Suspension of Judgment in Artificial Intelligence

Uncovering Uncertainty in Data-Based and Logic-Based Systems

Doctoral thesis for obtaining the

academic degree

Doctor of Philosophy (Dr. phil.)

submitted by

Daniela Schuster

at the

Universität Konstanz

Faculty of Humanities

Department of Philosophy

Konstanz, April 23rd 2024

Date of defense: July 11th 2024

1. Referee: Prof. Dr. Thomas Müller

2. Referee: Dr. Verena Wagner

Acknowledgments

First and foremost, I want to express my deepest gratitude to my supervisors Thomas Müller and Verena Wagner whose support could not have been greater. Thomas, thank you for being such a supportive and reliable mentor - always providing incredible feedback immediately, and leaving no stone unturned in furnishing me with valuable opportunities. Your depth of wisdom across various domains has been a constant source of inspiration throughout this journey. Verena, thank you for the many years of unconditional support, for introducing me to the world of suspension, and for managing to combine the most demanding notes with the most encouraging advice, academically, practically, and personally. I can confidently say that I would not be here without you.

I am profoundly grateful for the many chances to collaborate with brilliant scholars during my research stays. For my stay at Utrecht University, I thank my mentors Jan Broersen and Henry Prakken. For my time at the University of Maryland, Jeff Horty and Ilaria Canavotto. For my research stay at Stanford University, Krista Lawlor and Thomas Icard. Moreover, I extend my gratitude to the Studienstiftung des deutschen Volkes, the Baden-Württemberg Stiftung, Konstanzia Transition at the University of Konstanz, the Heidelberger Akademie der Wissenschaften, the Focus area Human-centered Artificial Intelligence at Utrecht University, the Doctoral Funds of the University of Konstanz, and the Krupp Foundation for their financial assistance.

I want to thank all colleagues, collaborators, advisors, fellow students, especially at the University of Konstanz, as well as my family and friends who accompanied me on this path. A special thanks to my PhD fellows and office mates Sahra Styger and Maud van Lier for all this time together, to Carl Eggen for daily check-ins, to Julia Weiss for proofreading, and to Elena Bayer for helping with the logistics during my stay at Stanford.

Most of all, thank you, Marco, for always backing me up, for challenging my views, for not accepting my insecurities, for being a role model, for pushing me to new challenges, for never even slightly doubting what I can do, and for making all of this less overwhelming.

I want to take the opportunity to last but certainly not least thank my parents and my siblings. Thank you for all your support throughout my education, starting from day one in primary school. Thank you for everything you somehow made possible and for always considering my education a priority. Thank you for your understanding without always understanding and for your slightly exaggerated belief in me. But most of all: Thank you for your pride.

Für Euch auf Deutsch: Ich möchte diese Gelegenheit nutzen, um nicht zuletzt meinen Eltern und Geschwistern zu danken. Tausend Dank für all die Unterstützung auf meinem gesamten Bildungsweg, beginnend bei meinem ersten Grundschultag. Danke für alles, was Ihr für mich möglich gemacht habt und dafür, dass Ihr meiner Bildung immer höchste Priorität eingeräumt habt. Danke für das Verständnis auch ohne immer alles zu verstehen und für Euren manchmal fast übertriebenen Glauben an mich. Aber vor allem: Danke für Euren Stolz.

Abstract

This thesis demonstrates how suspension of judgment can be integrated into systems of artificial intelligence (AI). Suspension of judgment is a crucial epistemological phenomenon that allows humans to remain neutral and to refrain from forming definitive opinions in unclear situations. A successful implementation of suspended judgment into AI systems presents a promising approach to mitigating erroneous outputs, particularly in high-stakes domains. Consequently, this research aims to analyze various AI systems to identify and enhance their ability to react neutrally when confronted with uncertain or conflicting information.

By exploring the nature of suspension and its epistemological norms, this thesis provides a philosophical analysis of fruitful implementations of neutrality in AI systems. It introduces various case studies of different AI frameworks, covering both logic-based and data-based systems, and critically assesses their current and potential capabilities to suspend judgment. The analysis reveals that while some architectures inherently possess mechanisms to communicate neutrality, others lack an appropriate capacity to respond constructively in the face of uncertain or conflicting information. As a solution, fundamental modifications are proposed for incorporating the option to suspend into existing AI architectures.

The thesis contributes significantly to the fields of epistemology, philosophy of mind, and artificial intelligence, providing a deeper understanding of the epistemic possibilities of AI systems. The findings have practical implications for the development of more robust and reliable AI systems, potentially capable of acknowledging and expressing uncertainties. Such advancements are essential for enhancing transparency, trustworthiness, and effective human-AI interaction.

Deutsche Zusammenfassung

Diese Dissertation untersucht, wie die Praxis der *Urteilsenthaltung* in Systeme der *Künstlichen Intelligenz* (KI) integriert werden kann. Urteilsenthaltung, ein zentrales erkenntnistheoretisches Konzept, erlaubt es uns, in unklaren Situationen unvoreingenommen zu bleiben und von festen Meinungen abzusehen. Die Einbindung dieses Konzepts in KI-Systeme bietet einen vielversprechenden Ansatz um Fehlentscheidungen, insbesondere in Hochrisiko-Anwendungen, zu reduzieren. Daher verfolgt diese Forschung das Ziel, die Fähigkeit verschiedener KI-Systeme, bei unsicheren oder widersprüchlichen Informationen eine neutrale Haltung einzunehmen, zu analysieren und zu verbessern.

Die vorliegende Arbeit bietet eine philosophische Analyse darüber, wie Neutralität erfolgreich in KI-Systeme eingebunden werden kann, indem sie die Natur der Urteilsenthaltung und ihre erkenntnistheoretischen Normen untersucht. Es werden verschiedene Fallstudien unterschiedlicher KI-Systeme, die sowohl logikbasierte als auch datenbasierte Systeme beinhalten, vorgestellt und ihre derzeitigen sowie potenziellen Fähigkeiten zur Urteilsenthaltung kritisch bewertet. Die Untersuchung zeigt, dass einige Architekturen von Natur aus Mechanismen zur Kommunikation von Neutralität besitzen, während es anderen an der erforderlichen Fähigkeit, angemessen auf unsichere oder widersprüchliche Informationen zu reagieren, fehlt. Um dieses Defizit zu beheben, werden grundlegende Anpassungen für die bestehenden KI-Architekturen vorgeschlagen, um die Urteilsenthaltung als Option zu integrieren.

Der Beitrag dieser Dissertation zu den Feldern der Erkenntnistheorie, der Philosophie des Geistes und der Künstlichen Intelligenz ist substanziell. Die Arbeit vertieft das Verständnis der epistemischen Fähigkeiten von KI-Systemen. Sie hat praktische Auswirkungen für die Entwicklung von robusteren und verlässlicheren KI-Systemen, welche die Fähigkeit besitzen, Unsicherheiten zu erkennen und auszudrücken. Diese Fortschritte sind entscheidend, um die Transparenz, die Glaubwürdigkeit und die Interaktion zwischen Mensch und KI zu verbessern.

Contents

Chapter 1

Introduction

1.1 Motivation

This thesis is about suspension of judgment in artificial intelligence. The term *Artificial Intelligence* (AI) refers to an artificial entity (i.e., non-human or non-biological, often a computer in common language) capable of solving tasks that typically require human intelligence. We find artificial intelligence in the most diverse forms, with the most diverse technologies, and in the most diverse areas of application whether in science, in education, in the professional environment, or in everyday life, (Luther, 2018; Kobl, 2020; Bundesministerium für Wirtschaft und Klimaschutz, 2019). The algorithms and application domains are continually advancing. Most notably, natural language understanding and processing has reached a milestone. Since November 2022, ChatGPT by OpenAI stands out as a remarkable development in *generative AI*, representing a sophisticated chatbot capable of diverse tasks, from composing poetry to coding algorithms (Hern, 2022). Moreover, computer vision is at the forefront, with AI models accurately identifying objects and actions in images and videos. Similarly, recommendation systems, as exemplified by platforms such as Netflix and Spotify, continue to refine their capabilities.

The omnipresence and extensive use of AI systems across a wide range of applications raise numerous philosophical and societal questions regarding the responsible creation and use of these systems. Many of these questions revolve around the topics of trust and responsibility. Do individuals have trust in AI? What constitutes the responsible use of AI? Is it responsible to unquestioningly rely on the results of AI systems? The underlying worry behind these questions is that one often cannot tell whether the results produced by AI systems are false, biased, discriminating, or harmful in some other way. This worry is indeed justified. AI systems make predictions and provide recommendations, inviting users to rely on their outputs. While these systems mostly provide correct and appropriate responses, instances exist where they produce erroneous, biased, or discriminatory outputs. The presence of such cases naturally creates skepticism and diminishes trust in AI systems, and we can find attempts to address this issue both on the industrial and on the governmental side. On the latter, the EU pioneered addressing this

topic via the EU Artificial Intelligence Act (European Parliament, 2023). In this risk-based approach, it is highlighted that transparency and trust in AI systems is especially crucial for those systems that entail significant *potential risks*, such as when issues related to "employment" or the "categorization of natural persons" are involved.

In this thesis, I investigate the possibilities of various AI systems to suspend judgment or exhibit neutrality in alternative manners. A primary motivation behind this project is to mitigate the occurrence of erroneous outcomes by attempting to make AI systems' responses more careful and deliberate. This objective is tackled by enhancing how AI systems manage and communicate when they are uncertain. Particularly in domains characterized by "high stakes" (like medical diagnostic), where inaccurate responses entail significant consequences, exploring how systems *process* uncertain data situations can prove exceedingly beneficial. Equally important is an examination of whether and how these systems *communicate* uncertainties to users. While there exist numerous methods to achieve this goal, with probabilistic outputs standing out as one of the most prominent, this thesis delves into the possibilities for AI systems to somehow "back out" from various types of ambiguous data situations. A central claim of this thesis is that an effective approach to appropriately *uncover* severe uncertainties involves encouraging AI systems to "suspend judgment" in unclear or critical situations, refraining from providing predefined answers and instead expressing uncertainty through a straightforward statement like "I don't know."

Suspension of judgment represents an appropriate response to situations involving missing or conflicting evidence, offering a promising possibility for a third course of action to avoid forcing critical decisions on artificial systems. Investigating the possibilities for suspension in AI systems contributes significantly to our understanding of the cognitive capacities that can be ascribed to AI systems. The exploration of which aspects of the human epistemic household can be replicated and effectively represented by AI systems is approached here from one significant perspective: the handling of uncertainties and indecision. This is of utmost importance for human epistemology, and it should be explicitly taken into account for the epistemic framework of AI systems as well.

In this regard, my research explores entirely new territory, as there have been no prior philosophical inquiries into suspension within artificial intelligence. Philosophically, investigations on suspension on the formal representation of suspension within formal philosophical frameworks can be found (see Zolfagharian, 2020), without the explicit application to AI though. Moreover, within the field of philosophy of AI, there has been a notable absence of philosophical discussions and assessments concerning AI systems that are, in some manner, capable of backing out from uncertain situations by reporting neutral responses. Within the field of AI, indecision and neutrality have not been investigated as explicit phenomena yet.

1.2 Introduction to the Topics

1.2.1 Suspension of Judgment

The first pivotal term in the title of this work, *Suspension of Judgment*, refers to a doxastic phenomenon currently under extensive investigation in epistemological discourse and related to other terms like indecision or doxastic neutrality. In political or social discussions, suspension of judgment often carries negative connotations. Individuals who refrain from forming a clear opinion on topics, such as whether the legalization of a certain drug will eliminate the black market, may face accusations of being ill-informed or of strategically concealing their true opinions.

In the philosophical discourse, indecision and suspension were historically given little explicit attention. Although these themes have roots in ancient philosophy, modern epistemology primarily focused on the accurate representation and normative profile of the propositional attitudes of belief and disbelief. Suspension was often characterized as mere non-belief, i.e., as the absence of both belief and disbelief. In this perspective, suspending concerning the drug legalization issue is simply marked by the absence of a belief that the legalization will eliminate the black market, coupled with the absence of a belief that it won't.

For a long time, there was no exploration into how to accurately represent suspension in epistemology, whether there are several distinct forms of neutrality, and what epistemic norms apply to these distinct doxastic states. However, suspension of judgment is not only often the appropriate attitude towards an issue but also an expression of active

deliberation and self-reflective behavior. When faced with contradictory evidence after a long deliberation of the drug legalization question, one may choose not to commit to a position on whether the drug legalization will eliminate the black market or not. Additionally, in situations where access to information is limited or when there is awareness of biased media sources on a given topic, the decision to refrain from forming a clear opinion can be based on the acknowledgment of scarce or inconsistent evidence.

Only recently has suspension enjoyed a renaissance in philosophical discourse and been taken seriously within the epistemological conversation as a legitimate, third option. A closer examination of the diverse forms of neutrality offers new insights for epistemology and philosophy of mind. Established models of graded beliefs (such as Bayesianism), fundamental assumptions regarding the connection between quantitative and qualitative beliefs, and foundational principles of doxastic logic are challenged by the explicit consideration of suspension.

However, there is still doubt about the utility of suspension and the actual value that a suspending attitude can contribute. It has been argued that, when one is faced with a conflict concerning a proposition, simply believing p or $\neg p$ and taking action based on this belief is better than suspending judgment, as "it is better to do something than nothing" (p. 63 Pollock, 1995). This is illustrated by the tale of "Buridan's ass," a donkey that starved to death standing between two equally tempting bales of hay because it could not decide which one to eat (p. 64 Pollock, 1995).

The story of Buridan's ass shows two misconceptions that commonly occur in the context of suspension. These misconceptions potentially misrepresent the concept of suspension but can be dispelled by means of the example of Buridan's ass. Firstly, epistemic and practical reasoning are confused here. There is a distinction between the beliefs one holds and how one carries out actions. As Pollock correctly points out, in the realm of practical reasoning, the argument for making an arbitrary choice when uncertain is valid, as captured by the tale of Buridan's ass. The donkey is well-advised to simply eat one of the balls of hay. Nonetheless, this conclusion does not imply that the donkey should also *form the belief* that one bale of hay is better than the other. It is one thing whether the donkey is convinced that the left ball of hay is better than the right, and

another thing whether the ass goes to the left ball of hay and eats it. While it is certainly appropriate for the donkey to simply eat the left one, it is epistemically irrational (and probably psychologically impossible) for the donkey to form the belief that the left ball of hay is better without having any evidence for that belief. Epistemic and practical considerations have to be separated.

If one accepts this, a second misconception in the story of Buridan's ass can still be found. If epistemic and practical considerations are separated, epistemic considerations always take precedence over practical ones, since actions typically rely on beliefs. The epistemic situation of a subject is determined first, and then the subject can be described to act based on it. Then, the question arises of how the actions should be based on the beliefs. The second misconception in Buridan's story tackles this question and involves the idea that a kind of "freezing" action must practically always follow the epistemic stance of suspension. The idea here is that if we do not know whether p, we cannot act and are *trapped* by our suspending attitude like the donkey that starved to death. This is misleading. It is an open and unresolved question what practical actions can follow a suspending attitude. It is clear, at least, that there is no set of instructions that always follows suspension; rather, the appropriate action depends heavily on the context.

Although the donkey should epistemically suspend about p, which is, whether say the left ball of hay is better, it should still act "as if p," as if the left ball of hay is preferable, and simply eat it. In this situation, not much is at stake if the donkey chooses the wrong action. In other situations, however, the best practical response to suspension may involve insisting on new information before deciding on an action. In yet other situations, waiting and collecting new information may not be suitable, but the preferred action would be whatever is the "more cautious" one, the one for which the consequences are less severe in the event of an error. For example, in the context of autonomous driving, faced with the epistemic question of whether a child is crossing the road or not, epistemically suspending might demand taking the more cautious action and braking, i.e., acting as if a child actually crosses the road.

The question of the appropriate practical response to the epistemic attitude of suspension is interesting and complex. However, it must not be

confused with the question of the appropriate epistemic reactions themselves which always take precedence and are the subject of investigation in this work. To describe the epistemic situation of a subject (whether human or artificial), it is unavoidable to engage with the third doxastic attitude, namely, suspension. Only then can we uncover the uncertainties that are present, properly describe the epistemic situation, and finally outline the possible actions based on it.

1.2.2 Data-Based and Logic-Based Artificial Intelligence

The second pivotal term of the thesis is *Artificial Intelligence*, which can be further specified as *Data-Based and Logic-Based Artificial Intelligence*. In this thesis, I aim to offer a *comprehensive* collection of the different possibilities for AI systems to suspend judgment. This does not mean that the thesis covers many different AI application areas within various industries (e.g., healthcare, finance, retail) or various sectors (e.g., security, manufacturing, marketing). The conclusions presented here are industry- and sector-agnostic and can be seamlessly applied across different domains.[1]

Rather, this objective of providing an overview is linked to the aim of incorporating a diverse range of AI frameworks and exploring algorithms with different technical compositions, structures, and underlying functionalities. The motivation behind this is the need to carefully examine the respective algorithms' compositions to reveal interesting insights regarding their capacity for suspension. To fulfill the goal of providing a comprehensive overview, I will explore both main subdomains of AI constructions: data-based AI and logic-based AI.

The distinction between data-based AI and logic-based AI has different names in the literature (Russell and Norvig, 2021). It is sometimes referred to as statistical versus rule-based AI, sub-symbolic (or non-symbolic) versus symbolic AI, or non-logistic versus logistic AI (Bringsjord and Govindarajulu, 2022). The distinctions are not necessarily always congruent. In the distinction between data-based and logic-based AI used

[1]To illustrate my results, I will employ specific application areas that are well-suited for my purposes. For this, application areas such as medical diagnostics are fitting, since they represent "high stakes" domains for which uncovering involved uncertainties seems particularly crucial.

here, I understand data-based AI as systems that use learned patterns and insights extracted from data to make decisions, while logic-based systems use logical reasoning and inference to make decisions.

Data-Based AI: What distinguishes data-based AI from logic-based systems is the fact that the model, the set of instructions on how to go from an input to an output, is not explicitly defined by the programmer but is *learned*. Therefore, data-based AI is mostly referred to with the term "machine learning" (ML). I use the term "data-based AI" to illustrate the relationship to logic-based AI, but I will refer with the term "machine learning" (ML) to the exact same set of systems. In simple terms, an ML system learns a connection between specific input parameters and specific outputs through data. It learns via examples.[2] Data-based AI comprises all "modern" AI architectures, including systems that are referred to with buzzwords like "deep learning" and "neural networks." These kinds of systems are nowadays the dominant systems in AI. In some contexts, the term artificial intelligence is even equalized with machine learning, i.e., data-based AI.

As previously defined, "artificial intelligence" refers to an artificial entity capable of solving tasks typically requiring human intelligence. Given that human intelligence encompasses various abilities, various abilities can be outlined for AI systems, too. Problem-solving strategies that are considered intelligent behavior include perception, communication, planning, and action. The development of machine learning (ML) systems has significantly advanced many of these areas. ML systems can outperform their purely logic-based counterparts in terms of speed and quality in various tasks as, for example, exemplified by the chess-playing engine AlphaZero (Silver et al., 2018). Moreover, the ML development introduces entirely new forms of intelligence, allowing artificial intelligence to tackle tasks that were previously unsolvable by logic-based systems.

A notable example is image recognition, where systems mimic the aspect of human intelligence that we would describe as seeing, perceiving, or recognizing. When an AI needs to determine what is displayed in a given image, it is not feasible to program explicit instructions that

[2] A typical process of how an ML system learns will be described in Chapter 6.

lead from individual pixels to a description. ML systems, though, can successfully learn relationships between images and descriptions from example data and apply these learned connections to new instances. Other crucial domains benefiting from ML include natural language processing, effectively enhancing the ability to communicate, and robotics, particularly known for progress in planning and action.

Logic-Based AI: In general, it can be noted that the two areas that appear in the term "logic-based AI" - "logic" and "artificial intelligence" - are fundamentally intertwined. Logical methods, logical thinking, and mathematical logic are essential for any type of programming, which forms the basis of every algorithm, in turn forming the foundation of any artificial intelligence. Without logic, artificial intelligence is not possible.

The general connection between logic and artificial intelligence is not the subject of this work, though. The term "logic-based AI" is associated with a more specific concept. At least in this work, I understand logic-based AI as a set of AI systems that exhibit intelligent behavior by making logical inferences. This means that the corresponding algorithms go from a given input or a given piece of information to an output or further information by following a sequence of set rules. These rules are determined by logical inferences based on a specific logical formalism. In this definition, logic-based AI is understood as the counterpart to data-based AI, where the rules of how to go from input to output are learned based on recognizing certain patterns in training data.

What is logic-based AI used for? Fundamentally, one could argue that logical reasoning is one crucial component of human intelligence per se. When artificial intelligence aims to replicate various aspects of human intelligence, logical reasoning becomes a significant facet for machines to emulate. However, a counterargument posits that logical reasoning can also be achieved by AI systems that are not logic-based, i.e., their internal structure is not designed explicitly for logical reasoning. Although these systems may not perform logical inferences per se, they can effectively mimic such processes, creating an appearance of engaging in logical reasoning for users. For instance, when presented with a logic puzzle, ChatGPT, a data-based system, often provides correct answers without relying on

explicit logical inference in its internal reasoning processes. Nonetheless, data-based systems can inherently produce erroneous outcomes to logical reasoning tasks due to their statistical nature.

Additionally, among the tasks that require intelligence, some specific tasks are particularly suitable for logical systems. Traditionally, these include systems meant to prove mathematical theorems, game-playing AI, expert systems, and simple chatbots. However, also within those tasks, logic-based AI faces limitations, especially when the complexity of the task is increasing. The most well-known example is the game of chess. Chess is a rule-based game in which logical reasoning seems essential. Nevertheless, the data-based system AlphaZero from DeepMind has defeated logic-based chess computers like Stockfish or Deep Blue which famously defeated the World Chess Champion Garry Kasparov in 1997 (BBC-News, 2017), highlighting the power of learning from data for complex tasks.

