hanser Verlag, München.

Niiniluoto, I. (2018). *Truth-seeking by abduction*, volume 400. Springer.

Nilsson, D.-E. (2009). The evolution of eyes and visually guided behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1531):2833–2847.

Novick, A. and Scholl, R. (2020). Presume it not: True causes in the search for the basis of heredity. *The British Journal for the Philosophy of Science*.

Oaksford, M. and Chater, N. (2020). Integrating causal bayes nets and inferentialism in conditional inference. *Logic and uncertainty in the human mind*, pages 116–132.

Okasha, S. (2001). What did hume really show about induction? *The Philosophical Quarterly*, 51(204):307–327.

Over, D. E. and Cruz, N. (2017). Probabilistic accounts of conditional reasoning. In *The Routledge international handbook of thinking and reasoning*, pages 434–450. Routledge.

Over, D. E. and Cruz, N. (2023). Indicative and counterfactual conditionals in the psychology of reasoning. In *Conditionals: Logic, Linguistics and Psychology*, pages 139–173. Springer.

Over, D. E., Hadjichristidis, C., Evans, J. S. B., Handley, S. J., and Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive psychology*, 54(1):62–97.

Paavola, S. (2004). Abduction as a logic and methodology of discovery: The importance of strategies. *Foundations of Science*, 9(3):267–283.

Paavola, S. (2006a). Hansonian and harmanian abduction as models of discovery. *International Studies in the Philosophy of Science*, 20(01):93–108.

Paavola, S. (2006b). On the origin of ideas: An abductivist approach to discovery. *Philosophical studies from the University of Helsinki*, 15.

Park, W. (2015). On classifying abduction. *Journal of Applied Logic*, 13(3):215–238.

Peirce, C. S. (1958). *Collected papers of Charles Sanders Peirce. 8 volumes. Edited by C. Hartshorne, & P. Weiss (Eds.) (Vol. 1–6) and A.W. Burks (Ed.) (Vol. 7-8).* Cambridge MA: Belknap Press.

Peirce, C. S. (1998). *The Essential Peirce, Vol. 2: Selected Philosophical Writings (1893-1913)*. The Essential Peirce. Indiana University Press.

Pfister, R. (2022). Towards a theory of abduction based on conditionals. *Synthese*, 200(3):206.

Pfister, R. (2024). The role of overdetermination and alternative implication in the evaluation of conditionals. *PhilSci-Archive*.

Pfister, R. (2025a). Beyond theory: A philosophical framework for decision-making in management. In Hoffmann, C. H., editor, *Artificial Intelligence, Entrepreneurship and Risk: Reflections and Positions at the Crossroads between Philosophy and Management*, Technikzukünfte, Wissenschaft und Gesellschaft / Futures of Technology, Science and Society. Springer VS Wiesbaden. Due April 8, 2025.

Pfister, R. (2025b). A representationalist, functionalist and naturalistic conception of intelligence as a foundation for agi. *arXiv:2503.07600*.

Pfister, R. and Jud, H. (2025). Understanding and benchmarking artificial intelligence: Openai's o3 is not agi. *arXiv:2501.07458*.

Popper, K. (1959). *The logic of scientific discovery*. Basic Books Inc.

Porter, M. L., Blasic, J. R., Bok, M. J., Cameron, E. G., Pringle, T., Cronin, T. W., and Robinson, P. R. (2012). Shedding new light on opsin evolution. *Proceedings of the Royal Society B: Biological Sciences*, 279(1726):3–14.

Prabhakar, A., Griffiths, T. L., and McCoy, R. T. (2024). Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. *arXiv:2407.01687*.

Preston, B. (1993). Heidegger and artificial intelligence. *Philosophy and Phenomenological Research*, 53(1):43–69.

Psillos, S. (2002). Simply the best: A case for abduction. In *Computational logic: Logic programming and beyond: Essays in honour of Robert A. Kowalski part II*, pages 605–625. Springer.