While it is possible that purely logic-based AI systems may become less prevalent in practical applications, the study of logic-based artificial intelligence remains vital, especially in the context of so-called *hybrid* or *neural-symbolic* AI systems. These systems are characterized by their fusion of elements from both logic-based and data-based AI, effectively combining reasoning and learning (Marcus, 2020; Besold et al., 2022). That the synergy between logic and machine learning can be efficient, was noticed already by pioneers like Alan Turing (Turing, 1950) and John McCarthy (McCarthy, 1959), both of whom already took "a combination of Logic and Learning as being central to the development of Artificial Intelligence research" (Muggleton and Marginean, 2000, p. 316). The concept of hybrid AI is not a new one (Marcus et al., 1992; Sun, 1996; D'Avila Garcez et al., 2009) but has gathered increasing attention in recent years, as exemplified in discussions like the debate between Gary Marcus and Yoshua Bengio about "the best way forward for AI" in Montreal 2019 (Marcus, 2020; Bengio and Marcus, 2021). While data-based machine learning is often referred to as the "second wave" of AI that marked a paradigm shift, many experts argue that hybrid, neural-symbolic AI represents the third wave of AI (D'Avila Garcez and Lamb, 2023). This is particularly pertinent considering the need to enhance the safety, robustness, trustworthiness,

and interpretability of contemporary AI systems. Notable examples of such hybrid systems, including AlphaGo and Google Search, demonstrate the potential of combining structured, logical reasoning with learning algorithms (Marcus, 2020).

This thesis delves into fundamental questions regarding the possibility of doxastic neutrality in both pure forms of artificial intelligence: data-based and logic-based systems. By thereby investigating both main components of hybrid systems, the thesis indirectly manages to address hybrid systems, too.

1.3 Scope of the Thesis

It is important to reiterate that the scope of this thesis is *purely epistemological*. As previously clarified, a strict differentiation is maintained between practical and epistemological reasoning. The goal is to investigate the epistemic household of AI systems and in how far we can make sense of our epistemic concepts, in particular suspension, when applying them to AI systems. Therefore, only what can be the object of epistemological investigation is eligible as an object of investigation for this thesis: knowledge, beliefs, suspension in the form of propositions or statements.

When exploring the possibility of suspension in data-based AI systems, the focus will be restricted to so-called "abstaining machine learning" systems. Firstly, this restriction is motivated by the limitation of the scope to epistemological considerations. The investigations in data-based AI are thus limited to those ML systems whose task is to answer questions, express opinions, make predictions, or otherwise make statements. These forms of AI tasks can be found in a wide variety of application areas. A translation system or an image recognition system can be described as "making statements" just as well as a recommendation system or a system that helps with cancer diagnostics. Negatively formulated: What is not in the scope of the thesis are all forms of ML that cannot be described as making statements purely. This is any form of ML that can be considered as involving actions, bodily movements, or any other kind

of physical involvement with the environment. Most prominently, the whole area of robotics will be excluded from the considerations of this thesis.

However, even within the realm of predicting ML systems the scope is further restricted. There is a multitude of systems that potentially qualify for being in scope for investigating how uncertainty can be represented. In the domain of ML, there are indeed numerous ways through which the systems can communicate their own uncertainty. Many ML systems compute soft probabilities for the different answer choices, which are then usually converted (via choosing the answer with the highest probability) into definite answers (hard probabilities) (Campagner et al., 2019). A more informative answer that respects uncertainties could then be generated when systems do not (only) output the unique definite answer, but the corresponding probability distribution over the possible answers (Ferri and Hernández-Orallo, 2004). An image recognition system could, for example, not only output "This is a dog", but "This is 65% a dog, 13% a cat, ..." Other possibilities consist in outputting the sets of likely answers or outputting intervals instead of precise points for when the answer is continuous. The usefulness of these other approaches compared to abstaining ML systems certainly depends on the specific application area and user profile and will not be discussed in more detail here.[3] This work deals exclusively with so-called abstaining ML (AML) systems, i.e., systems that in some way have the possibility to output a neutral answer, in saying something like "I don't know which object is displayed in the picture." This second restriction is motivated by the objective of this work to consider and emphasize AI systems that potentially qualify as suspending judgment. AML systems that are able to output one categorical, neutral output, are certainly the best candidate for these investigations.

The scope for logic-based systems will be on two frameworks that both fall under the category of "non-monotonic reasoning."

When logic is used to create artificial intelligence, one of the early demands placed on it is that the logic should be able to capture

[3]Certainly, from a philosophical point of view, there are also interesting epistemological questions that arise with respect to such probabilistic systems, especially in the area of formal, gradual epistemology.

common-sense reasoning of humans. For example, Minsky (1975) noted that human reasoning often involves making certain conclusions even when the inferences are not guaranteed, with the option to discard those conclusions when new information comes into play (Minker, 2000). Common-sense reasoning is defeasible. As classical logic cannot represent this behavior, non-monotonic reasoning often forms the basis for logic-based AI. Non-monotonic reasoning characterizes a set of logical systems that lack monotonicity. Monotonicity ensures that conclusions drawn from a set of premises remain valid even with the addition of new premises. In contrast, non-monotonic systems allow for the possibility of rejecting previously accepted conclusions when new premises are introduced. A widely known example is the case of Tweety the bird. If we only have the information that Tweety is a bird, we will conclude that Tweety can fly. However, if we receive additional information that Tweety is a penguin, we retract the previous inference. The three most prominent logical frameworks for non-monotonic reasoning are autoepistemic logic, circumscription logic, and default logic (Minker, 2000).

For my investigations, I will focus on Default Logic first. Default logic effectively captures the fundamental concept of non-monotonic reasoning, highlighting that the introduction of new information can change the outcomes of an inference. Default logic describes reasoning rules, so-called defaults, which possess two antecedents: a regular antecedent and an exception condition. The idea is that, first, we reason by default. Normally (by default), we can conclude certain things. However, when we get new information (this information is captured by the exceptions of the default), we have to withdraw from our previously made inference. Thereby, default logic manages to encompass common-sense, everyday reasoning as noted in Delgrande and Schaub (2000).

Moreover, the scope of this thesis will be limited to so-called *normal default theories*. With normal default theories, where the exception condition is the negation of the conclusion, we have a straightforward method for representing non-monotonic rules. In particular, the framework for constructing normal default theories outlined by Horty (2011) provides a contemporary and intuitive foundation for the forthcoming investigations on suspension.

As also described by Horty (2011, p. 8), in addition to the standard non-monotonic formalisms, there is a parallel related but different influential tradition within this field that focuses on arguments rather than on propositions: abstract argumentation theory (Dung, 1995). This framework constitutes the second logic-based framework that will be in scope of this thesis. Abstract argumentation theory centers on modeling arguments and their interrelations. In this context, it operates on a different level than formalisms like default logic, as it does not involve representing the internal structure of arguments. Rather than providing rules on when to draw specific conclusions, its objective is to determine which of the presented arguments should be accepted. It illustrates significant characteristics of non-monotonic reasoning, dealing with incomplete information and being based on the idea that additional information, in the form of extra arguments, can potentially lead to conflicts and the retraction of previously accepted arguments.

Considering the integration of suspension within the abstract argumentation theory frameworks allows me to explore what it means to suspend judgment (or remain undecided) regarding an argument or its acceptability. Moreover, findings derived at the abstract argument level can be transferred to propositions using structured argumentation approaches. This transfer facilitates a comparative analysis between the outcomes of default logic and argumentation theory.

Preceding the examination of the frameworks of default logic and argumentation theory, I delve into a specific case study of suspension in non-monotonic reasoning, investigating the phenomenon of floating conclusion. This phenomenon in non-monotonic reasoning exemplifies a prototypical case of suspension. In the scenarios under consideration, a conflict leads to suspension about the conflicting propositions. Then, additional conclusions can be drawn from the suspended propositions. In such cases, the general question of how to deal with conclusions derived from suspended propositions arises. In the specific case of floating conclusions, a proposition (the floating conclusion) follows from *both* conflicting propositions that are suspended.

The considerations concerning floating conclusions aim to provide an

initial understanding of the role of suspension in non-monotonic reasoning systems and highlight the potential stumbling blocks associated with the incorporation of suspension into these systems. Given that floating conclusions exemplify a typical case for suspension in non-monotonic reasoning (independently of the specific framework), these considerations act as a precursor to the subsequent examinations of default logic and argumentation theory.

The investigations of this thesis on the possibility of suspension in different AI systems reveal varying capabilities within each framework, both in terms of existing features and potential adaptations. Consequently, the results obtained from each framework differ. However, a unified summary of the respective results from the different frameworks will be provided based on three key questions, answered at the end of each of the corresponding chapters:

1. Does the considered framework allow for a way to deal with conflicting or uncertain information?

2. Is there something in the light of suspension of judgment present in the framework?

3. Can we find and distinguish different forms and epistemological norms of doxastic neutrality in the framework?

The diverse responses to the questions in the individual chapters highlight that the various frameworks are at different stages in their journey towards representing suspension.

1.4 Structure of the Thesis

The thesis is structured into seven chapters: the introduction (Chapter 1), five primary content chapters (Chapter 3 – Chapter 6), and the conclusion (Chapter 7). Chapter 2 is dedicated to outlining the philosophical background of the phenomenon of suspension of judgment. It draws upon significant insights from the philosophical domains of epistemology and philosophy of mind. In essence, this chapter establishes the theoretical groundwork for the thesis from a philosophical perspective. The first part

(Subsection 2.2) provides an overview of studies on epistemological norms related to suspension, while the second part (Subsection 2.3) summarizes theories in the philosophy of mind concerning the nature of suspension and related concepts. The third part (Subsection 2.4) covers two additional topics, one about different forms of suspension and one about suspension and its relation to inquiry, which are both situated in an overlapping region between epistemology and philosophy of mind.

The subsequent four chapters delve into the exploration of suspension in artificial intelligence. Chapter 3 serves as a case study, investigating the logical behavior of suspension in AI frameworks through the phenomenon of floating conclusions. This chapter provides valuable insights into the logic of suspension in AI systems by examining a prototypical situation for suspension, offering a preliminary understanding of one of the various modes in which suspension can manifest. The following primary content chapters, Chapter 4, 5, and 6, examine diverse AI architectures and explore how suspension manifests in these frameworks. In Chapter 4, I will investigate the framework of default logic and propose a constructive approach to adapt it. This adaptation aims to enable the representation of suspension within the existing framework. Chapter 5 explores argumentation theory, focusing on interpreting and elaborating state-of-the-art methods for representing suspension while identifying missing aspects and areas that require further development. Chapter 6 is the final primary content chapter, shifting the focus to data-based machine learning. Here, I will elaborate in detail how ML models operate and explore various methods for incorporating neutrality. I will portray the systems found in the literature on abstaining machine learning and organize them in a manner that facilitates the application of philosophical considerations on suspension. Additionally, I will address crucial topics such as autonomy and explainability within abstaining systems. Chapter 7 will serve as the conclusion, summarizing the findings and implications presented throughout the thesis and outlining promising paths for future research.

Chapter 2

Suspension

Contents

2.1 Introduction

Suspension of judgment is often considered the third doxastic stance that represents indecision and that is opposite to its two counterparts, belief and disbelief. Alternatively, it can be characterized as a response to a question that signifies neutrality (or indecision) about the answers to the question. In philosophical consideration, suspension has long been overlooked. Epistemology has primarily focused on the correct representation and norms concerning the propositional attitudes of belief and disbelief, characterizing suspension merely as the absence of both attitudes, i.e., *non-belief*. In the context of the Drug-Legalization example introduced in Chapter 1, if a person A suspends judgment on whether a certain drug legalization will eliminate the black market, suspension is, in this perspective, simply characterized by the absence of their belief that it will and the absence of their belief that it will not. For a long time, there was no investigation into how suspension can be adequately represented, whether there are different forms of it, or what epistemological norms apply to it.

This handling is insufficient. To suspend judgment is not only frequently the *appropriate* response to a topic but also often involves active deliberation and self-reflective behavior. If A has engaged with the question about the drug legalization for a long time, they may find themselves confronted with such contradictory evidence that they do not want to commit to whether the legalization of this particular drug will eliminate the black market or not. In another evidential situation, A may be aware that their access to information on this issue is very limited or that the media sources available to them are highly biased. Therefore, A could, due to limited or one-sided information, refrain from forming a definite opinion. This awareness of the scarcity or inconsistency of evidence involves a higher-order position that is cognitively often more demanding than simply forming a belief on a specific topic.

Rather recently, suspension has been taken seriously within the epistemological debate. Most prominently, since Jane Friedman published a series of papers (Friedman, 2013a,b,c) in which she places suspension at the center of investigation, many debates have arisen concerning the

accurate description of the phenomenon and the norms of rationality with respect to suspension. A closer examination of suspension provides new insights for epistemology and philosophy of mind. For example, we will, observe that when seriously considering suspension and attempting to integrate it into formal epistemological frameworks, it raises new questions concerning the representation of graded beliefs (e.g., in Bayesianism), the foundational principles of doxastic logic that govern categorical beliefs, and the relationship between graded and categorical beliefs. For instance, it remains unclear which credences (or values in general) in a model of quantitative belief could align with a state of suspension.

In this chapter, I will examine these debates. The chapter provides the theoretical background of the philosophical concept of suspension of judgment for my thesis. To do so, I will look at the phenomenon of suspension of judgment from two points of view. One is the epistemological point of view; the other is the point of view of philosophy of mind. The epistemological view focuses on the normative profile of suspension of judgment. I will distinguish different norms for suspension by considering situations in which suspension of judgment is an appropriate doxastic attitude. The studies from philosophy of mind focus on the nature of suspension of judgment and describe the psychological and phenomenological profile of this attitude. Here, I will distinguish suspension from other forms of neutrality, defining it in contrast to terms such as ignorance. I will present a spectrum of doxastic neutrality, with suspension representing the end of the spectrum. Furthermore, I will explore various accounts that seek to describe the nature of suspension.

We will see that it is not always possible to separate these two perspectives completely, i.e., to treat the different reasons[1] for suspension and the different psychological forms of neutrality independently from one another. Often a particular normative reason for suspension is tied closely to some form of neutrality, or some form of neutrality is in tension with some particular normative reason. Nevertheless, there is no clear correspondence between the different reasons and the different forms. Hence, despite the overlap between the two perspectives, I will structure this chapter alongside

[1]Unless stated otherwise, I will use the term "reason" as a *normative* reason, meaning a justificational reason, rather than a motivational or explanatory reason.

the two, presenting first the norms of suspension that epistemologists have described and secondly the nature of suspension. Additionally, at the end of this chapter, I will present two debates that do not neatly fit into either epistemological or philosophy of mind considerations. The first of these debates (Subsection 2.4.1) centers around various forms of suspension characterized by their normative profile. The second debate (Subsection 2.4.2) shifts the interpretation of suspension, deviating from the primary interpretation used in this chapter. This debate presents a different facet of suspension, particularly in its relation to inquiry, accompanied by specific norms. Those two topics, overlapping between epistemology and philosophy of mind, constitute the third main section (Section 2.4).

2.2 Epistemology: Norms of Suspension

Epistemology is the philosophical discipline that deals with epistemic and doxastic stances and their rationality. In epistemology, normative questions are asked about when a certain stance towards the truth of a proposition or the answer to a question is rational. Often this question has been asked in relation to taking a proposition to be true (believing) and taking a proposition to be false (disbelieving). Of course, it can also be asked in which situations a neutral doxastic stance, in particular suspending, is rational. As I will unravel the various variants of doxastic neutrality only in Section 2.3, I will exclusively use the term "suspension" within this section and discuss the rationality norms associated with suspension.

Questions of rationality are often also formulated as questions about the *justification* for a certain stance or the *reasons* or *evidence* for adopting that stance. Thus, in the light of an evidentialist picture, being justified in believing a proposition is equivalent to having evidence that supports that proposition (Feldman and Conee, 1985; Conee and Feldman, 2004; Owens, 2002).[2] In contrast, proponents of pragmatic encroachment argue that pragmatic reasons also play a role in the justification of epistemic stances (Fantl and McGrath, 2002; Armendt, 2010; Kim, 2017). According to this view, how important answering the question about p is, or how severe the consequences of being false are can influence which doxastic stance I should

[2]Being justified in believing should not be confused with having a justified belief. See the distinction between propositional and doxastic justification (Rosenkranz, 2018).

take about p.

The question about the influence of non-epistemic reasons for the justification of a doxastic stance has been freshly emphasized after scholars considered the normative profile and the nature of suspension more closely (Wagner, 2022; McGrath, 2021).

In the following chapters of this thesis (Chapters 3-6), I will apply different epistemological norms for suspension, which I will introduce in the following, to different frameworks of artificial reasoning. In doing so, I will focus on describing the evidential situation of these systems. Hence, in the following explanations about epistemology, I will focus on the evidentialist picture and ask about the evidence-based justifications and the evidential norms for suspension. Nevertheless, it might be interesting, too, to consider what a pragmatic reason for an artificial reasoner could consist of. Possibly, this could be something like the importance of the proposition for further query or the computational cost of gathering more information.

2.2.1 Privative and Positive Justification

Suspension offers a more complex normative profile than belief and disbelief do. While belief and disbelief can only be justified *positively*, suspension can, according to Zinke (2021b), be justified in two ways: *positively* and *privatively*.

A justification or a reason for a belief in p is always some sort of positive evidence for p. One can be positively justified in believing that a certain dog is a Husky, because the dog is black and white or because the dog has blue eyes. All these facts provide us with positive evidence for believing p. There is some piece of evidence or information (the black-and-white color of the dog) that speaks *for* believing p in a positive way. Likewise, I am positively justified in disbelieving that it will snow today because I have positive evidence for the belief that it will not snow today (e.g., the fact that it is currently 30 degrees Celsius and sunny).

In some cases, we are positively justified to suspend, too. Classic examples are cases of vagueness (Ferrari and Incurvati, 2022) or chance (Feldman and Conee, 2018; Zinke, 2021b). We might suspend about the proposition "this cup is blue" because the cup is a borderline case between

being blue and green. The cup being a borderline case of a blue cup, hence, provides us with positive evidence for suspending about whether the cup is blue. In a different scenario, we might suspend regarding the proposition "This is the winning lottery ticket" for a particular lottery ticket, or "The coin lands on heads" especially before the lottery was drawn or the coin was flipped. We suspend because we hold the belief that the lottery and the coin are fair, and success is solely determined by chance (for instance, with odds of 50:50 in the case of the coin). The belief that the lottery is a fair process is a positive reason for me to suspend about whether I will win and the belief that the coin-flipping is a fair chance process provides me with a positive reason for suspending about the coin landing on heads. The prototypical cases of positive justification of suspension are cases in which the neutrality is in some sense justified by the constitution of the proposition itself (and the facts about the world related to p) rather than by the (deficient) epistemic circumstances of the reasoner.

Additional instances of positive justification involve situations of higher-order evidence (Christensen, 2010). Higher-order evidence is not directed towards the proposition itself but rather towards the cognitive processes that lead to the formation of a doxastic stance towards the proposition. This type of evidence can serve as a direct basis for justifying a suspending attitude. The most used examples are cases where the reasoning subject initially acquires first-order evidence in favor of (or against) a particular proposition p, only to later discover that they were influenced by some substance that impairs[3] (or completely eliminates) their cognitive faculties essential for assessing the truth of p. A more realistic example, as described by Wagner (2021), could involve a person answering questions on a test, being confident in the correctness of their answers based on positive (first-order) evidence for each choice. However, they are later informed by the system that their answers were not entirely correct, introducing higher-order evidence that justifies suspending judgment about the answers to the questions.[4] Other typical examples include cases of impaired perception. Pollock (1995) famously argued that when a person sees a

[3]Note that in the discussion, scholars contend that the substance *may* impair cognitive abilities.

[4]In the example of Wagner (2021), the case is slightly more complicated, as the person is told that exactly one of the answers is incorrect, but they are equally sure of each of their answers. Additionally, there is some pressure to act.

red object but has higher-order information that the lighting in the room makes everything look red, they suspend judgment about the proposition p that the object really is red. In such scenarios, this higher-order evidence directly justifies suspending judgment regarding the proposition p, as it undermines the earlier evidence for or against p.

In the most frequent suspension cases though, suspension is justified in a different way. Usually, we do not suspend because we have positive evidence *for* suspension, but because we do not have enough positive evidence for believing or disbelieving. In an evidentialist picture, one could say that "suspension of judgment is the justified attitude when the person's evidence on balance supports neither a proposition nor its negation" (Feldman and Conee, 2018, p. 75). In such situations, suspension functions as a fallback position that we are justified to take if we are not justified in any other doxastic attitude. We are then justified privatively. This function as a rational fallback option is what distinguishes the normative profile of suspension from the one of belief and disbelief.

Two very basic norms discussed in the philosophical literature on suspension, both falling under privative justification, are the *Absence of Evidence Norm* and the *Balanced Evidence Norm*. Throughout the thesis, I will call those norms the *Absence Norm* and the *Balance Norm*. In some cases, we might just have no or barely any evidence for or against a proposition p. In other cases, we might have positive, first-order evidence both for belief and for disbelief, but the evidence is (almost) equally balanced. Both mentioned norms pertain to these situations in which a subject is privatively justified in suspending. The Absence Norm states that one should suspend[5] about a proposition p, when there is no evidence speaking in favor or against p (Friedman, 2013b, p. 60). The Balance Norm states that one should suspend about p, when the evidence for and against p is (more or less) equally balanced, see for example Schroeder (2012).[6] For example, we might have evidence for believing

[5]Plausibly any neutral doxastic position towards p and not only (qualified) suspension should be permissible. Full ignorance towards p should be permissible, too, if there is no evidence at all. For this distinction see Section 2.3.

[6]The evidence does not have to be perfectly equally balanced. How much overweight of evidence for one side is still permissible for the norm to apply might be dependent

the proposition "There is a Husky on the image," because the dog in the image seems black-and-white. We might also have evidence for disbelieving the proposition because the dog seems not to have blue eyes. Surely, the Absence Norm can be treated as a special case of the Balance Norm. When there is no evidence, the evidence can be viewed as balanced. To uphold the difference, I will, in the following, assume that the Balance Norm only covers cases where there is a minimum amount of evidence to start with.

To sum up, we see that epistemologists describe at least two different kinds of reasons for suspension. One type consists of reasons that positively speak for suspension, the other type covers cases when one has privative reasons for suspension, because one lacks enough reason for either believing or disbelieving. Two fundamental norms for suspension, the Balance Norm and the Absence Norm fall under the second type of reason.

The different justifications and norms can be displayed in the following figure.

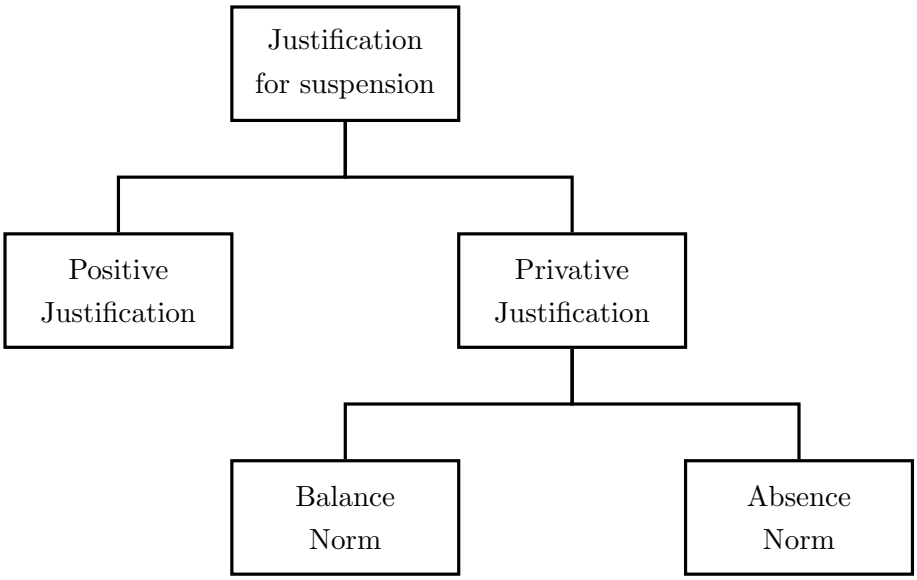


Figure 2.1: Different justifications for suspension.

on factors like the total amount of evidence or the stakes for deciding about the truth of p, which supervene again on non-epistemic factors that proponents of pragmatic encroachment would take into account.

2.2.2 The Logic of Suspension

Epistemological considerations about when a doxastic stance is justified or rational are characterized not only by norms that describe the specific situations in which the respective stance is permissible, mandated, or forbidden; additionally, they can involve an examination of the logical profile of the stance. In such an examination, one investigates the appropriate behavior of epistemic or doxastic attitudes within arguments. In particular, formal epistemology focuses on questions of rationality concerning the logical relationship between different beliefs or their propositional content: If one believes a certain set of propositions, should one rationally believe another proposition? If the set of beliefs is expanded or reduced by a proposition, what are the normative consequences for the remaining propositions? Specifically, it is asked how the doxastic attitude towards propositions can be transferred to other, logically related propositions. If I believe both p and q, should I also believe $p \wedge q$? With questions of this kind, we engage in epistemic and doxastic logic, a subfield of formal epistemology.

When investigating the epistemology of suspension, new challenges for epistemic logic arise. As I will describe in the following, simple principles like logical closure, which appear suitable for belief (at least under ideal circumstances), seem to fail in the case of suspension.[7] Moreover, when confronted with an argument for which we suspend judgment regarding its premises, it is not immediately clear which doxastic stance is rationally demanded concerning the conclusion. This is illustrated in the cases of floating conclusions that I will examine in detail in Chapter 3.

In Zolfagharian (2020), for instance, it has been argued that logical connectives such as conjunction and disjunction are not clearly defined when suspension comes into play. Similarly, regarding degrees of belief, Friedman (2013b) argued that Bayesianism, which many scholars consider

[7]For example, an ideal reasoner, adhering to logical closure, would be rationally obligated to believe any disjunction involving the proposition that it is sunny, if they already believe that it is sunny. For example, they should believe the disjunction that it is sunny or rainy. On the contrary, a reasoner who suspends judgment about whether it is sunny cannot be rationally obligated to suspend judgment about the disjunction that it is sunny or rainy. The reasoner could rationally believe the disjunction while simultaneously suspending judgment about both disjuncts.

to be an appropriate framework for describing degrees of belief, fails to adequately represent the concept of suspension of judgment.

Given that the reasons for suspension (and its various forms, as will be elaborated in Subsection 2.3.2) are diverse, logical principles and formal frameworks are applicable in some scenarios while they fail in others. Different forms or reasons for suspension are accompanied by different logical profiles, as also noted by Zinke (2021a).

For instance, when we encounter suspension due to chance, Bayesianism appears to model this reasonably well. The degree of belief in the proposition that a fair coin will land on heads is appropriately described by the value 0.5 which aligns with the objective chance of the event, as advocated by Lewis' principal principle (Lewis, 1980). According to the *Lockean Thesis* (see e.g. Foley, 1992), which aims to transform degrees of belief into categorical doxastic attitudes, a degree of belief of 0.5 could be interpreted as a suspending attitude. The rules of Bayesianism also correctly represent that the degree of belief in a disjunction, such as "the coin will land on heads or tails," should be the sum of the degrees of the two disjuncts, in this case, 1.0. However, closure under conjunction immediately fails for suspension due to chance. If I suspend judgment about whether a coin landed on heads and also suspend judgment about whether it landed on tails, it does not follow that I should suspend judgment about the proposition "the coin landed on heads and on tails." Rationally, I should disbelieve this contradicting conjunction.[8]

Conversely, in other cases of suspension, closure under conjunction often seems appropriate. If I suspend judgment regarding whether it is rainy in Berlin and also suspend judgment about whether it is rainy in New York, it is rational to suspend judgment whether it is rainy in both Berlin and New York (the conjunction). However, in this case, it appears less clear that the rules of Bayesianism (and probability theory) hold, as demonstrated

[8]There are numerous adaptations of this example that do not involve a straight contradiction. For instance, one can suspend judgment about whether toss 1 will be heads, whether toss 2 will be heads, and so on, but disbelieve the conjunction that all tosses 1-50 will be heads, as the probability for this event is only 0.5^{50}. Moreover, the issue of closure under conjunction is not restricted to chance processes. This is particularly obvious in the context of answers to polar questions that are not independent of each other. Analogous to the case of the coin, when I, for example, suspend judgment about "Christ did go to the Party" and about "Christ did not go to the Party," it seems inappropriate to suspend judgment about the conjunction of these propositions.

by Friedman (2013b). Friedman (2013b) challenges the Lockean Thesis, which asserts that belief, disbelief, and possibly suspension can be reduced to certain credences. She contends that, when dealing with a substantial number of independent propositions p_i an individual suspends judgment upon, the laws of Bayesianism, combined with the Lockean Thesis, fail to accurately represent the individual's credence in the conjunction or disjunction of the propositions. For instance, if an individual suspends judgment about whether it is rainy in New York and holds a middling credence of 0.5 for the proposition, and also suspends judgment about whether it is rainy in Berlin with a credence of 0.5 for this proposition, the credence in the proposition "it is rainy in New York and in Berlin" will be only 0.25, even though the subject still suspends judgment. As Friedman (2013b) increases the number of independent propositions, the credence of a long conjunction of independent propositions becomes infinitesimally small and close to 0 (if there are infinitely many p_i). According to this argument, suspension would encompass the entire spectrum of credences, leaving no room for belief or disbelief.

Considering this lack of well-defined rules, Rosa (2019) also investigates different logical principles for suspension. He argues against the general logical principle that posits if $\phi_1, \ldots, \phi_n \models \psi$, then if I suspend judgment about $\phi_1, \ldots, \phi_n$, I should have reason not to believe $\neg\psi$. He deems this principle incorrect since ψ could be false for other reasons, such as being contradictory (as in the example above where ψ would be the proposition that the coin lands both on heads and on tails). Even this seemingly mild principle fails clearly.

The two principles he argues for are in fact negative principles for suspension. One states that if $\phi_1, \ldots, \phi_n$ entail ψ, and you believe all of ϕ_i, then you *should not suspend judgment* about ψ. However, in my view, this principle is more a principle about belief than about suspension, as the logic of belief suggests that in this case, you should believe ψ. The only additional step here is to assert that believing is incompatible with suspending judgment, which is at least from a normative perspective widely accepted. A principle Rosa (2019) considers valid is the one that states if $\phi_1, \ldots, \phi_n$ entail ψ, then you have reason not to believe all of $\phi_1, \ldots, \phi_{n-1}$, suspend judgment about ϕ_n, and disbelieve ψ. He argues that of the stricter

principles concerning suspension he investigates, all that demand a specific doxastic attitude ultimately fail. In another paper, Rosa (2021) attempts to capture the most fundamental rationality principles for suspension. In this context, we encounter rationality requirements like the following:

$$S(p), S(q), \neg B(\neg(p \wedge q)) \models S(p \wedge q).$$

$S(p)$ is to be read as "p is suspended" and $B(p)$ as "p is believed." This principle allows exactly for the exceptions described earlier. In many cases, suspension is not closed under conjunction because it seems more reasonable to disbelieve the conjunct. Nevertheless, the principle of Rosa (2021) allows us to at least exclude one possibility: When we suspend judgment about two propositions, we should not believe their conjunction. (Conversely, for disjunction, it seems that if we suspend judgment about two propositions, we should at least not disbelieve the disjunction.)

In conclusion, there are only a few limited logical principles that are firmly established for suspension. For instance, it seems unquestionable that suspension should be closed under negation. In other words, if I suspend judgment about p, I should also suspend judgment about $\neg p$. However, the general logical behavior of suspension is an interesting and unanswered question. Since I will also investigate logical frameworks in the following chapters, especially in the chapters about default logic (Subsection 4.3.3) and floating conclusions (Section 3.4), I will provide a partial answer to this question.

2.3 Philosophy of Mind: Nature of Suspension

This section primarily focuses on describing the nature of suspension of judgment. Its structure is more complex compared to the previous section, given the number of different debates and variations in terminology in the literature. The section comprises four subsections. Before I start to describe the nature of suspension, in the first subsection (Subsection 2.3.1), I will delve into a fundamental question about what serves as the *object* of suspension. Following this, the two subsequent subsections will aim to unravel the nature of suspension. In Subsection 2.3.2, I will position suspension within the broader context of doxastic neutrality by

distinguishing it from other forms of doxastic neutrality. In Subsection 2.3.3, I will present accounts from various scholars attempting to define suspension as such.

2.3.1 The Object of Suspension

Before exploring the nature of the mental state of suspension, it is important to clarify what the object of suspension is. Recall that suspension is commonly identified as the third doxastic stance, alongside belief and disbelief. Consequently, it has conventionally been assumed that the object of suspension aligns with that of belief and disbelief, namely a proposition. In this picture, given a proposition p, one can adopt one of three distinct doxastic stances: believing p, disbelieving p, or suspending judgment about p.

However, the very phrasing of this description already suggests that the object of suspension is at least linguistically distinct from the object of belief and disbelief. The linguistic relationship between suspension and its object appears to be different from the relationship between belief and its object. We cannot accurately state that a subject suspends *that* white mice exist in the same manner as a subject believes that white mice exist. Instead, we say that the subject suspends about questions like *whether* white mice exist or about *when* a certain train departs. Thus, as argued by Friedman (2013a), suspension seems to directly apply to a question, rather than to propositions.

The different objects of belief and suspension can be reconciled if we consider propositions as answers to *questions that are under discussion*, i.e., considered in a certain context.[9] The questions under discussion (QUD) that I discuss in this thesis will encompass a set of well-defined, complete answers A.[10] Ferrari and Incurvati (2022) then, for instance, categorize belief and disbelief as *gnostic* attitudes, while considering

[9] As Ferrari and Incurvati (2022) adopts the term "question under discussion" from Roberts (1996), it is predominantly used in contexts involving multiple interlocutors who align on a common goal by accepting a question as under discussion. My considerations are limited to one single subject. Still, I employ the term "question under discussion" (or QUD) to *fix* a specific question I wish to be seen as the object of epistemic consideration for the moment, occasionally also to differentiate it from other potential questions within the context.

[10] In the subsequent chapters, I will at times refer to the members of this set as "defined answers."

suspension as an *agnostic* attitude one can adopt towards the complete answers of a question, which are expressed by *propositions* $p_1, \ldots, p_n$ in the answer set A.

In the simplest scenario involving propositional, polar questions, only two complete answers exist — one affirmative and one negative.[11] Belief and disbelief are characterized by taking one of the complete answers to a question, in this case the affirmative or the negative answer, to be true. Suspension (as well as other forms of doxastic neutrality as we will see in the following subsections) emerges from not choosing or not committing to the truth of any of the defined or complete answers, i.e., being neutral towards the complete answers. In this vein, suspension is not defined by providing an explicit answer to the question. Instead, it is characterized as an alternative form of responding to or addressing the question that is characterized by its neutrality.[12]

Friedman (2017) builds upon this idea and offers a unified terminology, not only expressing that a subject suspends towards a question Q, but also expressing that a subject *knows* Q, which means that the question is answered and closed. As Archer (2018) correctly points out, the terminology of "knowing a question" can be potentially misleading.[13] In such cases, it is often more precise to resort to propositions and use the expressions of "believing or knowing a proposition p" and the supplementary construction of "suspending *about* p."

Throughout this thesis, I will also employ this construction and often refer to suspension *about* p, whenever suspension is compared to belief or disbelief about a certain proposition and the respective framework suggests a propositional terminology. At other times, the question under discussion itself will be more relevant, such that I will use formulations like "suspending about Q" or "suspending with respect to Q.

[11] In this case, transitioning from questions to propositions is straightforward. The traditional framework, wherein disbelief corresponds to the belief in the negation of the proposition, encounters challenges only when considering non-polar questions.

[12] To differentiate between addressing a question with a full answer and addressing it in another way (for example, with neutrality or with the denial of the presupposition), I will use the term "response" as a broad term for any form of addressing the question and "answer" as addressing the question with a defined, complete answer, see Wagner (2023b).

[13] Unlike suspension, Friedman does not consider knowing to be an "interrogative attitude" as elaborated in Section 2.4.2.

2.3.2 Suspension and General Doxastic Neutrality

When studying doxastic positions, recent philosophy has redirected its attention from just belief and disbelief (or graded versions of belief or disbelief) to also considering the other doxastic positions one might take towards a proposition. In this thesis, I am concerned with those other doxastic positions towards a proposition p that involve and are substantially characterized by being neutral about the truth of p.

The focus of this work is on suspension of judgment, which I understand as a form of doxastic neutrality. While I consistently refer to suspension as a type of doxastic neutrality, it is important to recognize that there are other forms of doxastic neutrality that warrant distinction. Therefore, I will begin by broadening the perspective and examining doxastic neutrality and its various manifestations.

Two questions arise regarding the distinctions I propose. Firstly, there is the question of the different phenomena of doxastic neutrality. It is widely accepted that doxastic neutrality can manifest in diverse ways. However, there is ongoing debate over whether these variations represent merely gradual differences or if they are categorically distinct, and how these manifestations can be grouped into different subtypes of doxastic neutrality. Secondly, there is a lack of consensus on the terminology. Various terms, such as "suspension," "agnosticism," "indecision," "ignorance," "non-belief," "withholding," "and withdrawing," are used to describe doxastic neutrality, without agreement on which term refers to which phenomenon or grouping of phenomena.

In this work, I aim to establish groupings and use terminology that aligns reasonably well with the conventional usage (as seen in scholarly discourse and everyday language), but also provides a solid foundation for the examination of AI systems. Regarding the terminology question, I will maintain flexibility in my use of different terms and adapt my terminology based on the context. As I explore various formal frameworks and engage in different debates, each with its own established vocabulary, I will adhere to the terminology most commonly used within each specific framework. For instance, in the realm of argumentation theory, the term "indecision" is prevalent, while in machine learning, "abstention" is more common. In

philosophical contexts, I predominantly use "suspension" in a rather broad sense. Only when explicitly referring to a different phenomenon of doxastic neutrality, I will employ alternative terms. In particular, I will consistently use the term "suspension" when engaging in epistemological discussions about norms, as evident in Section 2.2.

Regarding the first question, I will in the following offer both a categorical and a gradual analysis of the different phenomena. The primary categorical distinction I will make throughout this project involves differentiating suspension from two other phenomena: ignorance and (mere) non-belief. Subsequently, I will present a gradual perspective, viewing suspension as the highest form of neutrality on a spectrum of neutrality, with ignorance positioned at the opposite end.

2.3.2.a Suspension versus Non-Belief and Ignorance

As already mentioned, for the longest time, epistemology predominantly concentrated on belief and disbelief, with the third option of doxastic neutrality being typically described in relation to these two. In this context, suspending judgment regarding p was defined as neither believing p nor disbelieving p. Suspension was characterized as an absence or lack of something, specifically the absence of belief coupled with the absence of disbelief. The term "non-belief," which is mostly used for this characterization, captures this idea properly – there is neither belief in p nor belief in $\neg p$.

Since doxastic neutrality has gained more attention, this characterization has come under increasing pressure. Most philosophers who have explored suspension argue that there is a distinction between suspension and non-belief. Many also posit that there is already a difference between non-belief and more fundamental forms of doxastic neutrality, such as indecision (Wagner, 2022). Although there is some disagreement on this matter, which I will delve into further in Part 2.3.2.b, rejecting the characterization of suspension as mere non-belief is now a common stance among most philosophers.[14] The most convincing argument against this identification can be found in Wedgwood (2002, p. 272). If we define

[14]Possible exceptions can be found in van Fraassen (1998); Zinke (2021b).

suspension simply as the absence of belief and disbelief, the term can be applied too broadly. Wedgwood (2002) notices that non-belief is not even an attitude, a position one can take, or a mental state at all. If suspension were to be identified with mere non-belief, even a stone could be described as suspending regarding any proposition p because a stone neither believes nor disbelieves p. But the ascription of suspension to a stone is clearly wrong. A stone does not suspend about whether Paris is in France or whether one plus one equals two, even if it is true that the stone neither believes nor disbelieves all those propositions.

A similar example, illustrating the absurdity of this characterization, is presented in Hájek (1998, p. 205, footnote). If we were to equate suspension with non-belief, we would end up describing a caveman as being in a state of suspension concerning the proposition that quarks exist. While, in principle, we would be more inclined to ascribe suspension about certain propositions to a caveman than to a stone, it is evident that they would not suspend judgment about this specific proposition since they lack understanding of the central concept, the concept of "quarks", of the proposition. Similarly, young children (unless they have an exceptional grasp of physics) do not suspend judgment about the proposition that quarks exist. Presumably, most readers will not suspend judgment about propositions like "Selenocysteine has the EC number 808-428-7" either.[15]

What these examples have in common is that a fundamental level of understanding of the relevant proposition and the concepts involved is

[15]In the following, I will use the example proposition about Selenocysteine as an illustration of a proposition that one may not understand due to a lack of familiarity with the involved concepts. If the reader understands this proposition, they are encouraged to substitute it with a different proposition from another field. According to a sociolinguistic hypothesis by Putnam (1973), finding such a sentence might turn out to be not that easy. Putnam (1973) claims that understanding the terms in a sentence and being able to differentiate them from others is not necessary to competently use the terms. Rather, it is sufficient that there are *some experts* in our linguistic community who could define the distinctive properties of the terms involved in the sentence, for all speakers to competently use them. This is what he calls the linguistic division of labor. For example, can I competently utter the sentence "This necklace is made of Gold" without being able to discriminate the properties of gold from other metals. Surely, we find experts on Selenocysteine in our linguistic community, too. Still, there will be sentences for every speaker that will not make sense to them, as the concepts involved are too alien. For many speakers, an example could even come from continental phenomenological philosophy. An obscure sentence from the large language model GPT3 (OPEN AI, 2024) was: "Upon encountering the apophatic hermeneutics of phenomenological eschatology, one grapples with the liminal interstices of existential praxis."

missing. This could be because the subject possesses no mental capacities whatsoever (as in the case of the stone) or because the subject lacks understanding of the relevant concepts in the proposition (as in the case of the caveman or in the case of Selenocysteine). Non-belief does not necessitate any form of understanding, and thus, it is not enough on its own to constitute suspension. Nevertheless, the majority of scholars maintain that there remains a relationship between non-belief and suspension. Although this relationship is questioned by Friedman (2013c), most scholars would argue that while non-belief alone is not sufficient, it is still a necessary condition for suspension. In other words, one cannot suspend judgment about p if one is not in a state of non-belief about p.[16]

Another phenomenon frequently encountered in the discourse on neutral doxastic states, which should not be confused with suspension, is described by the term "ignorance." As I will explain in the following, within this realm of neutral doxastic states, I intend to use the term "ignorance" to describe instances of non-belief that *cannot* be characterized as suspension, as exemplified by the caveman example. I am focusing on a more specific aspect of ignorance rather than using the term in its broadest sense.

The traditional view on ignorance identifies ignorance with lack of knowledge (Goldman and Olsson, 2009; Le Morvan, 2011).[17] According to this traditional view, ignorance appears to be a concept that stands in contrast to knowledge, unlike suspension or non-belief, which contrast with belief.[18]

However, Peels (2010, 2012) advocates for a "new view" on ignorance,

[16]The majority of scholars would generally agree to this essential condition for the type of suspension I am describing here. However, if suspension is employed in a zetetic way, linked to the idea of inquiry and double-checking, it has been argued that suspending can coexist with belief, (Wagner, 2023a; Lord and Sylvan, 2021). I will comment on this briefly in Section 2.4.2.

[17]"Ignorance" is a term that does not have a perfect equivalent in other languages, at least not without reference to other concepts. In German, for instance, the linguistically related term "Ignoranz" fails to encapsulate the same content as its English counterpart. Instead, terms like "Unwissen" or "Unkenntnis" are more often used, which directly translates to a lack of knowledge.