Psillos, S. (2011). An explorer upon untrodden ground: Peirce on abduction. In *Handbook of the History of Logic*, volume 10, pages 117–151. Elsevier.

Putnam, H. et al. (1981). *Reason, truth and history*, volume 3. Cambridge University Press Cambridge.

Qiu, L., Jiang, L., Lu, X., Sclar, M., Pyatkin, V., Bhagavatula, C., Wang, B., Kim, Y., Choi, Y., Dziri, N., et al. (2023). Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv:2310.08559*.

Raidl, E., Iacona, A., and Crupi, V. (2022). The logic of the evidential conditional. *The Review of Symbolic Logic*, 15(3):758–770.

Ramsey, F. P. (1929/1990). General propositions and causality. *Philosophical papers*, pages 145–163.

Reutlinger, A. (2018). Extending the counterfactual theory of explanation. *Explanation beyond causation: Philosophical perspectives on non-causal explanations*, pages 74–95.

Rosenthal, S. (2004). Peirce's pragmatic account of perception: Issues and implications. *The Cambridge Companion to Peirce*, pages 193–213.

Rott, H. (2022a). Difference-making conditionals and the relevant ramsey test. *The Review of Symbolic Logic*, 15(1):133–164.

Rott, H. (2022b). Evidential support and contraposition. *Erkenntnis*, pages 1–19.

Rott, H. (2023). On the logical form of evidential conditionals. *Logic and Logical Philosophy*, pages 1–18.

Rott, H. (2025). Conditionals, support and connexivity. In *60 Years of Connexive Logic*, pages 149–199. Springer.

Russell, S. J. and Norvig, P. (2022). *Artificial intelligence: A modern approach, global edition*. Harlow: Pearson.

Safranski, R. (2014). *Ein Meister aus Deutschland: Heidegger und seine Zeit*. Carl Hanser Verlag.

Schaffer, J. (2005). Contrastive knowledge. *Oxford studies in epistemology*, 1:235–271.

Schickore, J. (2018). Scientific discovery. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2018 edition.

Schlimm, D. (2011). On the creative role of axiomatics. the discovery of lattices by schröder, dedekind, birkhoff, and others. *Synthese*, 183:47–68.

Schneider, W. J. and McGrew, K. S. (2018). The cattell-horn-carroll theory of cognitive abilities. *Contemporary intellectual assessment: Theories, tests, and issues*, pages 73–163.

Scholl, R. (2013). Causal inference, mechanisms, and the semmelweis case. *Studies in History and Philosophy of Science Part A*, 44(1):66–76.

Schurz, G. (2008). Patterns of abduction. *Synthese*, 164:201–234.

Schurz, G. (2016). Common cause abduction: The formation of theoretical concepts and models in science. *Logic Journal of the IGPL*, 24(4):494–509.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

Searle, J. R. (1992). *The rediscovery of the mind*. MIT press.

Sebben, S. and Ullrich, J. (2021). Can conditionals explain explanations? a modus ponens model of b because a. *Cognition*, 215:104812.

Seff, A., Cera, B., Chen, D., Ng, M., Zhou, A., Nayakanti, N., Refaat, K. S., Al-Rfou, R., and Sapp, B. (2023). Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8579–8590.

Semmelweis, I. P. (1861). *Die aetiologie, der begriff und die prophylaxis des kindbettfiebers*. Hartleben. Available online: `https://www.deutschestextarchiv.de/book/show/semmelweis_kindbettfieber_1861`.

Shanahan, M. (2005). Perception as abduction: Turning sensor data into meaningful representation. *Cognitive science*, 29(1):103–134.

Shanahan, M. and Mitchell, M. (2022). Abstraction for deep reinforcement learning. *arXiv:2202.05839*.

Shapiro, L. (2019). *Embodied cognition*. Routledge.

Skinner, B. F. (1948). 'superstition' in the pigeon. *Journal of experimental psychology*, 38(2):168.