[18]When ignorance is understood as a lack of knowledge, and knowledge is classically analyzed based on the conditions of 1) belief, 2) truth, and 3) justification, one can be considered ignorant by failing to satisfy any of these conditions. Consequently, there exist (at least) three pathways to ignorance.

in which ignorance is not regarded as the opposite of knowledge but rather as the mere absence of belief in a true proposition p. While this new perspective has faced criticism (Le Morvan, 2011), I share the view presented by Peels (2010, 2012), who argues that there are situations where a subject lacks knowledge of proposition p but should not be described as ignorant of it. Such cases arise when a subject genuinely believes in p but lacks knowledge due to the absence of sufficient justification for their belief. According to the new view of Peels (2010, 2012), ignorance is best situated at the level of belief rather than at the level of knowledge. A person A can be considered ignorant of a proposition if either A does not believe p while p is true, or A does not disbelieve p while p is false. According to this definition, we can say that a person is ignorant of a proposition p if that person[19] either disbelieves p while p is true (or possibly believes p while it is false) or if they maintain a *neutral* doxastic position[20] regarding p.

As the focus of this work centers exclusively on neutral doxastic states, I will restrict my examination to those instances of ignorance in which an individual maintains a neutral position regarding the truth of the proposition.[21] Peels (2010) introduces various terms to distinguish between the different ways in which a person can experience ignorance. In the context of this work, the relevant form of ignorance aligns with what Peels (2010) refers to as "deep ignorance." We can ascribe deep ignorance regarding a proposition p to a person A, when they maintain a neutral doxastic position about p "because [the person] lacks the cognitive capacities or the relevant knowledge necessary to grasp p" (Peels, 2010, p. 62). This aligns with the example of the caveman.

Related to this is the distinction of Le Morvan (2011), which is part of his critique of Peels (2010), between what he terms "factive" and "propositional" ignorance. Being propositionally ignorant is being ignorant

[19] It is particularly intriguing to observe that such ascriptions of ignorance are invariably made from a third-person perspective. If person A believes that the Earth is flat, I can assert that A is ignorant of the fact that the Earth is round. A cannot ascribe ignorance to themselves in the same manner.

[20] Peels (2010) in fact uses the term "suspends judgment" here, equating it with mere non-belief.

[21] Given the symmetry of neutrality, where neutrality about p also implies neutrality about $\neg p$, the actual truth value of the proposition need not be a concern, for the purposes of this discussion.

of the content of the proposition, which is incompatible with believing p. Being factively ignorant is being ignorant of the truth of the proposition: "For example, of proposition q that Hillary Clinton is not identical to her autobiography, someone who has never considered q and who therefore has no doxastic attitude towards q is both propositionally and factively ignorant with respect to q" (Le Morvan, 2011, p. 341). The person is factively ignorant because they do not believe q to be true and propositionally ignorant because they are ignorant towards q itself as they have never considered q.

Le Morvan (2011) describes cases of propositional ignorance as scenarios in which an individual is ignorant because they have never considered the relevant proposition. While these cases differ from what Peels (2010) defined as "deep ignorance," where ignorance results from a lack of cognitive capacities, these two terms converge when examining typical cases of propositional ignorance, as emphasized by Peels (2012, p. 742). In his reply to Le Morvan, Peels (2012) asserts that a "bushman" would be propositionally ignorant regarding the statement that "glucose is composed of $C_6H_{16}O_6$" because he lacks knowledge of the truth conditions of this proposition. This is akin to his cases of deep ignorance and closely resembles the prototypical case of the caveman's deep ignorance about quarks. Both instances of deep ignorance and instances of propositional ignorance are cases of *non-belief* (or neutrality), while being clearly distinct from those cases of non-belief that are typically classified as suspension.[22]

In this sense, deep ignorance (which is the term I will use henceforth instead of the term "propositional ignorance") captures exactly those instances of non-belief that are typically *not* considered cases of suspended judgment. Those are cases like the caveman example.[23] The term "deep ignorance" is hence suitable to be used as the counterpart of the term "suspension" when investigating the sphere of doxastic neutrality.

Throughout this thesis, especially when applied to various AI

[22] Peels (2010, p. 62) also explicitly distinguishes cases of deep ignorance from situations where an individual lacks sufficient evidence to believe or disbelieve, which aligns more closely with our understanding of suspension.

[23] It is important to note that not all cases of non-belief can be neatly categorized into either ignorance or suspension. The classification often hinges on the specific definition of suspension, and we will explore cases in the middle of the spectrum in the following Part 2.3.2.b.

frameworks, I will thus employ the term "deep ignorance," or simply "ignorance" to describe a rudimentary form of neutrality accompanied by a certain lack of understanding or consideration, considering ignorance as the counterpart to suspension.[24] Along the spectrum of neutral doxastic states (all of which fundamentally hinge on non-belief as a basic, essential condition), ignorance can be regarded as the polar opposite of suspension. A visual representation of this concept, alongside typical examples, will be illustrated in the subsequent Part 2.3.2.b, in Figure 2.2 and Figure 2.3.

2.3.2.b Different Degrees of Engagement

As previously elaborated, most of the scholars investigating the nature of suspension of judgment agree with the distinction between suspension and non-belief. Some scholars go even further, postulating the existence of nuanced variations among various forms of doxastic neutrality. These forms of doxastic neutrality and the conditions that must be met in order to speak of one or the other form can be organized along a spectrum of neutrality, where ignorance marks one end and suspension represents the opposite end.

Example 2.1. Let us consider the following example propositions a person A is neutral about:

1 Selenocysteine has the EC number 808-428-7.

2 The guava fruit has a lot of vitamin C.

3 The COVID-19 pandemic will be over in 2022. (Stated in 2021.)

4 God exists.

Firstly, consider a person A, who is unaware of both Selenocysteine and what an EC number signifies. Secondly, A has only briefly heard about the guava fruit and has never thought about its nutritional value. Thirdly, in the year 2021, A found themselves occupied with the COVID-19 pandemic, wondering whether it would come to an end soon. However, they could

[24] Under this negative definition, one might argue that even a stone is ignorant about various propositions. In my opinion, attributing ignorance to objects like stones is not a critical concern. Since ignorance is defined by its negative aspect, it can be ascribed to various entities. Alternatively, one might contend that a proposition like "the stone is ignorant about $2+2=4$" involves a presupposition failure, as the possibility of not being ignorant is not inherent, akin to propositions like "7 is not green."

not resolve this question based on the available evidence and continued to search for additional information. Lastly, suppose A also spent much time thinking about the existence of God but ultimately arrived at the conclusion that there is no definitive means to find out. Consequently, they chose to remain neutral on this matter.

As discussed in Part 2.3.2.a, the situation regarding Proposition 1 typifies a classic instance of mere non-belief or deep ignorance, similar to the caveman example from Hájek (1998, p. 205, footnote). In this situation, the caveman could not even grasp the concepts involved in the proposition. Similarly, A finds themselves unable to grasp the concepts involved in Proposition 1. Conversely, the scenario concerning Proposition 4 offers an illustration that often serves as a prototypical case of "true agnosticism" or sophisticated suspension of judgment.

How does the neutrality concerning the other propositions relate to each other? It is evident that the situations involving Propositions 2, 3, and 4 all exhibit a more advanced form of neutrality compared to the situation involving Proposition 1. For Propositions 2, 3, and 4, the subject A understands the propositions and can attribute meaning to the concepts they involve. This understanding distinguishes A's state regarding Proposition 1 from the others.

Nonetheless, A's states concerning Propositions 2, 3, and 4 are different for each proposition. What distinguishes A's state for Proposition 2 from their state about 3 is what is sometimes referred to as "cognitive contact" (Friedman, 2013; Wagner, 2022). In the situation involving Proposition 3, A considered Proposition 3 and thought about its truth value. This is not the case for Proposition 2. There are numerous propositions akin to Proposition 2, which A (and other subjects) theoretically understand yet lack cognitive contact with. Examples include propositions such as "there are 7 zebras in Berlin Zoo," "San Rafael is a city in Argentina," "Japan won the last World Figure Skating Championship," or "Game of Thrones is available on Netflix." In all these cases, A is neutral about the proposition, understands the propositions, and can correctly outline its truth conditions. However, they never thought about them until just now. (At least this is the case for me.)

This is different for A's state concerning Proposition 3. This is a

proposition that A has been in cognitive contact with. They analyzed Proposition 3 and thought about its truth value. Nevertheless, back in 2021, they remained neutral, neither believing nor disbelieving it.

Some scholars have argued that having this cognitive contact is not only necessary but, together with non-belief, also sufficient for suspending judgment (Wedgwood, 2002; Hájek, 1998; Bergmann, 2005). In contrast, other scholars draw the line at a later stage. While they generally acknowledge cognitive contact as a necessary condition for suspension, they do not consider it sufficient.[25] What distinguishes genuine suspension (or agnosticism) from other forms of doxastic neutrality, according to scholars like Wagner (2022); Raleigh (2021); Friedman (2013c), is exemplified by the contrast between Propositions 3 and 4.

For Proposition 3, even though A remained undecided (in 2021), in the example, they continued to think about whether the COVID-19 pandemic would come to an end in the following year. In fact, almost every day, they encountered new evidence both supporting and opposing this proposition. Furthermore, they might have been strongly motivated to seek out information about this matter, possibly due to planning a trip for 2022. This is what distinguishes their state regarding Proposition 3 from their state regarding Proposition 4, as argued by Friedman (2013c) and Wagner (2022). Regarding Proposition 4, they "settled" their indecision and closed the question regarding the existence of God. This aligns with Sturgeon (2010), who describes suspension as "committed neutrality." Friedman (2013c) argues that this settlement[26] involves forming a *sui generis* attitude of suspension. Wagner (2022) characterizes the element of settling as an endorsement of one's de facto indecision. I will delve into these differing perspectives in Subsection 2.3.3. Regardless of how each specific account interprets this settling element, it is evident that A lacks this kind of settlement or cognitive closure concerning Proposition 3. As I described the example, A was highly motivated to find out about the

[25] Friedman (2013c) argues that cognitive contact is not necessary for suspension, as one might potentially enter a state of suspended judgment through means other than considerations. Nevertheless, she acknowledges that an attitude-based approach to suspension (which she advocates for) naturally includes cognitive contact.

[26] This contrasts with her account in Friedman (2017), where she portrays suspension not as an attitude that settles a question but, conversely, as an attitude that opens questions and inquiry.

proposition and constantly searched for new evidence for or against it.

To illustrate the differences, it can be beneficial to order the various forms of doxastic neutrality, ranging from mere non-belief or deep ignorance, to sophisticated (committed) suspension, as previously described.

The four different situations, in which A is neutral about the Propositions 1, 2, 3, and 4 respectively, can be referred to with Situation 1, Situation 2, Situation 3, and Situation 4. As we progress from one situation to the next, an additional condition that some consider constitutive of suspension is satisfied. Situation 2, unlike Situation 1, fulfills the condition of understanding.[27] Situation 3 goes a step further by satisfying the condition of cognitive contact (while still meeting the condition of understanding).[28] Finally, Situation 4 satisfies understanding, cognitive contact, and settlement.

Of course, there may exist example situations of doxastic neutrality that fall between the examples I have presented. For instance, one could argue that the caveman example should be put underneath the Selenocysteine example (since, unlike the caveman, I could potentially discover information about Selenocysteine), and that the stone example might be positioned even further down than the caveman. As the different forms of neutrality involved in the different situations build upon one another and further examples in between are always possible, it is helpful to represent the distinction of the different situations in this *gradual* ordering.

Overall, we can visualize the concept of ordering different forms of doxastic neutrality with the help of an axis representing a varying degree of engaging with the respective proposition A is neutral about. With a higher engagement to the proposition comes a more sophisticated form of neutrality.

[27] Understanding is here taken to be dispositional. If A were to consider proposition p, they would understand it.

[28] It can be debated whether one can be in cognitive contact with a proposition without understanding the proposition. In the sense in which cognitive contact can be realized by, for example, merely looking at a sentence with the respective propositional content, one can possibly be in cognitive contact even without understanding. I use a slightly more demanding concept of cognitive contact. Having cognitive contact with a proposition presupposes the (dispositional) understanding of the concepts within the proposition. Hence, the conditions on the axis build upon each other. The conditions for Situation 2 are also met by Situation 3, and the conditions for Situation 3 are met by Situation 4.

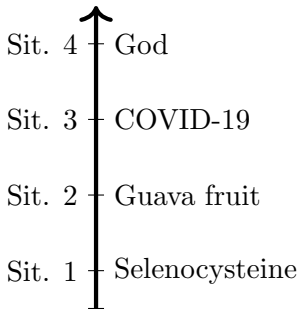


Figure 2.2: Neutral states along an axis of the degree of engagement.

When stepping from one situation to the next in this spectrum, an additional aspect is added. A had no engagement with the proposition concerning Selenocysteine, as they do not even understand it. While A does understand the proposition regarding the nutritional value of the guava fruit, they still do not truly engage with the proposition since it has never even crossed their mind. Nevertheless, this form of neutrality is more sophisticated than the first one. In comparison, in 2021, A actively engaged with the proposition concerning the COVID-19 pandemic to the extent that they continually contemplated it. Situation 4 deviates somewhat from this pattern. On one hand, one could argue that A is actually less engaged with the proposition about the existence of God than with the proposition about the COVID-19 pandemic, as they no longer think about the former. On the other hand, they are still, in some sense, more bound to the proposition in Situation 4, as they explicitly incorporate it into their doxastic sets, more precisely into the third neutral set of their doxastic household. They identify with being neutral about the proposition. In this regard, even though they do not think about it extensively anymore, they are more engaged with the proposition. They are farther along in the evaluation process compared to where they are (in 2021) with the proposition regarding the COVID-19 pandemic. They have concluded the inquiry about God's existence and have integrated the proposition into their doxastic framework by settling on neutrality. Consequently, Situation 4 represents the most sophisticated form of neutrality and must be positioned at the far right of the axis.

In a simplified form, suspension within doxastic neutrality, along with its relation to the doxastic counterparts belief and disbelief, can be illustrated in the following graph:

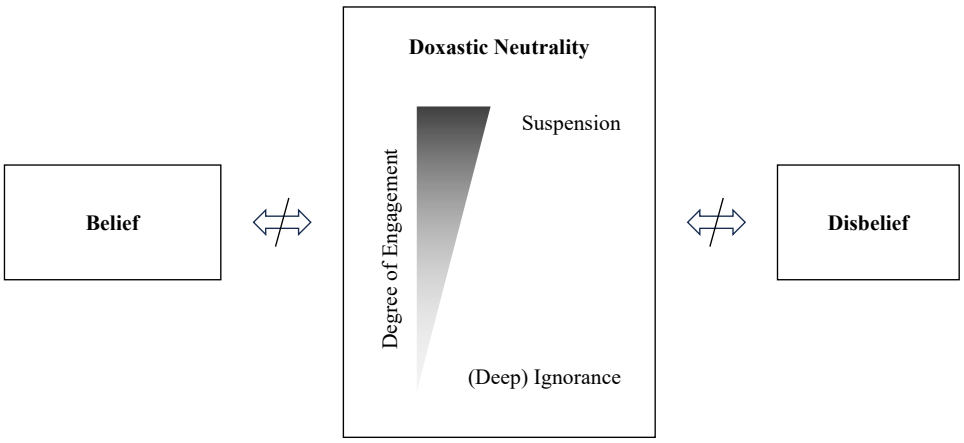


Figure 2.3: The doxastic sphere: Belief and disbelief are contrasted with doxastic neutrality. All forms of neutrality involve non-belief. Suspension and (deep) ignorance fall within doxastic neutrality, are categorically distinct from one another, but also mark opposite ends of a gradual scale of degrees of engagement. All forms of doxastic neutrality are also cases of *non-belief* as illustrated by the crossed-out arrows towards belief and disbelief.

2.3.3 The Nature of Qualified Suspension

2.3.3.a Different Accounts for Suspension

As the third doxastic stance garners increasing attention, numerous efforts have emerged to describe the psychological and phenomenological characteristics of this stance. Various theories aiming to provide a descriptive account of this mental state have been proposed. Just to name a few: Friedman (2013c); Wagner (2022); Raleigh (2021) offer different accounts of what suspension consists of. It is crucial to emphasize that these different and conflicting theories do not merely categorize different forms of doxastic neutrality (as described above) and thereby conflict each other. Rather, their proponents argue that they describe *the* sophisticated form of neutrality which is conventionally referred to by the terms “suspension” or

"agnosticism."[29] Within the spectrum of neutrality outlined in Figure 2.2, they aim to describe the version at the top, the most (or one of the most) sophisticated and engaging form, characterized by a state of settlement or commitment. They want to describe *the third doxastic attitude* towards proposition p, rather than any other neutral state.[30] Consequently, the key inquiry for these scholars is: How can we best describe this doxastic attitude? Can it be analyzed with the help of other mental vocabulary or is it not further reducible?

The leading response to this query is proffered by Jane Friedman (Friedman, 2013c), who asserts that the attitude of suspension of judgment cannot be reduced further. She argues that all attempts to reduce suspension to other concepts (such as the meta-cognitivist approaches that I will describe in the following) ultimately fail and that suspension must be understood as a *sui generis attitude*. This sui generis attitude is taken to be in itself indecision representing.

Other scholars try to unravel the nature of suspension or agnosticism with the help of other concepts. Wagner (2022), for example, defends an account of suspension[31] that she calls "endorsed indecision." She proposes a two-component analysis and claims that a subject is agnostic about p iff the subject is (a) undecided about p and (b) the subject endorses their indecision about p. This endorsement involves acknowledging the state of indecision and embracing it as appropriate and as the correct doxastic stance towards p. While indecision towards p is in a sense the first step towards suspension, it is only through endorsement that the subject settles the question about p and commits to their indecision. Besides the undecided state of mind, this settlement is necessary to obtain a suspending attitude (or to be in the agnostic stance as Wagner (2022) would call it.)

[29]The term "suspension" is occasionally used in a different, zetetic sense, as seen in Friedman (2017), where it denotes an interrogative attitude that initiates inquiry rather than settling it. I will briefly comment on this in Section 2.4.2.

[30]Some scholars would not define suspension as a doxastic attitude or stance at all. For example Crawford (2022) and McGrath (2021) take suspension of judgment to be an act rather than an attitude. Crawford (2022) characterizes suspension, for example, as "intentionally omitting [i.e., refraining] to judge whether p." To be precise, also Wagner (2022) uses the term "suspension" for the act of committing to one's indecision, which then leads to the agnostic attitude.

[31]Wagner (2022) calls the third doxastic attitude she aims to describe "agnosticism."

Another common approach to examining the nature of suspension is the so-called *meta-cognitive* account. This account describes a suspending person as having a meta-cognitive attitude. The condition of having a meta-cognitive attitude is also fulfilled in the endorsement account of Wagner (2022). Although an endorsement about one's indecision naturally involves a meta-cognitive attitude, endorsement requires an additional, more psychologically demanding dimension. Endorsement is a distinctly human attribute, whereas meta-cognition can potentially be exhibited by entities other than humans. Consequently, a meta-cognitive account is particularly suitable for application to artificial reasoning systems. Comparisons can be drawn between the neutral behavior of artificial reasoners and this description of suspension. As I will demonstrate, it is also viable to incorporate something similar to a meta-cognitive version of suspension into certain frameworks of artificial reasoning. Therefore, a more detailed elaboration on the meta-cognitive account is provided in the following.

2.3.3.b Meta-Cognitive Accounts of Suspension

Authors that subscribe to meta-cognitive accounts take suspension to be a *higher-order attitude*. In particular, the meta-cognitivist accounts agree that suspension towards p is a second-order belief. While a first-order belief is a belief about a proposition p, a second-order belief is a belief about a belief about p, a belief about a disbelief about p, a belief about the lack of a belief about p, and so on. Meta-cognitivists also agree that the object of the relevant second-order belief is some insight into one's own evidential or doxastic situation towards p. For example, Crawford (2004) states that

> Suspension of judgment necessarily involves thoughts about one's own epistemic perspective on whether or not p, namely, that one's epistemic perspective falls short of establishing whether p and thus that one does not know whether p. (Crawford, 2004, p. 226)

For Crawford (2004), the object of the second-order belief is the proposition that one does not know whether p and that one's own situation falls short of establishing whether p.

Bergmann (2005) emphasizes that withholding (which he takes to be the

third stance that I call suspension or agnosticism here) is not to be identified with non-belief but involves an attitude towards an attitude towards p, i,e., a second-order attitude:

> Withholding p, then, is a propositional attitude distinct from mere failure to take up any attitude towards p. Like believing or disbelieving, it is taking an attitude towards a proposition. What more can one say about withholding? As I shall be using the term, withholding p involves resistance, voluntary or involuntary, to believing p and to disbelieving p. The only thing one must consider in order to believe p or to disbelieve p is p (or its denial). But to withhold p (in the sense I have in mind) one must, in addition, consider the prospect of one's believing p as well as the prospect of one's disbelieving p; otherwise, one will not be able to resist both believing p and disbelieving p. So withholding p involves not only an attitude towards p but also attitudes towards attitudes towards p. (Bergmann, 2005, p. 214)

Furthermore, Rosenkranz (2007) describes a stable form of suspension that he calls "true agnosticism," which is a statement about what one is in a position to know:

> There is a genuine third stance which is identifiable by means of the assertion that we are neither in a position to know p nor in a position to know $\neg p$. (Rosenkranz, 2007, p. 63)

> [True agnosticism is] stable enough to generate commitments with respect to the debate's future course, and thus is more than a mere refusal to adopt any stance at all. (Rosenkranz, 2007, p. 101).

Raleigh (2021) also subscribes to what he calls a "meta-cognitive belief view," which states that

> suspending whether p constitutively requires having a belief or opinion that one cannot yet tell whether or not p, based on one's evidence. (Raleigh, 2021, p. 2455)

Here the object of the second-order belief is the proposition that one does not have the required evidence to tell whether p.

For the meta-cognitivists, suspension about p is constituted, first, by a second-order attitude (most plausibly a belief) that is about one's own problematic evidential or epistemic situation towards p *and*, second, by the respective evidential or epistemic situation towards p itself. Thus, the second-order belief itself is not sufficient. In addition, there must be some kind of doxastic neutrality or indecision towards the proposition p itself. Non-belief towards p is therefore also a necessary and minimal criterion for suspension. Moreover, it is sometimes claimed that the indecision and the second-order belief should stand in an appropriate relationship towards each other.[32] The relationship is spelled out differently in the different accounts, though. While for Crawford (2004), indecision is the primary state that is assessed with the second-order belief, for Raleigh (2021, p. 2457), the second-order belief is constitutive for the indecision. This means, that the belief about one's own evidential state somehow explains the indecision.

One can also see from these quotes that the scholars are all trying to describe suspension as a rather settled, sophisticated attitude that comes with a certain demand.[33] The idea that suspension is in a sense more demanding than believing or disbelieving can be found in the different descriptions of a meta-cognitive account.[34] For example, Bergmann (2005) states in the quote above that for suspension one has to consider not only the proposition itself but one's own attitudes towards it. Also, Crawford (2004, p.226) states that "it is one thing to believe (or disbelieve) something; it is quite another thing to suspend belief about that thing. The suspension of belief, or the reservation of judgement, has a degree of

[32]In the same sense, it is important for the account of Wagner (2022) that the endorsement and the indecision are in the right kind of relationship.

[33]Indeed, Raleigh (2021) distinguishes his account from that of Rosenkranz (2007) by asserting that suspension does not necessarily require "any such settled, strong commitment about the futility of future inquiry." However, the kind of settlement that is meant in this quote refers to settling or terminating future inquiry, which is not precisely the type of settlement I am describing here. Raleigh still regards suspension as a somewhat sophisticated attitude, as will become evident from his quote at the end of this section.

[34]This idea is not only present for the meta-cognitive account but also, for example, in the account of Wagner (2022). In fact, her criterion of endorsement of one's own indecision can be said to be even more demanding than having a second-order belief about one's evidential situation. Even more, the act of endorsing can be said to *pre-suppose* the forming of a meta-cognitive attitude.

cognitive sophistication beyond mere believing, in that it arguably involves the ability to have beliefs about what one does not believe." And Raleigh (2021, p. 2455) argues that "such a meta-cognitive opinion about what one can currently tell concerning some questions plausibly requires some degree of cognitive sophistication".

2.4 Overlapping Considerations: Nature and Norms

Finally, two topics should be considered that cannot be clearly categorized into either epistemology or philosophy of mind considerations yet remain relevant for the subsequent application chapters. As discussed earlier, the two levels (about the nature and the norms) often intertwine. This is particularly evident in the considerations that I will present in Subsection 2.4.1. Here, a differentiation between various forms of (qualified) suspension from the literature will be discussed. Although this inherently pertains to the nature of the phenomenon (i.e., philosophy of mind), the differentiation of forms is made based on the *normative profile*.

In Subsection 2.4.2, the perspective will be broadened once again to another, categorically different use of suspension. Thus far, I have used the term "suspension" as a neutral response to a given question. This usage will be maintained throughout the remainder of the thesis, particularly in its application to AI systems in Chapters 3 – 6. In Subsection 2.4.2, however, I will briefly mention another prevalent use of the term "suspension" that is characterized by its close connection to *inquiry*. While this use will not be the primary focus of this work, it can still be effectively used for demarcation. In Chapters 3 and 4, it helps to demonstrate that not every interpretation and facet of suspension is equally implementable in AI systems. These considerations in Subsection 2.4.2 about the different usage of suspension do not clearly fit into either the category of epistemology or the category of philosophy of mind either. On the one hand, one could argue that they belong to the field of philosophy of mind, since the nature of the term "suspension" is characterized differently. On the other hand, this fundamentally different characterization is also associated with different norms, which are discussed in this context.

2.4.1 Different Forms of Suspension

In addition to the discourse between meta-cognitivists and other approaches regarding how to accurately describe the nature of suspension, there is a separate debate concerning the potential existence of different forms of suspension or agnosticism. This conversation is largely independent of the debate about suspension and its relationship to other kinds of doxastic neutrality. Focusing *solely* on suspension itself, different variations of this phenomenon can be identified.

In the upcoming discussion, I will focus on distinctions between different forms outlined by Ferrari and Incurvati (2022). The authors use the term "agnosticism" for what I call "suspension" in this analysis. For consistency reasons, throughout this discussion, I will utilize the term "suspension" when referring to what Ferrari and Incurvati denote as "agnosticism." As mentioned already in Subsection 2.3.1 about the object of suspension, they draw a distinction between *gnostic* and *agnostic* attitudes towards a question that is under discussion. Here, belief and disbelief represent *gnostic* attitudes, as they involve the acceptance of one of the complete answers to the question under discussion. In contrast, suspension constitutes the *agnostic* attitude, leading to a response like "can't say" regarding the question under discussion.

They propose that there is a minimal agnostic attitude that can manifest itself through different stances.[35] They argue that the different stances reveal variations in certain epistemic and metaphysical commitments.

Ferrari and Incurvati (2022) draw three distinctions between different stances of agnosticism. The first distinction, and for them the most basic one, is between grounded and ungrounded suspension. This distinction concerns the question whether the subject inquired into the question under discussion. Simply speaking, a subject who somewhat inquired into the question under discussion and suspends is *grounded* in their suspending attitude, while a subject who suspends without having inquired into the question at all is *ungrounded* in their suspension.

[35]Notably, in this framework, Ferrari and Incurvati (2022, p. 371) consider an agnostic *attitude* to be more fundamental than an agnostic *stance*. *The* agnostic attitude can manifest in different agnostic stances.

The second distinction is, according to Ferrari and Incurvati a sub-distinction among the grounded stances of agnosticism. To account for the different perspectives of a subject on *further* evidence or further inquiry, Ferrari and Incurvati introduce the distinction between *pessimistic* and *optimistic* suspension. In the case of pessimistic suspension, an individual does not anticipate that further inquiry will definitively resolve the question in either direction. Conversely, in optimistic suspension, one does expect that further inquiry will play a relevant role in determining the truth-value of a proposition p. Ferrari and Incurvati (2022, p. 374) also introduce the concept of a "hesitant suspender" who responds with "can't say" when questioned about whether further inquiry will settle the matter.

Within pessimistic suspension, Ferrari and Incurvati (2022) again identify two different suspending stances. They introduce a distinction between *epistemic* suspension and *indeterminacy* suspension. This distinction is best explained when we recall that suspension is regarded as a response to a question under discussion. The differentiation between epistemic and indeterminacy suspension primarily concerns the individual's stance on whether the question under discussion can be answered or not. Indeterminacy suspension is at work when the subject views the question under discussion as fundamentally unanswerable. In contrast, epistemic suspension emerges when the subject believes that the question under discussion is, in principle, decidable or answerable but that they cannot find out about it.

Here, it is evident how the philosophy of mind and epistemology perspectives intersect. The delineation between various forms of suspension closely correlates with the normative dimension of suspension and the different reasons for suspension. For instance, recognizing a question under discussion as unanswerable provides a clear reason for suspending judgment. In fact, Ferrari and Incurvati claim that their focus is on describing the different stances of suspension that come with a particular *normative commitment*. Still, the commitments they take to be relevant can be both epistemic and metaphysical (as obvious in the case of indeterminacy suspension).

Cases for indeterminacy suspension can be identified when considering

the various justifications for suspension discussed in Section 2.2. Cases of positive justification are characterized by the explicit evidence supporting a suspending attitude. The most classic examples include cases involving chance (e.g., "The coin will land on heads") or vagueness (e.g., "The cup is blue"). In these instances, there is a clear correlation between the justification for suspension and the type of suspension adopted. If a subject possesses positive justification for suspension, they are likely to adopt a stance of what Ferrari and Incurvati call indeterminacy suspension.

More "extreme" cases of indeterminacy suspension arise in cases of mathematical indeterminacy, where a subject can deduce that the proposition is neither definitively true nor false but rather ontologically indeterminate or undecidable. The most notable example is the continuum hypothesis (Gödel, 1947; Cohen, 1966), which has been demonstrated to be independent of the standard set of axioms, meaning it is neither provably true nor provably false. Other cases encompass self-defeating propositions and liar-like sentences (Caie, 2012), often characterized as having either no truth value or an intermediate one (Schuster, 2023).

In contrast, epistemic suspension comes into play when the subject is in a deficient epistemic situation regarding proposition a p (or a question Q) to the extent that they cannot ascertain whether p is true or false. In contrast to indeterminacy suspension, here, suspension does not arise from the nature of the proposition but rather from the characteristics of the subject. The subject lacks the necessary evidence, cognitive resources, time, or other prerequisites to ascertain the truth or falsity of the proposition p. Again, one can find a clear correspondence with the different justifications for suspension. For example, when we suspend according to the Balance Norm, we have some evidence that speaks for p and some additional evidence that speaks for $\neg p$. In those situations, it seems mostly clear that we suspend not due to p itself being indeterminable but due to some epistemic limitation. Hence, in most of those cases, epistemic suspension is present.

Different Forms of Suspension	Distinguishing Feature
grounded vs. *ungrounded*	Has A inquired into the QUD? • if *yes*, then *grounded* • if *no*, then *ungrounded*
optimistic vs. *pessimistic* vs. *hesitant*	Does A think that further inquiry will resolve the QUD? • if *yes*, then *optimistic* • if *no*, then *pessimistic* • if A *cannot say*, then *hesitant*
epistemic vs. *indeterminacy*	Does A think that the QUD is answerable? • if *yes*, then *epistemic* • if *no*, then *indeterminacy*

Table 2.1: Three distinctions of Ferrari and Incurvati (2022) between different forms of suspension (or agnosticism in terms of Ferrari and Incurvati (2022)).

It is crucial to acknowledge that the various distinctions proposed by Ferrari and Incurvati (2022) represent just one possible framework for unraveling different forms of suspension. Additionally, in their framework, the distinctions between grounded and ungrounded, between optimistic and pessimistic, and between epistemic and indeterminacy suspension are presented as subcategories of one another, rather than as independent distinctions. According to this categorization, the broadest division is between grounded and ungrounded suspension. *Within* grounded stances, the further differentiation is made between optimistic and pessimistic attitudes. And again, *within* pessimistic stances, the distinction between epistemic and indeterminacy suspension is drawn. However, I consider

these distinctions to be somewhat independent, such that they can be combined more flexibly. For instance, whether one has already inquired into a question seems independent of whether one believes that further inquiry will resolve it. Hence, I find it too restrictive to assume that only when one has already inquired into a question (thus is in a grounded stance), can one be labeled as optimistic or pessimistic about further inquiry.

In particular, I do not believe that being pessimistic about the outcome of further inquiry is a prerequisite for adopting an epistemic suspension or indeterminacy suspension stance. It is possible to acknowledge that further inquiry may resolve the question while still recognizing one's own epistemic limitations, leading to a state of epistemic suspension.

However, I find one aspect of the restriction to pessimistic case appealing. Pessimistic stances often entail a sense of closure or settlement, which I consider crucial to the way I consider suspension in this thesis, as discussed in Subsection 2.3.2. With this, I diverge from other interpretations of suspension, such as that proposed by Friedman (2017), which views suspension not as a settling attitude but as one that initiates inquiry. Ferrari and Incurvati (2022) also observe that the inquiry-opening form of suspension described by Friedman (2017) does not encompass the pessimistic stances of suspension.

In fact, the nature of the connection between suspension and inquiry is an open debate. Different usages of the term "suspension" could potentially be differentiated by the various roles that suspension plays for inquiry. Hence, in the following subsection, I will briefly describe some findings about the connection of the two.

2.4.2 Inquiry and Suspension

The discussions and distinctions I have presented thus far have all centered around suspension (or doxastic neutrality) as a potential stance or attitude towards a question or proposition, with the intention of responding to the respective question or proposition. I have used the term "suspension" to describe a (more or less) settled attitude one can adopt towards a proposition, signifying neutrality.

However, there is some discussion regarding this characterization of suspension. A significant debate revolves around the relationship between

suspension and inquiry, a topic prominently addressed by Friedman (2017).

"Inquiry" is a term that, Ferrari and Incurvati (2022, p. 372), for example, define as "the practice of gathering and assessing evidence in order to answer a question under discussion." A prototypical case of an inquirer, one might think about, is a detective who performs the actions of collecting evidence from a crime scene and talking to witnesses. However, Friedman (2017, p. 10) makes clear that in this debate, the term "inquiry" is supposed to be used in a sense in which it makes the behavior of such a detective neither necessary nor sufficient for inquiry. Inquiring into a question means being in a "inquiring state of mind."[36] Inquiring is goal-directed in the sense that when inquiring into a question, we aim to find out the answer to the question. On the one hand, the actions of the detective are neither necessary, since inquiry could also consist in *only* being in a certain state of mind, without actively moving. On the other hand, they might also be not sufficient, if the detective is, for example, only performing these actions because she is told to do so but is not really aiming to find out who committed the crime, i.e., is not in an inquiring state of mind.

The prevailing consensus among most scholars is that inquiry and suspension are *closely interlinked*. Nevertheless, there is no clarity regarding the specific nature of this connection. Philosophers hold different viewpoints on *when* suspension comes into play during the process of inquiry. As I have used the term so far, suspension is one possible *outcome* of an inquiring process, alongside the other possible outcomes of belief and disbelief. On the contrary, the alternative view states that suspension is the *initial attitude* that does not terminate but rather start the process of inquiry. Lord (2020) and Lord and Sylvan (2021) present a unifying view that brings together the two perspectives. They assert that there are two distinct suspension of judgment attitudes: an anti-interrogative attitude, which aligns with what I have been describing thus far, signifying the end of an inquiry, and an interrogative attitude that starts (or continues) inquiry. As I will use this second, alternative view only sporadically in the discussions on suspension in AI, I will only briefly summarize the key elements of this perspective here.

[36]Still, the term "inquiry" refers to a (mental) *activity* that has the inquiring state of mind at its core.

The underlying idea of this alternative approach is that suspension is the doxastic attitude that represents neutrality, and neutrality is arguably an ideal (if not a necessary or obligatory) starting point for open inquiry. For Friedman (2017, p. 1), the association between suspension and inquiry is not only normative but also conceptual; she contends that "one is inquiring into a matter if and only if one suspended on the matter." Both normatively and conceptually, these two notions are inherently intertwined, according to Friedman (2017). Since suspension provides the relevant commitment to keeping the question open, it is necessary (but also sufficient) for genuine inquiry. She takes suspension and inquiry both to belong to a group of attitudes that she calls interrogative attitudes. Wondering, doubting, or being curious are also interrogative attitudes. She claims that not only inquiry, but all other inquiring attitudes imply suspension.

Support for the relation between inquiry and suspension is provided, according to Friedman (2017) and others, by means of the reference to historical examples of philosophical investigations about suspension, most prominently, by referring to Descartes (1641), who suspended judgment at the start of his methodological skepticism.

When suspension is considered within this close relationship to inquiry, Friedman (2020) claims that one is no longer operating within the doxastic realm but has entered the domain of *zetetic reasoning*. Zetetic reasoning places significant emphasis on the process of seeking evidence, investigating, double-checking, questioning, and the like.

The normative perspective put forth by Friedman (2017) concerning zetetic reasoning is encapsulated by her *Ignorance Norm*. According to this norm, one should refrain from inquiring into a question (or adopting any interrogative attitude) when already possessing knowledge of the question's answer. In a different set of papers, Friedman (2019a,b) takes this idea further by proposing the *DBI* (don't believe and inquire) norm. This norm suggests that one should not simultaneously hold a belief and engage in inquiry. To genuinely engage in an inquiry, even if it is a matter of double-checking, one is compelled to drop one's belief about the question. However, Friedman's norms have been subject to critique by scholars like

Archer (2018), Lord and Sylvan (2021), Wagner (2023a), among others. I will not delve into the discussion surrounding these zetetic norms in this context, as my focus (as previously indicated) will remain within the doxastic realm of suspension and its role in closing questions or inquiry.

2.5 Conclusion

In this chapter, the philosophical foundations for the thesis regarding suspension of judgment have been established. I elaborated on epistemological considerations regarding the normative aspects of suspension, descriptive considerations on the nature of suspension, and overlapping considerations that bridge both domains.

The epistemological perspective focuses on the normative profile of suspension of judgment. I particularly emphasized two distinct justifications for suspending judgment: positive and privative justification. This differentiation, as explicitly outlined in Zinke (2021b), highlights the complex nature of the normative profile of suspension, contrasting with belief and disbelief, which only allow for positive justification. Additionally, I introduced two standard norms governing suspension: the Absence Norm and the Balance Norm, both of which I regard as forms of privative justification for suspension. Furthermore, I discussed the intricacies of the logic of suspension, elucidating why it is more intricate than the logic of belief, as can be seen by the limited number of established logical principles governing suspension.

In exploring the nature of suspension, I provided an overview of the complex and dynamic discussions within philosophy of mind. I emphasized that there is no consensus on how to categorize and differentiate between the various phenomena of doxastic neutrality, nor is there clarity on the appropriate terminology to use for each grouping. In navigating this landscape, I aimed to utilize terminology and conceptual frameworks that would facilitate an investigation into the diverse frameworks of artificial intelligence, which is the focus of the subsequent chapters.

To this end, I delineated both categorical distinctions among different forms of doxastic neutrality and presented a gradual spectrum of these

forms. For the application to AI systems, it is crucial to distinguish between mere non-belief and suspension of judgment. Additionally, the term "ignorance" holds relevance, as it can describe cases of non-belief that do not meet the criteria for suspension. However, I argued against a rigid separation of suspension from all other forms of doxastic neutrality, advocating instead for a nuanced understanding that recognizes a spectrum of doxastic neutrality. This spectrum ranges from clear cases of ignorance at one end to clear instances of committed suspension at the other end.

I provided a closer examination of clear cases of qualified suspension, by introducing different accounts to characterize qualified suspension. In particular, I concentrated on the meta-cognitive account. This approach is demanding but still offers a valuable framework for delineating what constitutes qualified suspension. As I will demonstrate, it can potentially be extended to artificial agents, serving as a suitable reference for assessing the suspension capabilities of AI systems.

In addition to exploring topics in epistemology and philosophy of mind, Section 2.4 introduced two topics that do not neatly fit into either category. The first topic in this section investigates different forms of qualified suspension, most importantly the distinction between epistemic and indeterminacy suspension, adapted from Ferrari and Incurvati (2022). I showed how this distinction between the two forms of suspension is motivated by variations in the normative profiles underlying suspension.

The second topic of overlapping considerations involves a characterization of suspension that views it not as a settled response to an inquiry but rather as the initial stage of inquiry. This perspective, prominent in the work of Friedman (2017), diverges from my primary usage of the concept but is nonetheless significant to mention for a comprehensive understanding of the phenomenon. Furthermore, we will observe in Chapters 3 and 4 that this aspect of suspension is currently absent from the therein proposed framework for incorporating suspension.

Chapter 3

Floating Conclusions

This work is to a large extent based on and taken over from my paper "On floating conclusions," which is joint work with Jan Broersen and Henry Prakken (Schuster et al., 2023).

Contents

3.1 Introduction

This chapter addresses a key problem in non-monotonic reasoning related to suspension: Floating Conclusions. With the examination of floating conclusions, I illustrate a specific, easily understandable situation where suspension is at play, providing a foundational understanding of how the logic of suspension operates within AI systems. The investigation into floating conclusions offers insights into one of the various ways in which suspension manifests in non-monotonic reasoning systems, providing a specific context for reasoning with suspended propositions. Importantly, these investigations are independent of a specific logical framework, making them suitable as a case study for drawing broader conclusions about handling suspended propositions in logic-based AI in general.

As will become evident, cases involving floating conclusions lack a clear directive on how to proceed with reasoning, highlighting the significant logical complexity inherent in suspension, as discussed in Chapter 2. In the subsequent Chapter 4, an extension of default logic is introduced to derive specific rules for logical reasoning with suspended propositions. Building on the comprehensive examination of floating conclusions in this chapter, Chapter 4 outlines how the logic presented there deals with floating conclusions.

Cases of floating conclusions illustrate one classic scenario in non-monotonic reasoning where suspending judgment is demanded. Non-monotonic logic models defeasible reasoning, where every line of reasoning is fallible. Consequently, situations can arise where two (non-monotonically valid) lines of reasoning come into conflict. One famous example of a floating conclusion is presented in the Nixon case as discussed for example by Ginsberg (1993) and Horty (2002). In this example, one line of reasoning begins with the fact that Nixon is a Republican, leading to the defeasible conclusion that Nixon is a hawk, and ultimately concluding that Nixon is politically extreme. Another line of reasoning starts with the fact that Nixon is a Quaker, leading to the defeasible conclusion that Nixon is a dove, and again concluding that Nixon is politically extreme. These two lines of reasoning conflict because Nixon cannot be both a hawk and a dove. We must reject one line of reasoning. Nevertheless, both lines of reasoning,

despite their conflict, arrive at the same conclusion: the so-called floating conclusion that Nixon is politically extreme. Should we accept this floating conclusion? This is the immediate question that arises and is in focus for this chapter. The term "floating conclusion" from Makinson and Schlechta (1991) nicely captures that the conclusion "floats" above the conflicting arguments.

The question of whether we should accept floating conclusions is closely tied to the question of whether we should accept at least one line of reasoning from a set of conflicting lines of reasoning. This question arises primarily in defeasible reasoning and hinges on the notion that all the non-monotonic reasoning lines involved, while fallible, possess a certain degree of credibility or plausibility. When a conflict emerges, it becomes evident that at least one line of reasoning fails at some point. Given the uncertainty regarding *which* line of reasoning is faulty, we cannot simply accept one while not accepting the other. We should suspend about both. However, can we assume that *there is* at least one valid line of reasoning among them? That is, can we assume that while we suspend about both lines of reasoning individually, can we assume that we should believe that at least one of them is valid? If so, we should accept a floating conclusion; if not, we should not and suspend about the floating conclusion as well.