Skovgaard-Olsen, N. (2016). Motivating the relevance approach to conditionals. *Mind & Language*, 31(5):555–579.

Skovgaard-Olsen, N. (2020). Relevance and conditionals: A synopsis of open pragmatic and semantic issues. In *Logic and uncertainty in the human mind*, pages 192–206. Routledge.

Skovgaard-Olsen, N., Kellen, D., Hahn, U., and Klauer, K. C. (2019). Norm conflicts and conditionals. *Psychological Review*, 126(5):611.

Skovgaard-Olsen, N., Kellen, D., Krahl, H., and Klauer, K. C. (2017). Relevance differently affects the truth, acceptability, and probability evaluations of "and","but","therefore", and "if–then". *Thinking & Reasoning*, 23(4):449–482.

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition*, 150:26–36.

Smith, D. W. (2018). Phenomenology. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2018 edition.

Spelke, E. (2022). *What babies know: Core knowledge and composition volume 1*, volume 1. Oxford University Press.

Stalnaker, R. C. (1968). A theory of conditionals. In *Ifs: Conditionals, belief, decision, chance and time*, pages 41–55. Springer.

Suk, H., Lee, Y., Kim, T., and Kim, S. (2024). Addressing uncertainty challenges for autonomous driving in real-world environments. In *Advances in Computers*, volume 134, pages 317–361. Elsevier.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Thagard, P. (2012). Creative combination of representations: Scientific discovery and technological invention. *Psychology of science: Implicit and explicit processes*, pages 389–405.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., and Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological review*, 115(1):155.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. (2024). Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Tromp, C. (2024). Creativity from constraint exploration and exploitation. *Psychological reports*, 127(4):1818–1843.

Tschaepe, M. (2014). Guessing and abduction. *Transactions of the Charles S. Peirce Society: A Quarterly Journal in American Philosophy*, 50(1):115–138.

Tye, M. (2021). Qualia. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition.

Valentinčič, T., Wegert, S., and Caprio, J. (1994). Learned olfactory discrimination versus innate taste responses to amino acids in channel catfish (ictalurus punctatus). *Physiology & behavior*, 55(5):865–873.

van Rooij, R. and Schulz, K. (2019). Conditionals, causality and conditional probability. *Journal of Logic, Language and Information*, 28:55–71.

van Rooij, R. and Schulz, K. (2022). Causal relevance of conditionals: semantics or pragmatics? *Linguistics Vanguard*, 8(s4):363–370.

Vidal, M. and Baratgin, J. (2017). A psychological study of unconnected conditionals. *Journal of Cognitive Psychology*, 29(6):769–781.

Ward, J. (2013). Synesthesia. *Annual review of psychology*, 64(1):49–75.

Webb, B. (1993). *Modeling Biological Behaviour or" Dumb Animals and Stupid Robots"*. University of Edinburgh, Department of Artificial Intelligence.

Welling, H. (2007). Four mental operations in creative cognition: The importance of abstraction. *Creativity research journal*, 19(2-3):163–177.

Wheeler, M. (2008). Cognition in context: phenomenology, situated robotics and the frame problem. *International journal of philosophical studies*, 16(3):323–349.

Wolpert, D. H. (2013). Ubiquity symposium: Evolutionary computation and the processes of life: What the no free lunch theorems really mean: How to improve search algorithms. *Ubiquity*, 2013(December):1–15.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82.

Woods, J. (2011). Recent developments in abductive logic— abductive cognition. the epistemologic and eco-cognitive dimensions of hypothetical reasoning. lorenzo magnani: Springer, heidelberg/berlin; 2009, p. 536, price£ 135 hardback, isbn 978-3-642-03630-9.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. (2023). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv:2307.02477*.

Yan, J., Wang, C., Huang, J., and Zhang, W. (2024). Do large language models understand logic or just mimick context? *arXiv:2402.12091*.

Yong, E. (2022). *An immense world: How animal senses reveal the hidden realms around us*. Knopf Canada.