Interestingly, there no single definitive answer to this question. In different examples of floating conclusions, we appear to have conflicting intuitions about whether the floating conclusion should be accepted or not. In other words, when faced with conflicting situations, there are times when we believe that at least one line of reasoning is valid, while in other scenarios, we believe that the conflict destroys *both* conflicting lines of reasoning. This chapter aims to explain the diverse intuitions regarding the acceptability of floating conclusions across different examples. I will do so by examining different kinds of examples of floating conclusions from the literature, the associated intuitions about their acceptability, and existing explanations for these differing intuitions.

It is crucial to note that my approach relies on intuitions. I strive to establish a theory that can account for pre-theoretic intuitions across different examples and contexts. This approach is not without criticism. As observed in Veltman (1985) and again in Prakken (2002), employing

intuitions in logic encounters at least two challenges. Firstly, it raises questions about whose intuitions should carry weight, as individuals may have varying intuitions. Secondly, the assumption that intuitions should always be taken at face value can be questioned as such. Veltman (1985, p. 10) argues that when seeking intuitions, we are typically interested in the pre-theoretic judgments of "laypeople" who lack expertise in the field and have not been exposed to theories on the topic. However, this means we cannot ascertain whether these judgments reflect knowledge or merely some form of fallible belief. Consequently, such judgments are fallible and do not provide an "rock bottom empirical basis for testing logical theories" (Veltman, 1985, p. 13). Furthermore, (good) theories and arguments can certainly shape and alter the judgments that one has had pre-theoretically. Hence, it is important not to uncritically rely on intuitions. Nonetheless, theories should also align with broadly accepted common-sense judgments. Although individuals may differ in their intuitions regarding the acceptability of specific floating conclusions, it is undisputed that certain floating conclusions are generally regarded as acceptable, while others are not, particularly when considering that people must *take action* based on these conclusions. A theory that universally accepts all floating conclusions is as unsatisfactory as one that unconditionally does not accept any. In this chapter, I do not blindly embrace any one intuition. Rather, following the suggestion of Prakken (2002), I seek to identify underlying patterns in commonly shared intuitions, aiming to explain both the similarities and differences in these intuitions.

I will start by presenting several different examples of non-monotonic arguments that involve floating conclusions. I will demonstrate how these various examples can trigger different intuitions about whether we should accept the respective floating conclusion. Following this, I will introduce different hypotheses aimed at explaining these conflicting intuitions. I will test the validity of these hypotheses using the examples presented. After evaluating the various hypotheses, I will argue that no single explanation can account for all the diverse intuitions. Instead, my proposed solution will incorporate elements from different explanations. As a default, I argue that floating conclusions should be accepted. However, there are

circumstances that justify deviating from this default and not-accepting a floating conclusion. I will present two distinct reasons for deviation, which together comprehensively explain and cover all the examples presented. One reason applies if there is a possible "compromise" between the conflicting elements of the arguments; the second reason applies if the conflict is harmful not only to the conflicting part of the argument but also to other non-conflicting parts because, e.g., the conflict questions the credibility of the sources of information altogether. Both these reasons are grounded in the fact that in situations involving floating conclusions, the conflicting propositions are *contrary*.[1] This means that these propositions cannot both be true simultaneously but *can* be false simultaneously. This distinction sets such situations apart from dilemma scenarios involving *contradictory* propositions. If $p \rightarrow q$ and $\neg p \rightarrow q$, we can conclude that q. Here it is not possible that both antecedents are false, i.e., it is not possible that $\neg(p \vee \neg p)$. Hence, there is no third possibility besides p being true and $\neg p$ being true. This is different in the contrary case in which floating conclusions occur. The two explanations that give us reason to deviate from the default both spell out a way in which the conflicting propositions are both false, offering a third alternative beyond the two alternatives that are addressed in the argument, namely that one proposition is true and the other one false or vice versa.

In the final section (Section 3.5), I will concentrate on the connection between floating conclusions and the logic of suspension. Floating conclusions are only relevant when we consider skeptical reasoning approaches in non-monotonic reasoning.[2] According to Horty (2002), the skeptical approach can be interpreted as advocating for suspension when one is uncertain about a proposition. Considering this perspective, the question of whether a floating conclusion – one that follows from propositions we are supposed to suspend judgment upon – should be accepted, becomes a broader inquiry into how we should treat propositions stemming from other

[1] Thanks to Michael De for pointing me towards this.

[2] As will be precisely explained for the example framework of default logic in Chapter 4, skeptical reasoning is one approach to unify a plurality of possible results (extensions) in non-monotonic reasoning, as opposed to credulous reasoning. Skeptical reasoning essentially demands that only those entities (propositions or arguments) should be accepted that are accepted in *every* extension. This will be elaborated in detail in Chapter 4 and Chapter 5.

propositions for which we suspend judgment. Therefore, establishing an appropriate treatment for floating conclusions will provide insights into the general logic of suspension. The inquiries made at the end of this chapter provide a natural transition to the broader investigations into incorporating suspension within the non-monotonic framework of default logic, and the respective logical consequences, which will be presented in Chapter 4.

3.2 Examples of Floating Conclusions

In the following, I will provide examples of arguments that involve a floating conclusion. While some examples are drawn from existing literature, others are created for the purpose of this chapter to offer a nuanced understanding of the phenomenon. In a subsequent step, I will categorize them based on the various intuitions regarding the acceptance of the corresponding floating conclusion.

These investigations are to be considered in the context of non-monotonic reasoning. That is, the inferences leading to floating conclusions are generally non-monotonic, which means that the acceptance of a premise does not guarantee the acceptance of the conclusion. Additional information may override the conclusion derived initially. However, there exists a plausible, albeit defeasible, connection between the premise and the conclusion in non-monotonic inferences. While I will delve into a specific non-monotonic framework in detail in Chapter 4 on default logic, the following considerations are independent of any specific formalism and aim to facilitate general statements about non-monotonic reasoning and logic-based AI.

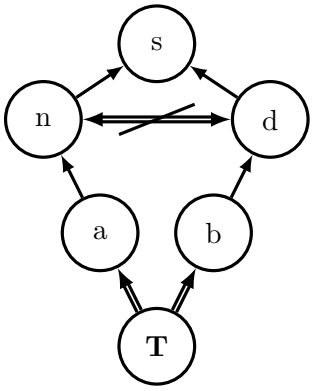
3.2.1 Introduction of the Examples

In the upcoming illustrations, I will employ small letters as abbreviations for propositions. The arguments will be depicted using arrows, where double arrows signify deductive, monotonic reasoning, and single arrows denote defeasible inferences. A double-sided crossed-out arrow indicates a conflict between two propositions, and the symbol **T** represents "truth." Known sentences are represented as inferred from **T**.

In total, I will present ten examples, enumerated from (I)-(X). The first

eight can be found in the literature, either in their exact form or in a slightly adapted version. The last two examples are own examples. While I provide explanations for some examples, others are self-explanatory.

(I) Ice-Skating (Prakken, 2002)



s Brigt Rykkje likes ice-skating.

n Brigt Rykkje is Norwegian.

d Brigt Rykkje is Dutch.

a Brigt Rykkje has a Norwegian name.

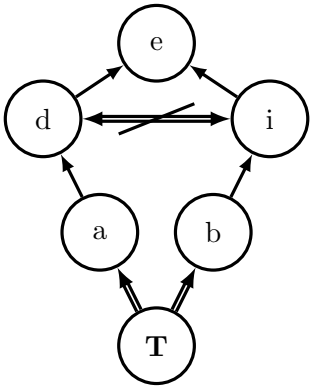
b Brigt Rykkje was born in the Netherlands.

Figure 3.1: Floating Conclusion graph for the Ice-Skating example.

The argument depicted in the illustration can be articulated as follows: It is true (it deductively follows from (**T**)) that Brigt Rykkje was born in the Netherlands (b) and that he has a Norwegian name (a). The argument on the right asserts that Brigt Rykkje is Dutch (d) because he was born in the Netherlands (b). Simultaneously, the argument on the left contends that Brigt Rykkje is Norwegian (n) because he has a Norwegian name (a). However, the statements (n) and (d) contradict each other. Both cannot be simultaneously true.[3] In light of this situation, we suspend judgment concerning both (n) and (d). The argument proceeds on both sides: On the right, Brigt Rykkje likes ice-skating because he is Dutch (d). On the left, Brigt Rykkje likes ice-skating (s) because he is Norwegian (n). Consequently, both lines of argument lead to the floating conclusion that Brigt Rykkje likes ice-skating (s), a conclusion derived from two contrary propositions, both of which are subject to suspended judgment.

[3]In nearly all examples, empirical assumptions are made, as in this case: Holding dual citizenships is not possible. This reflects our engagement in a non-monotonic setting with incomplete knowledge.

(II) Economy (Horty, 2002)



e We will have economic downturn.

d We will have deflation.

i We will have inflation.

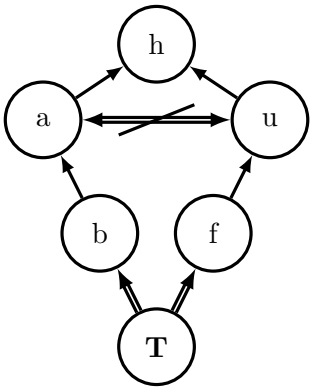
a Economist A says we will have deflation.

b Economist B says we will have inflation.

Figure 3.2: Floating Conclusion graph for the Economy example.

In this example of Horty (2002, p. 69), you find yourself in the setting of a macroeconomics conference "during a time of economic health with low inflation and strong growth." Once expert (A) at the conference uses a reliable model to predict declining inflation rate leading to dangerous economic deflation, while the other expert (B) uses a slightly different reliable model to predict that the current strong growth rate will lead to higher inflation. Both experts follow from their predictions that there will be an economic downturn.

(III) Student Housing (Broersen, 2017)



h It will be difficult to get housing.

a Susan wants to study in Amsterdam.

u Susan wants to study in Utrecht.

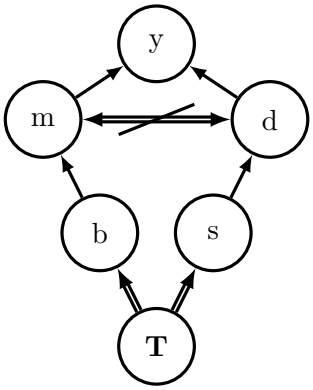
b Susan's boyfriend studies in Amsterdam.

f Susan's best friend studies in Utrecht.

Figure 3.3: Floating Conclusion graph for the Student Housing example.

In this example, desires are involved, distinguishing it from other non-monotonic inferences where the relationship between antecedents and consequences is factual, describing interrelations in the world. These inferences can be epistemically interpreted—if one were to believe the antecedents, one would believe the consequence. However, in the present case, the relationship is different, as desires are involved. For instance, Susan *wants* to study in Amsterdam (a) because her boyfriend studies in Amsterdam (b). Yet, the inference that housing will be expensive because she studies in Amsterdam is again of a purely factive or epistemic (non-practical) nature.

(IV) Yacht (Horty, 2002)



y I put a high deposit on a Yacht that costs half a million dollars.

m I will get half a million dollars from my mom.

d I will get half a million dollars from my dad.

b My brother tells me that Dad will give his money (half a million dollars) to him (my brother), but Mom will give her money to me.

s My sister tells me that Mom will give her money (half a million dollars) to her (my sister), but Dad will give his money to me.

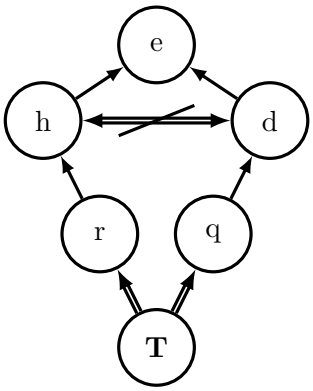
Figure 3.4: Floating Conclusion graph for the Yacht example.

This slightly complex example from Horty (2002) is about a situation where I have a brother and sister who are both equally reliable sources of information and have never lied to me. Although they are defeasible sources, their reliability is emphasized. As outlined by Horty (2002), my parents jointly possess a million-dollar fortune, split for tax reasons, with each having half.

Afflicted by the same terminal disease, they are certain to pass away within a month. Before falling into a coma, both parents shared different statements about which child would receive their respective halves, as reported by my siblings. My sister tells me: "Mom will give her money to me, but dad will give his money to you." My brother, however, tells me: "Dad will give his money to me, but mom will give her money to you."

Now, with my parents and siblings unreachable to me, and to cope with the grieving process, I contemplate purchasing a yacht I have been eyeing. Adding to this dramatic situation, there is a limited-time offer, and the yacht will be sold unless I place a deposit today. While I can afford the deposit, the total half-million-dollar payment will be due in six weeks, after my parents' passing. It is assumed that the optimal use of the inheritance, if received, is to purchase this specific yacht. Failure to inherit the half-million dollars will render me unable to pay for the yacht, resulting in the loss of the deposit.

(V) Nixon (Horty, 2002)



e Nixon is politically extreme.

h Nixon is a hawk.

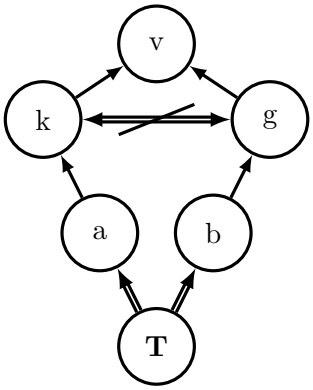
d Nixon is a dove.

r Nixon is a Republican.

q Nixon is a Quaker.

Figure 3.5: Floating Conclusion graph for the Nixon example.

(VI) Murderer (Prakken, 2002)



v Peter killed the victim.

k Peter killed the victim with a knife.

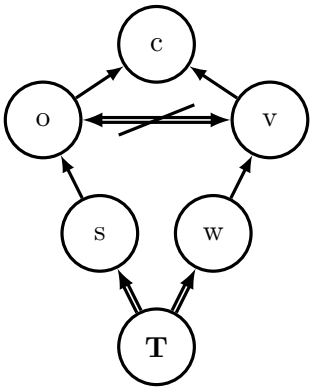
g Peter killed the victim with a gun.

a Witness A says that Peter killed the victim with a knife.

b Witness B says that Peter killed the victim with a gun.

Figure 3.6: Floating Conclusion graph for the Murderer example.

(VII) Mary in Europe (Reiter, 1980) and (Horty, 2002)[4]



c Mary lives in Berlin or in Madrid.

o Mary lives in Berlin.

v Mary lives in Madrid.

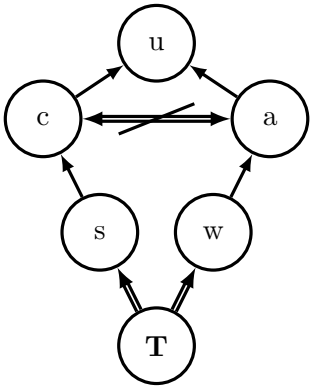
s Mary's spouse lives in Berlin.

w Mary's work is in Madrid.

Figure 3.7: Floating Conclusion graph for the Mary in Europe example.

[4]The cities involved in the next two examples (VII) and (VIII) have been changed in Schuster et al. (2023) compared to the original in order to introduce a European setting.

(VIII) Carol in Germany (Horty, 2002)



u Carol lives in Düsseldorf or in Bonn.

c Carol lives in Düsseldorf.

a Carol lives in Bonn.

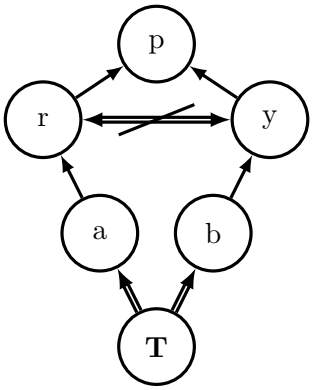
s Carol's spouse lives in Düsseldorf.

w Carol's work is in Bonn.

Figure 3.8: Floating Conclusion graph for the Carol in Germany example.

Certainly, most examples could be theoretically extended to include more than two lines of reasoning that are in conflict. This is especially evident in the previous two examples. Perhaps, for example, Mary's hobby takes place in London.

(IX) Primary Color (Schuster et al., 2023)



p The cup is colored in a primary color.

r The cup is red.

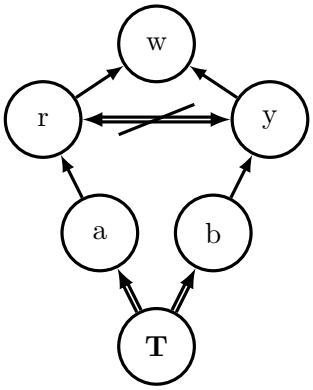
y The cup is yellow.

a Anna says that the cup is red.

b Ben says that the cup is yellow.

Figure 3.9: Floating Conclusion graph for the Primary Color example.

(X) Wavelength Color (Schuster et al., 2023)



w The color of the cup has a higher wavelength than the wavelength of blue.

r The cup is red.

y The cup is yellow.

a Anna says that the cup is red.

b Ben says that the cup is yellow.

Figure 3.10: Floating Conclusion graph for the Wavelength Color example.

3.2.2 Intuitions about the Examples

While individual intuitions may vary, evidence from presenting this material and engaging in discussions with colleagues across different countries suggests that a majority tends to share the following intuitions about the acceptance or non-acceptance[5] of the floating conclusion in each respective example.

[5]In the literature, people typically distinguish between *accepting* and *rejecting* the floating conclusion. However, "rejecting" a floating conclusion does not imply "disbelieving" it by believing the negation, as there is still no evidence for the negation of the floating conclusion. Instead, rejection in these cases signifies suspension. As I will demonstrate in Chapter 4, in non-monotonic reasoning frameworks, there is often no differentiation between actual rejection and suspension. "Rejection" can typically encompass both. Since this thesis focuses on suspension, I will be more precise and refer to "non-acceptance" instead of "rejection."

Floating conclusion accepted	Floating conclusion not accepted
(I) Ice-Skating	(II) Economy
(III) Student Housing	(IV) Yacht
(VII) Mary in Europe	(V) Nixon
(X) Wavelength Color	(VI) Murderer
	(VIII) Carol in Germany
	(IX) Primary Color

Table 3.1: Table of accepted and not accepted floating conclusions in different examples.

Some examples require contextualization and rely on certain empirical assumptions. For instance, in the (I) Ice-Skating example, assumptions include the impossibility of holding dual citizenship (in particular Dutch and Norwegian simultaneously), the popularity of ice-skating among the majority of Dutch and Norwegian people, and the understanding that being born in the Netherlands does not *necessarily* grant Dutch citizenship.

The (III) Student Housing example presupposes that both Utrecht and Amsterdam face challenging housing markets. In the cases of (VII) Mary in Europe and (VIII) Carol in Germany, the assumptions hinge on the geographical understanding that Köln is situated between and close to both Bonn and Düsseldorf and that Berlin and Madrid are approximately two thousand kilometers apart. In the (VI) Murderer example, it is assumed that it is impossible to kill someone with *both* a knife and a gun. The two color examples, (IX) Primary Color and (X) Wavelength Color, are based on the assumption that the cup is only (and fully) covered in *one* color.

In the (V) Nixon example, the empirical assumptions are straightforward once one understands that in the American political context, a "dove" opposes any military use, while a "hawk" believes that any military use is justified in a conflict. Additional empirical assumptions include the tendencies for Republicans to lean towards hawks and for Quakers to lean towards doves. Besides that, the Nixon case is probably the most controversial and therefore the most interesting one. In former literature,

(Ginsberg, 1993) people argued that the floating conclusion that Nixon is politically extreme should be accepted. However, I think that it should not be accepted. Especially when I explain my reasons *why* the Nixon floating conclusion should not be accepted, people seem to sometimes change their intuitions and admit that in fact, one cannot conclude that Nixon is politically extreme. Without presuming additional knowledge about Nixon as a person, it is natural to think that his Quaker and his Republican side "balance each other out" such that he ends up with no politically extreme stance. This is in line with what I said in the introduction. Intuitions are not infallible, and they are not necessarily stable. Sometimes good explanations can change our assessments contrary to the first intuitions we had.

3.3 Hypotheses

In this section, I present different hypotheses that aim to explain why some but not all floating conclusions seem acceptable. Some of the hypotheses can be found in the literature, others are new. Afterwards, I will then evaluate the hypotheses by my examples.

3.3.1 Presenting the Hypotheses

The following hypotheses aim to explain the different assessments of the various examples of floating conclusions. Some of the theses can be found in the literature, others have been developed in Schuster et al. (2023). While some of the theses may seem immediately less convincing than others, particularly in terms of their all-inclusiveness, I will present all theses encountered. One reason for this is completeness, and another reason is that each thesis is often motivated by one or two concrete examples, thereby contributing to our understanding of the nuanced aspects of the examples.

1. **Vagueness**: One possible explanation for the different assessments refers to the concept of vagueness. Some conflicts can be seen as borderline cases for vague concepts that are involved in the corresponding arguments. If a vague concept is involved, and the conflicting propositions incorporate clear, non-borderline cases of

the concept, it has to be tested whether the floating conclusion also follows from the borderline case.

- If the floating conclusion does not follow from the borderline case, it should not be accepted.
- If the floating conclusion does follow from the borderline case, it should be accepted.
- If there is no vague concept, the floating conclusion should be accepted.

2. **The direction of fit** (Broersen, 2017): The difference could stem from different directions of fit. Beliefs concern propositions that aim to describe the world, hence the direction of fit can be described as proposition-to-world. Desires and intentions, on the other hand, concern propositions that describe how the world ought to be, so the direction of fit is world-to-proposition.

   - If the conflict is between two beliefs, the conflicting beliefs "cancel each other out," resulting in the non-acceptance of the floating conclusion.
   - If the conflict is between two desires or intentions, the conflicting desires or intentions do not cancel each other out. Thus, at least one of the desires will remain intact and the floating conclusion is to be accepted.

3. **Hidden Defaults** (Prakken, 2002): This explanation states that the perceived unacceptability of certain floating conclusions stems from implicit "hidden defaults" that are not explicitly stated but must be inferred in the respective examples.[6] These hidden defaults undermine[7] the presented defaults that lead to the alleged floating conclusions, rendering them not floating conclusions but conclusions of defeated defaults. This can be compared to examples that do not

[6] Prakken (2002) employs the terminology of default logic, referring to the non-monotonic reasoning rules as defaults. However, these thoughts apply to other non-monotonic reasoning systems as well.

[7] In fact, the hidden defaults *undercut* the presented default. Undercutting is a distinct form of defeat that destroys the relationship between the antecedents and the consequence, leaving no evidence for or against the consequence. I will elaborate on this form of defeat in Chapter 4.

involve floating conclusions at all. For instance, if a default asserts that the song "Wind of Change" is an 80s song because it is included in the playlist "Best of 80s," this default could be defeated by another default stating that the playlist "Best of 80s" includes songs from other decades. Consequently, we would no longer infer that "Wind of Change" is an 80s song based on this information. According to Prakken (2002), such a hidden default exists and defeats the defaults leading to the apparent floating conclusion whenever the floating conclusion appears not acceptable.

- If hidden defaults exist that defeat both arguments supporting the apparent floating conclusion, it should not be accepted.
- If no such hidden default exists, the floating conclusion is to be accepted.

4. **Possible Compromise**: This explanation suggests examining the compatibility of conflicting propositions. If a possible "compromise" or intermediate position exists between conflicting propositions, this compromising position is likely the case. In such a situation, one needs to check if the floating conclusion also follows from the compromising case.

   - If the floating conclusion follows *only* from the presented "extreme" cases but not from the compromising one, it must not be accepted.
   - If the floating conclusion follows *also* from the compromising case, it must be accepted.
   - If there is no plausible compromise between the conflicting propositions, one is justified in thinking that at least one of the conflicted propositions is true, and hence the floating conclusion is acceptable.

5. **Harmfulness of the Conflict**: This explanation analyzes the conflict and the sources of information. Sometimes, it appears that the conflict is only harmful to the conflicting propositions themselves. In other cases, however, the conflict seems to destroy the credibility of the sources of information more generally.

- If the conflict destroys the credibility of the sources of information in general, then there is no longer a reason to assume that at least one line of reasoning is valid, leading to the non-acceptance of the floating conclusion.
- If the conflict is only harmful to the conflicted propositions themselves, the floating conclusion is to be accepted.

3.3.2 Testing the Hypotheses

In this section, I will evaluate the presented hypotheses using the previously introduced examples. While most hypotheses explain certain examples well, none manages to account for the intuitions behind every presented example.

1. **Vagueness:** The vagueness hypothesis is motivated by the different intuitions in examples like (X) Wavelength Color versus (IX) Primary Color. These involve a vague concept (color), where conflicting propositions ("The cup is red" and "The cup is yellow") can be made compatible by a third proposition ("The cup is orange"), representing the borderline case. In the (IX) Primary Color example, the floating conclusion does not follow from the borderline case, and it is thus not accepted. Whereas for (X) Wavelength Color, the floating conclusion *does* follow from the borderline case, and it is thus accepted. However, vagueness alone does not explain all intuitions, as there are examples without vagueness for which we still have varying intuitions. These examples cannot be explained solely by this hypothesis. For instance, (IV) Yacht is an example for which we want to suspend judgment about the floating conclusion, and (I) Ice-Skating is an example which we want to accept, yet neither involves a vague concept. In particular, the thesis that floating conclusions should be accepted in the absence of vague concepts is contradicted by cases like (IV) Yacht or (VI) Murderer.

2. **The direction of fit:** The idea that different directions of fit can yield varying intuitions about the acceptability of floating conclusions was initially proposed in Broersen (2017) based on the different intuitions in the examples (II) Economy and (III) Student Housing. In the latter example, the conflict arises from conflicting desires — Susan wants to study both in Utrecht and in Amsterdam. Although one

desire will eventually prevail, the desires do not cancel each other out, unlike in the Economy case, where the conflict is between beliefs. However, this explanation falls short in other examples. (I) Ice-Skating is an example without desires or intentions, relying solely on beliefs. However, we still find it appropriate to accept the floating conclusion in (I) Ice-Skating. The thesis fails in the opposite direction as well. If one were to adapt (VIII) Carol in Germany to an example about Carol's *desires* to live in one city or the other, the hypothesis would suggest accepting the floating conclusion, although we want to suspend about it.

3. **Hidden Defaults:** Prakken (2002) argues that the examples (IV) Yacht and (VI) Murderer as such do not provide sufficient grounds to argue for a suspension about the floating conclusions. The conflicting propositions and the resulting floating conclusions in both cases are, according to Prakken, defeated. This defeat occurs because the defaults leading to the conflicting propositions are undercut by other defaults that are not explicitly mentioned in the theory. In the case of (VI) Murderer, the alleged floating conclusion's unacceptability is attributed to the hidden default: If two witnesses provide contradictory statements, their credibility is dismissed. This hidden default undercuts both the default inferring that Peter killed the victim with a gun and the default inferring that Peter killed the victim with a knife, resulting in no floating conclusion. Similarly, in the (IV) Yacht example, a hidden default would undercut both arguments that rely on the testimonies of my sister and my brother.[8] This strategy is applicable in other examples as well. In the Nixon case, an additional hidden default could state that if someone is both a Quaker and a Republican, one cannot make any assumptions about their stance on military operations. This hidden default would then undercut both defaults inferring that Nixon is a dove or that Nixon is a hawk. Consequently, the floating conclusion that he is

[8]Prakken (2002) describes the example slightly differently than Horty (2002) does. In his version, both my sister and my brother tell me that they spoke to both parents and that my mom (respectively my dad) told my sister (respectively my brother) that she (he) will give me her (his) money. Prakken argues that this example relies on the additional default that people tend to speak the truth about their intentions, which is undercut as soon as people (in this case both mom and dad) tell conflicting things about their intentions.

politically extreme would not follow. The rather clear case of (I) Ice-Skating also supports this hypothesis. There is no apparent hidden default that should be visualized in the example, leading to the intuitively correct conclusion that the floating conclusion is acceptable.

However, the strategy becomes more questionable when considering examples like (VII) Mary in Europe versus (VIII) Carol in Germany (or the Color examples), where the same defaults in one case lead to seemingly acceptable floating conclusions and in another case to unacceptable ones. Why should there be hidden defaults in one case but not in the other? Prakken (2002) also admits that this strategy might not be valid for all possible examples, such as conflicts arising from different interpretations of legal norms. Moreover, I think that, although this approach might be applicable to many examples, it does not explain *why* in certain situations a floating conclusion is to be accepted and in others not. By referring only to possible missing defaults, we might find a way out of the unequal treatment of different floating conclusions in given examples. However, it still shifts the burden of explanation only to the question of why we feel that defaults are missing (or hidden) in some cases, while in other cases, this is not the case. After all, in all examples, there is always some additional context one can provide to make the example more informative. The thesis does not tell us, though, which parts of this additional information will be necessary to possibly introduce additional undercutting defaults, and which can be validly ignored.

4. **Possible Compromise:** The idea behind this thesis is best illustrated by the differing intuitions in the (VII) Mary in Europe and (VIII) Carol in Germany case. Although the rules leading to the conflict and the floating conclusions are of the exact same form, the floating conclusion seems justified in one case and not in the other (as observed by Horty (2002)). What explains the difference in this particular case? It seems that the conclusion that Mary lives either in Madrid or in Berlin is acceptable because there is no viable alternative option in the "middle." Since the cities are extremely far away from each other, it is unlikely that Mary could live somewhere in the middle and commute between the places daily. This differs

in the case of Carol in Germany. Since both cities, Düsseldorf and Bonn, are not very far away from each other, and there is a good "compromise," Köln, in the middle, it is likely that Carol neither lives in Düsseldorf nor in Bonn but went for the compromise, the city in between. This idea can be applied to other examples, too. In the (II) Economy case, there is a "compromise"[9] between (strong) deflation and inflation, namely that there will be none of both. Similarly, in the (V) Nixon case, the compromise between Nixon being a hawk and Nixon being a dove lies clearly in the middle, describing Nixon as not having a clear or extreme opinion on military use. In both cases, we do want to suspend about the floating conclusion because the compromise is too likely, and the floating conclusion does not logically follow from the compromise.[10]

This is different for cases such as (I) Ice-Skating and (III) Student Housing. There is no attractive student town between Amsterdam and Utrecht, and it is even less plausible that Brigt Rykkje can have a citizenship "in between" Norwegian and Dutch. Therefore, we should adhere to the conclusion that, even if he cannot have both, he has at least one of the citizenship, allowing the floating conclusion to be drawn. While we still suspend about his exact citizenship, we believe that one of them is the correct one and that no third alternative comes into play. Hence, we do not suspend about the floating conclusion, but believe it. The general idea is that if there is no compromise between conflicting propositions, then it is likely that at least one line of reasoning is correct, and the floating conclusion will follow. If there is a plausible compromise, then it must be tested if the floating conclusion also follows from this compromise. This is also illustrated through the two color examples. In the identically

[9]The term "compromise" may be somewhat unusual in this context. Not all the described cases involve a compromise in the sense of people agreeing on something. What is meant by compromise is rather an unignorable possibility or relevant alternative. The term "compromise" is used because it nicely suggests that this alternative or possibility lies somewhere *in the middle* on a spectrum at the end of which the two conflicting options lie (and is not simply some additional alternative that lies outside the spectrum considered so far).

[10]Horty (2002, p. 69) already suggests something similar in his considerations of (II) Economy and (V) Nixon: "Perhaps the extreme predictions are best seen as undermining each other, and the truth lies somewhere in between."

constructed examples (X) Wavelength Color and (IX) Primary Color, the compromise (that the cup is orange) entails the one floating conclusion (that the cup is colored with a higher wavelength than blue) but not the other floating conclusion (that the cup is colored in a primary color).

However, the examples of (VI) Murderer and (IV) Yacht cannot be perfectly explained by this hypothesis. The reason why we do not want to accept the floating conclusion in those cases is not that there seems to be a compromise or intermediate position between the two conflicting propositions. Rather, it appears that the mere existence of a conflict undermines the credibility of both argument lines, leaving us with no argument for the floating conclusion and thereby suspending about it.

5. **Harmfulness of the conflict:** The thesis regarding the harmfulness of the conflict is founded precisely on this observation concerning the (IV) Yacht and (VI) Murderer example. The fundamental idea is that conflicts come in various forms, some of which are harmful to the floating conclusion, while others are not. The cases of (IV) Yacht and (VI) Murderer, for instance, involve a conflict where two witnesses assert different things that, although conflicting in one aspect, are consistent with each other in another aspect. In the (VI) Murderer case, the witnesses' testimonies conflict regarding the murder weapon Peter used, but they agree that it was Peter who killed the victim. In the (IV) Yacht case, the siblings' testimonies conflict about Mom and Dad's intentions with their half a million dollars, but they agree that I will end up inheriting half a million dollars from one of them. However, we would not want to conclude that Peter killed the victim or that I would inherit half a million dollars. We suspend about these conclusions. Why is this? The conflict involved seems to be *harmful not only to the conflicting part* itself but to the *entire situation*. The existence of the conflict raises doubts about the credibility of the witnesses, suggesting that something more general has gone wrong. We might suspect that the two witnesses or the siblings have coordinated their statements, that the conditions for witnessing Peter's actions were not ideal, or that our parents have no intention of revealing anything about who gets their money.

This explanation can be extended to the (II) Economy example as well. In other instances, such as (III) Student Housing or (I) Ice-Skating, the conflict does not appear to damage anything beyond the conflicting components themselves. We have information favoring Brigt Rykkje being Norwegian and other information favoring him being Dutch. However, these different sources of information seem independent of each other and are not harmed by the conflict. In cases where the conflict is harmful to the overall argument, it is because the *credibility* of the sources of information or the *authority* of the experts is undermined by the conflict.

It is not clear, however, how this explanation succeeds in explaining the different intuitions about (VII) Mary in Europe and (VIII) Carol in Germany, or (X) Wavelength Color and (IX) Primary Color. The conflict involved is exactly the same in form, and thus, it is not apparent why the conflict is harmful to one floating conclusion but not to the other.

In conclusion, it can be said that none of the presented hypotheses is suited to explain the intuitions about all examples. Still, a positive outlook suggests that some hypotheses manage to describe well what is going on with a subset of the examples. In fact, the two hypotheses, "Possible Compromise" and "Harmfulness of the Conflict," when combined, can describe the heart of the matter for *all* the different examples. For instance, "Possible Compromise" explains the varying assessments for (VII) Mary in Europe and (VIII) Carol in Germany, as well as for (X) Wavelength Color and (IX) Primary Color by referring to the compromising proposition. "Harmfulness of the Conflict" captures the basic intuition of, for example, (IV) Yacht or (VI) Murderer where different testimonies seem to harm the whole argument. This motivates the introduction of my proposed framework for handling floating conclusions, which I will discuss in the following Section 3.4.

3.4 A Default Framework for Floating Conclusions

Here, I propose a solution that combines different hypotheses, manages to hit our intuitions in the different examples, and provides a straight-forward

way to evaluate the treatment of a floating conclusion for a given example.

The basic idea is that a floating conclusion should be accepted by default. Then, there can be different reasons to deviate from the default and suspend about a floating conclusion. One such reason stems from the "Harmfulness of the Conflict" thesis. If a conflict is not only harmful to the conflicting propositions but destroys the credibility or authority of the sources of information entirely, then the floating conclusion is to be suspended. Another reason to deviate from the default and not to accept the floating conclusion is described by the "Possible Compromise" thesis. If there is a compromise between the conflicting propositions, and the floating conclusion does not follow from this compromising proposition, then the floating conclusion is to be suspended. According to my solution, evaluating whether a floating conclusion is to be accepted can be visualized by walking through the following flowchart.[11]

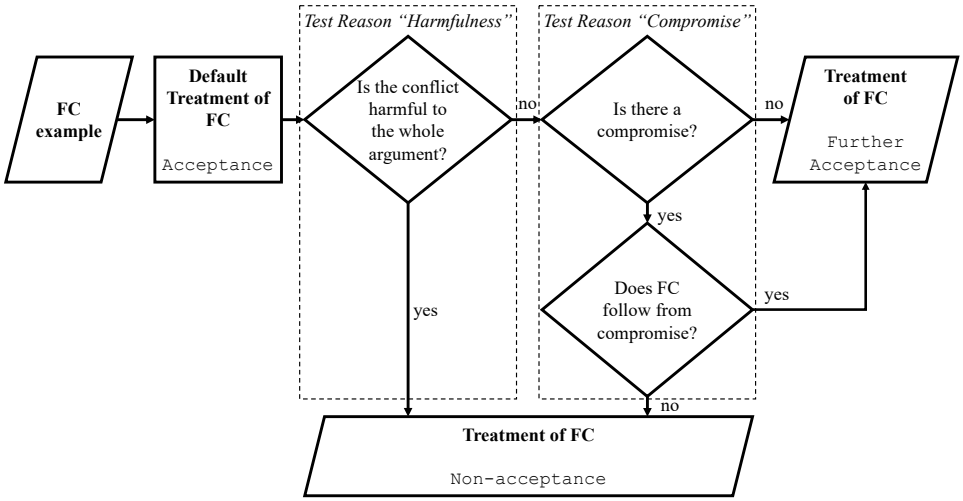


Figure 3.11: Flowchart for evaluating the status of a floating conclusion (FC).

How is this solution and the choice of these two hypotheses serving as deviating reasons motivated? In all cases of floating conclusions, the conflicting propositions are contrary to each other. Although they cannot

[11]It is easy to see how the flowchart could be translated into the framework of a fast and frugal tree (Phillips et al., 2017), a commonly used decision tree in which there is an end node (in this case, a treatment for the floating conclusion) at every decision node.

both be true at the same time, they can both be false at the same time. That is, in addition to the possibility that one proposition is true and the other false (or vice versa), there is a *third possibility*: Both propositions are false. The two reasons to deviate from the default each describe one version of this third possibility. Either we reject both propositions because the credibility of their justification has been undermined, or because there is a third proposition available as a compromise. Incorporating these two hypotheses, moreover, allows including all the other potential theses that have been discussed to a certain extent. For example, the cases of vagueness describe special cases of a possible compromise, but the "Possible Compromise" thesis is broader as it allows for different kinds of compromise. Moreover, "Harmfulness of the Conflict" subsumes the idea of the "Direction of Fit" thesis, which also states that some conflicts (involving beliefs) cancel each other out while others (those that involve desires) do not. These are special cases of describing whether a conflict is harmful to the whole argument or not. Moreover, the hidden default thesis can also be incorporated into the framework. The two explanations "Harmfulness of the Conflict" and "Possible Compromise" describe different reasons why, in some examples, a default still seems to be missing or hidden.

With this differentiated approach, we can precisely elucidate the dynamics in different examples and expound the underlying patterns of intuitions without resorting to an artificial "one size fits all" approach. In cases like (IX) Primary Color, (V) Nixon, or (VIII) Carol in Germany, it is genuinely the plausibility of the compromise between conflicting propositions (whether the cup is orange, Nixon is politically in the middle, or Carol lives in between) that leads intuitively to suspension about the floating conclusion. Conversely, in instances such as (IV) Yacht, (VI) Murderer, or (II) Economy,[12] the intuitive non-acceptance of the floating conclusion stems from a lack of trust in any line of argument, as the credibility of the sources is compromised. For instance, in the (VI) Murderer case, the credibility of the witnesses is undermined by their conflicting statements about the weapon.

[12]Notably, (II) Economy can be explained by both referring to a potential compromise *and* the harmfulness of the conflict for the credibility of sources. This example demonstrates that there can be more than one reason to deviate from the default of accepting floating conclusions.

Furthermore, it is not my claim that the reasons for deviating from the default of accepting the floating conclusion are limited exclusively to the compromise and the harmfulness of the conflict thesis. Undoubtedly, there may be additional factors at play. However, this does not pose a challenge to my theory. The framework can easily be extended to incorporate numerous other reasons for deviating from the default. This can also be seen in Figure 3.11. It is straightforward to extend the flowchart to include a third or fourth test for deviation.

With this adaptable default framework, I manage to describe the underlying pattern of our intuitions regarding the floating conclusions. However, some intricacies of the possible examples point towards open questions. For example, consider the following example from practical reasoning. Imagine there was a robbery where jewelry was stolen. Later, the police stop a man in a car and find the stolen jewelry. The police have reason to believe that the occupant stole the jewelry. However, the man claims to have bought the jewelry from someone else. Both activities (stealing and the so-called "Hehlerei"/"heling," i.e., the purchase of stolen goods, which can colloquially also be called "fencing") are punishable in the Netherlands as well as in Germany. The German legal system, on the one hand, allows the suspect to be convicted of the crime with the lesser penalty since it is clear that they committed one of the two crimes. The Dutch legal system, on the other hand, apparently cannot convict the suspect unless there is evidence that clearly shows which of the crimes was committed.[13] The acceptance of the practical floating conclusion (the suspect is punishable) here does not depend on intuitions but on the legal system one is referring to.

The dependency on context and on stakes can also be visualized by other examples. While the conflict destroys the credibility of the witnesses in (IV) Yacht or (VI) Murderer, the conflict does not seem to destroy the credibility of Anna and Ben in the color examples. In these contexts, where

[13]For the German system this can be found in the following commentary to the criminal code: Fischer (2023) in "Becksche Kurz-Kommentare Strafgesetzbuch," 70th edition 2023, § 1, Rec. 32 cont. For the specific topic of fencing, "Hehlerei," see Fischer (2023, §1, Rec. 42a). This principle of German law (which is called "Wahlfeststellung") is apparently not present in the Dutch criminal law, according to Henry Prakken's personal communication with Hans Nijboer. I could not find written evidence supporting the absence of this principle.

they are simply telling us the color of a cup, we have no reason to be suspicious because the context offers us no reason why they should lie about the color of the cup.[14] Since whether or not we want to choose the third alternative, deviate from the default, and not accept the floating conclusion seems to depend heavily on the stakes and on the context, I consider it a hardly solvable challenge to represent the appropriate handling of floating conclusions in a formal logical system. Additionally, these examples imply a potential distinction between purely theoretical, epistemological reasoning and practical reasoning. Considering, moreover, that intuitions might become even more comparable when actions are involved, further research on the influence of practical reasoning for floating conclusions appears promising. However, for the investigation in this thesis, especially in the next Section 3.5, where floating conclusions are treated as a prototypical case of suspension, it is reasonable to confine the scope to epistemological examples.[15]

3.5 Floating Conclusions and Suspension

In this section, I will more explicitly explore the relationship between floating conclusions and suspension. Specifically, I will demonstrate how a particular approach to handling floating conclusions gives rise to a specific perspective on the logic of suspension. As previously mentioned, floating conclusions pose a challenge within the context of skeptical reasoning. Skeptical reasoning, in contrast to credulous reasoning,[16] is characterized by the acknowledgment that in the presence of a conflict between p and $\neg p$ (or when there are equally compelling reasons both in favor of and against p), suspension concerning p is deemed the only appropriate stance. In such cases, neither p nor $\neg p$ is considered to be affirmed (Pollock, 1995).

[14] Thanks to Joris Graff for coming up with this point.

[15] In this context, it is interesting to refer back to the discussion of evidentialism and pragmatic encroachment from Chapter 2. Particularly in the example of (IV) Yacht, it becomes evident that practical reasons play a role when deciding between acceptance and suspension on the floating conclusion.

[16] I will explain the distinction between skeptical and credulous reasoning in the subsequent chapters, Chapter 4 and Chapter 5. At this point, the precise definitions are not crucial.

When dealing with floating conclusions, the conflict takes on a slightly different form. As illustrated in the examples, the conflict does not arise between a proposition p and its negation, but rather between two propositions p and q. In logical terms, these two propositions are not contradictory but rather contrary — meaning that while both cannot be true simultaneously, both can be false simultaneously. In skeptical reasoning approaches, neither p nor q is then accepted. A skeptical reasoner would suspend judgment towards both p and q. A floating conclusion f is then a conclusion that follows from both p and q. Consequently, the question of whether we should accept floating conclusions is intricately linked to how one should epistemically handle the implications of a proposition about which we suspend judgment.

Framed in this manner, one might initially conclude that the most plausible response would be to suspend judgment about the consequence of a suspended proposition as well. If I suspend judgment about whether p is true, and I am aware that f logically follows from p, it might seem that (ceteris paribus) suspension towards f is the only applicable stance.[17] However, the situation is slightly more complicated. Not only does the relevant conclusion f follow (often even deductively) from p, but it also follows from another proposition q about which I also suspend judgment. In isolation, this might not significantly alter the situation. Nevertheless, when we consider that I possess some form of positive evidence for both propositions, p and q, and the only evidence against them is their mutual conflict, the situation starts to appear less straightforward. Does it still follow that only suspension towards f is warranted? In some trivial examples, this seems incorrect. If, for instance, the floating conclusion is a mathematical truth, i.e., $2 + 2 = 4$ and $p \rightarrow 2 + 2 = 4$ and $q \rightarrow 2 + 2 = 4$, suspension about the floating conclusion $2 + 2 = 4$ does not seem to be demanded.

To illustrate the situation with a non-trivial example, let us reconsider one of the examples involving a floating conclusion discussed earlier. For example, in the (I) Ice-Skating example presented in Figure 3.1, the

[17] Of course, there are circumstances in which I might have additional evidence for or against f, as discussed in Chapter 2. The situation I consider here is to be understood in a context where, prima facie, I lack other evidence for or against the conclusion. This is why I include the ceteris paribus clause.

conflicting propositions are p, "Brigt Rykkje is Norwegian" and q, "Brigt Rykkje is Dutch" and the floating conclusion is f, "Brigt Rykkje likes ice-skating." We saw that, although we suspend about p and q, at least intuitively we want to be justified in believing f. Hence, at least in some cases, suspending about regarding p and q is not, in itself, sufficient to determine whether I should believe f or also suspend judgment about f.

In my default framework, I argued that what determines whether we should believe or suspend about f is whether or not we think that there is a third alternative between p being true and q being false and vice versa. As mentioned earlier, this can be illustrated by restating the fact that p and q are contrary to each other. This implies that, in terms of the possible truth values of p and q, only $p \wedge q$ can be ruled out. Consequently, three additional possibilities are still viable:

(i) $p \wedge \neg q$,

(ii) $\neg p \wedge q$,

(iii) $\neg p \wedge \neg q$.

As stated above, the issue of whether a floating conclusion should be accepted can be reformulated by the question whether at least one of p and q is true, despite the conflict. This, in turn, can be reformulated as: Can we assume that either (i) $p \wedge \neg q$ or (ii) $\neg p \wedge q$ is the case? If yes, the conclusion f should be accepted. If we do not assume that (i) $p \wedge \neg q$ or (ii) $\neg p \wedge q$ is the case, and take the third alternative, i.e., (iii) $\neg p \wedge \neg q$, to be more likely, we should not accept f and suspend about it.

Certainly, we typically cannot definitively answer this question. The logical circumstances do not furnish us with any means to address this issue conclusively. However, given that we are engaging in common-sense, defeasible reasoning, we can still inquire whether there is more convincing evidence in favor of (i) or (ii), or in favor of (iii). As this evidence is typically not represented within our formal theory, the existence of floating conclusions suggests the necessity of examining the broader evidential context to effectively address certain phenomena. Furthermore, as demonstrated in the example of (IV) Yacht, there are instances where it

seems necessary to even consider practical reasons.

This implies that the information contained in a formal theory may occasionally prove insufficient, casting doubt in the closed-world assumption, which is captured in non-monotonic frameworks like default logic, to which I will come back in Chapter 4.

A general observation concerning the logic of suspension that might be drawn from these findings is the following. Floating conclusions provide us with an indication of how we can continue reasoning with suspended propositions. In situations in which we suspend about some propositions $p_1, \ldots p_n$ and in which for all those p_i we have $p_i \rightarrow f$, we have to check whether we find their common disjunction[18] $\bigvee p_i$ more likely (then we can believe f) or whether we find the conjunction of their negations $\bigwedge \neg p_i$ more likely (then we should suspend upon f).

Of course, this rule is only applicable to the rather specialized type of situation. Exploring other related scenarios is indeed an interesting topic for further research. For instance, consider situations where two conflicting propositions are not contrary (as in floating conclusions) but rather contradicting, say p and $\neg p$. In such cases, one might be inclined to accept a proposition f that follows from both p and $\neg p$ because, logically, either p or $\neg p$ must hold. However, in the realm of non-monotonic reasoning, this claim may appear too simplistic. Non-monotonic inferences do not guarantee conclusions given the premises. Consequently, the conflict between the lines of arguments might ultimately defeat the arguments, particularly in situations involving testimony. Still, it appears that in almost every natural example where a proposition follows from both p and its negation, this proposition is typically accepted. This may be explained by the fact that the inferences are always monotonic/deductive rather than non-monotonic when a proposition follows from both p and $\neg p$. Consider an adapted version of the Murderer example, where A asserts that Peter killed the victim (p) and B insists that Peter did not kill the victim ($\neg p$). Anything that reasonably follows from both Peter killing the victim and Peter not killing the victim appears to follow non-defeasibly. Examples might include presuppositions like "Peter exists." Consequently, we end up

[18]Note that option (i) $p \wedge \neg q$ *or* (ii) $\neg p \wedge q$ translates to an *exclusive disjunction* of p and q. It is sufficient, though, to check the disjunction, since $p \wedge q$ is ruled out from the beginning.

in a classical dilemma scenario where the conclusion is deductively valid.

Nevertheless, from the specific investigations about floating conclusions we can distill already a more general rule about reasoning based on suspended beliefs: If we want to determine the appropriate doxastic stance towards a proposition that follows from other suspended propositions, we should look left and right and consider the broader evidence base. This idea will recur in Chapter 4, when the logic of suspension is analyzed more precisely within a framework of adapted default logic.

3.6 Conclusion

In this chapter, I considered the phenomenon of floating conclusions as a prototypical case of suspended judgment in defeasible reasoning. The question about the acceptability of floating conclusions can be reformulated as the question of how we should doxastically behave towards a proposition that follows from two distinct propositions we suspend upon and that are in conflict with each other.

I presented an overview of different examples of floating conclusions from the literature and extended the list with new examples. Furthermore, I examined different hypotheses that aim to explain our non-uniform intuitions about whether floating conclusions should be accepted or not and tested them via my examples. I argued that no hypothesis succeeds in explaining our intuitions concerning all the presented examples. Instead, I presented an approach that manages to provide an overarching explanation for the acceptability of floating conclusions. The approach starts with the basic idea that floating conclusions ought to be accepted by default. The framework then allows several reasons to deviate from the default and not to accept the floating conclusion. These reasons come into play when there seems to be a third alternative besides one of the conflicting propositions being true and the other one false or vice versa.

I presented two possible ways how this alternative can arise. If there is a compromise between the conflicting propositions from which the floating conclusion does not follow or if the conflict is harmful to the sources of information, one can deviate from the default and refrain from accepting

the floating conclusion. We saw that these two reasons exhaustively cover and explain all the examples investigated in this chapter. With this, I managed to provide a framework that explains our varying intuitions but is still flexible enough to allow for further examples and explanations.

Still, there are open topics that have to be considered in further research. As illustrated by the differing treatment of the "fencing or stealing" case in the Dutch and German legal systems, examples from practical reasoning cannot be adequately addressed yet.

Nevertheless, when it comes to investigating the behavior of suspension in non-monotonic reasoning, specifically through the prototypical case of floating conclusions, it appears reasonable to confine our focus to purely epistemic reasoning. As demonstrated with examples of floating conclusions, continuing reasoning with suspended propositions is not straightforward. Suspension, unlike belief, lacks clear guidelines on how the doxastic status of premises translates to conclusions. In particular, our examination of the acceptability of floating conclusions suggests that when a conclusion derives from two different propositions, and we suspend judgment about both, it becomes necessary to consider the broader evidential context to determine the appropriate doxastic response to the conclusion. The framework presented indicates that if we possess more evidence favoring the disjunction of the two propositions over the conjunction of their respective negations, we can accept, i.e., believe the conclusion. If the evidence leans towards the conjunction of the two negations, suspension regarding the conclusion is appropriate. The case of floating conclusions generally underlines the importance of considering the overall body of evidence when determining the proper doxastic stance towards inferences from suspended propositions.

In the subsequent Chapter 4, we will see that this observation also holds for the broader cases, e.g., when the conclusion follows only from one proposition about which we suspend judgment.

Here, I will broaden the perspective beyond the specific case of floating conclusions to more general instances where implications from suspended propositions are made. This comprehensive exploration will be undertaken by employing the formal framework of default logic. Notably, when

formalized within a default theory, floating conclusions will highlight the constraints in modeling suspension, illustrating that not every facet of suspension is representable.

Chapter 4

Default Logic

Contents

4.1 Introduction

In this chapter, I will expand my logical investigations beyond the phenomenon of floating conclusions and delve into other scenarios involving reasoning with suspended propositions within the realm of logic-based AI. To achieve this, I will approach these situations in a more formalized manner, employing a formal framework. Non-monotonic logics are particularly well-suited to provide a framework for logic-based AI, as they capture everyday reasoning well. In this chapter, I will explore the non-monotonic framework of default logic and show how suspension can be incorporated into it.

Default logic, along with autoepistemic logic and circumscription, stands as one of the three most prominent logical frameworks for non-monotonic reasoning (Minker, 2000). Autoepistemic logic serves the purpose of formalizing self-knowledge, specifically knowledge about one's own knowledge. It introduces an operator, $\Box$ (also known from modal logic), into propositional logic, indicating that any formula following this operator is known. The logic incorporates default beliefs, assumed to be valid until counter-evidence emerges. This property makes it non-monotonic because new information can result in the retraction of previously established beliefs. Autoepistemic logic primarily concentrates on self-knowledge and thus may be somewhat limited when attempting to represent the broader scope of non-monotonic reasoning (Straßer and Antonelli, 2019).

Circumscription, dating back to John McCarthy (McCarthy, 1980, 1993), stands as one of the earliest formalisms in non-monotonic reasoning. It operates on the idea that each non-monotonic reasoning rule, in addition to the regular antecedent, includes an extra condition, indicating that this specific case is not "abnormal." For instance, in order to non-monotonically infer that a bird can fly, a rule might include three predicates: the predicate for being a bird, the predicate for flying, and the predicate for being abnormal. This rule can be formulated as $Bird(x) \wedge \neg Abnormal(x) \rightarrow Fly(x)$. In circumscription, the goal is to minimize the abnormal predicate, assessing the normal entailment in the model using the minimal abnormal predicate. The statement "Tweety flies" can then be concluded (given a certain knowledge base) if it can be deduced

from the model with the minimal abnormal predicate. Circumscription thus streamlines the definition of semantic consequence to those models incorporating the minimal abnormal predicate (Horty, 2001).

Default logic shares similarities with circumscription, presenting rules (defaults) with two antecedents: a regular antecedent and an exception condition. The default rule asserting that birds fly can be similarly expressed as $Bird(x) \wedge \neg Exception(x) \rightarrow Fly(x)$. Once again, the exception condition is what imparts non-monotonicity to the system. Thereby, default logic encapsulates common-sense and everyday reasoning, as highlighted by Delgrande and Schaub (2000). A specific subclass of defaults, known as normal defaults, involves defaults where the negation of the conclusion serves as the exception to the rule. In this scenario, the exception condition of the above rule is $\neg Fly(x)$. This can be interpreted simply as: If x is a bird, then x can fly, unless it can't. In simpler terms, birds fly *by default*. With these *normal default theories*, we possess a straightforward and uncomplicated method for representing non-monotonic rules. The framework for constructing normal default theories outlined by Horty (2011), which I will use in this chapter, provides a contemporary and intuitive foundation for the forthcoming investigations on suspension.

Horty's default logic considers propositions as the fundamental elements of the framework. These propositions are combined to compose defaults, with one proposition serving as the premise and another serving as the conclusion of a default. While Horty (2011) primarily focuses on deontic reasoning (Horty, 1994), which deals with reasons for actions, the original interpretation of default logic by Reiter (1980) is epistemic. Given the focus of my work on epistemic matters, I will interpret default logic from an epistemic perspective. In this context, defaults represent reasons for inferring one proposition from another, indicating an evidential relationship between the premise and the conclusion.

Another feature of default logic is that it operates under the *closed-world assumption* (Reiter, 1980). The closed-world assumption states that all propositions that are *not known to be true* are considered false, or conversely, everything true is also known to be true. By adopting this assumption, default logic allows for drawing conclusions based

on incomplete knowledge or information. As noted by Reiter (1980), particularly in computational contexts, the closed-world assumption enables concentrating purely on the explicit expression of positive information while disregarding negative information. This simplifies the representation of the domain under consideration. Reiter (1980, p. 84) illustrates this with an example of an airline database queried about whether "Air Canada flight 113 connects Vancouver with New York." If the system cannot deduce this information from its rules and knowledge base, it will output "no" to the question. Such an output is only possible within a closed-world setting, where it is assumed that all relevant airline information is stored in the system. Statements that cannot deduced from the system are considered to be false.

While this principle is certainly reasonable from this computational standpoint, it inherently contradicts the concept of doxastic neutrality. It assumes that all relevant information is present within a default theory, and that this information is complete. Consequently, if a proposition cannot be derived from the default theory, it is automatically considered false. This binary approach adopted by default theory leaves no room for neutrality or any intermediate state between true and false.[1] The proposal of this chapter, to include the possibility of suspension into default logic, can serve as a tool to overcome this shortcoming and provide room for neutrality.

Moreover, default logic is well suited for examining the possibility of suspension in logic-based AI because default logic allows for the explicit modeling of conflicts, which are typical scenarios where the need for suspension arises. In this chapter, I will illustrate how the current framework of default logic represents and handles conflicts and explain why this representation is inadequate. I have developed a way to incorporate suspension into the framework of default logic as a mechanism for effectively addressing conflicts and representing neutrality. I will also describe the logical profile of suspension within this adapted theory. Consequently, this chapter begins with a critical diagnosis but is ultimately constructive in character.

[1]This aspect will be further highlighted in Subsection 4.3.4 when the framework is examined in cases of floating conclusions, and in Subsection 5.3.2 of Chapter 5 on argumentation theory. Through the translation of default logic into argumentation theory, this "black or white" nature of default logic becomes apparent again.

I will start with Section 4.2 by introducing fundamental notions and definitions from default logic. Examples will be provided to illustrate these definitions, and I will address the topic of conflicts. This will involve explaining how conflicts in default logic arise and delving into the state-of-the-art methods for dealing with conflicts.

In Section 4.3, I will present the constructive part of this chapter. Here, I will introduce a logic that incorporates suspension, ignorance, and higher-order reasoning within the framework of default logic. After justifying this specific approach (Subsection 4.3.1), I will present the relevant definitions for adjusting default logic (Subsection 4.3.2) and the logical principles resulting from this adjustment (Subsection 4.3.3). The adjustments and logical rules will be further illustrated through additional examples.

I will particularly focus on the logical profile of suspension, demonstrating how suspended propositions logically interact with other propositions within my proposed framework. To do this, I will show how we can continue reasoning from suspended propositions in general, and in particular for cases of floating conclusions (Subsection 4.3.4), as investigated in Chapter 3. The handling of floating conclusions within my adapted framework will be put to the test using the observations on floating conclusions from the Chapter 3.

While all these considerations are made in an epistemic interpretation, I will briefly outline how my framework can be interpreted deontically as well (Subsection 4.3.5).

4.2 Default Logic - Background

4.2.1 Definitions

Default logic is a formal framework of non-monotonic reasoning that goes back to Reiter (1980). The version of the framework that I consider in the following was detailed by Horty (2011) and corresponds to the subclass of default logic of what is called *normal default logic* in the original of Reiter (1980). In Reiter's default logic, a default is composed of a regular premise, an exception condition, and a conclusion. It should be interpreted as follows: If the premise is true, and the exception condition is not satisfied, then conclude the specified conclusion. Normal defaults are a specific

type of default in which the negation of the conclusion itself serves as the exception condition, eliminating the need to explicitly state the exception condition.

In default logic, we are dealing with *default theories* that help us to describe how to continue reasoning from a knowledge base via some reasoning rules (defaults) to further knowledge in a defeasible matter. For this, Horty starts with a logical language (in our case a propositional language) $\mathcal{L}$ based on which default theories can be defined.

Definition 4.1 (Default Theory)**.** A default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ consists of $\mathcal{W}$, which is a set of propositions in $\mathcal{L}$, and a set of defaults $\mathcal{D}$. The set $\mathcal{W}$ represents a set of premises, world descriptions, or a knowledge base. A default $\delta \in \mathcal{D}$ is a reasoning rule and has the form $\delta : a \rightarrow b$, with $prem(\delta) = a$ being the premise and $con(\delta) = b$ (both propositions in $\mathcal{L}$) being the conclusion of the default.[2]

In order to reason with the defaults, it is important for Horty to grade them according to their strength. For this, Horty (2011, p. 22) extends the idea of a default theory to a *prioritized default theory.*

Definition 4.2 (Prioritized Default Theory)**.** A prioritized default theory $\Delta = \langle \mathcal{W}, \mathcal{D}, < \rangle$ is a default theory with a priority relation $< \subseteq \mathcal{D} \times \mathcal{D}$ among the defaults in $\mathcal{D}$, with $\delta_i < \delta_j$ meaning that the default δ_j has to be prioritized over the default δ_i. The relation $<$ is required to be a strict partial ordering.

Horty introduces a *scenario* $\mathcal{S}$ of a default theory, which is a subset of the set of defaults, $\mathcal{S} \subseteq \mathcal{D}$. He then defines the premises and the conclusions of a whole scenario as the premises or the conclusions of the defaults that the scenario consists of, i.e, $prem(\mathcal{S}) = \{prem(\delta) : \delta \in \mathcal{S}\}$ and $con(\mathcal{S}) = \{con(\delta) : \delta \in \mathcal{S}\}$.
Then, one goal is to find so-called *proper scenarios*, which are meant to capture those defaults a reasonable agent can accept. Given any original scenario, $\mathcal{S}$, a proper scenario (with respect to $\mathcal{S}$) is defined as the set of all

[2]Strictly speaking, $\mathcal{L}$ does not consist of propositions but of formulas or sentences. In alignment with Horty (2011, p. 16), I will speak of propositions instead of sentences, as the goal is to speak of a premise of a default as a *reason* for the conclusion. Statements like "the proposition a" have to be interpreted as referring to the proposition that is expressed by the sentence a.

defaults that are *bound* by $\mathcal{S}$. To define bound defaults, it is necessary to establish three further key definitions of *triggered*, *defeated*, and *conflicted* defaults first. A default δ is then bound by a scenario $\mathcal{S}$, if it is *triggered* by the scenario, *not conflicted* with the scenario, and *not defeated* by the scenario (Horty, 2011, p. 27–32).

Definition 4.3 (Triggered Default)**.** Given a prioritized default theory $\Delta = \langle \mathcal{W}, \mathcal{D}, < \rangle$ and a scenario $\mathcal{S} \subseteq \mathcal{D}$, a default $\delta_i \in \mathcal{D}$ is *triggered* by $\mathcal{S}$, iff the following two conditions hold *both*:

(i) $\mathcal{W} \cup con(\mathcal{S}) \vdash prem(\delta_i)$

(ii) $\delta_i \notin Excluded_{\mathcal{S}}$, i.e., $\mathcal{W} \cup con(\mathcal{S}) \nvdash Out(d_i)$.

Definition 4.4 (Defeated Default)**.** Given a prioritized default theory $\Delta = \langle \mathcal{W}, \mathcal{D}, < \rangle$ and a scenario $\mathcal{S} \subseteq \mathcal{D}$, a default $\delta_i \in \mathcal{D}$ is *defeated* (rebutted) by $\mathcal{S}$ iff there is some default δ_j that is triggered by $\mathcal{S}$ with $con(\delta_j) = \neg con(\delta_i)$ and $\delta_i < \delta_j$.

Definition 4.5 (Conflicted Default)**.** Given a prioritized default theory $\Delta = \langle \mathcal{W}, \mathcal{D}, < \rangle$ and a scenario $\mathcal{S} \subseteq \mathcal{D}$, a default $\delta_i \in \mathcal{D}$ is *conflicted* in the context of $\mathcal{S}$ iff $\mathcal{W} \cup con(\mathcal{S}) \vdash \neg con(\delta_i)$.

The concept of being triggered involves two conditions. A triggered default is generally one that is *activated* by the scenario. If the scenario already contains a default arguing for a proposition a, or if a is already present in the world descriptions $\mathcal{W}$, then a default δ with $prem(\delta) = a$ is preliminarily triggered by the scenario. This condition is defined by (i) in Definition 4.3. However, there are situations in which a default can lose its triggered status. This occurs when a default is *defeated before* it can be *properly triggered* due to what are known as *undercutting defeaters*. This is captured in (ii) in Definition 4.3. A default δ_i is undercut by another default δ_j if δ_j entails that the connection between the premise and the conclusion of δ_i cannot be justifiably drawn.[3] Horty (2011, p. 124–125) extends his definition of triggered defaults to include the requirement

[3] I will explain the concept of undercutting in more detail and provide an example in Subsection 4.2.2.

that a default should not be excluded (due to undercutting defeaters) in order to be triggered. To achieve this, the background language $\mathcal{L}$ must be enriched with a unique name d_i for every default δ_i. Moreover, the background language should contain a predicate Out, which, when applied to a name of a default, $Out(d_i)$, denotes that the default δ_i is excluded (due to undercutting). A default δ_i is then excluded (i.e., $\delta_i \in Excluded_{\mathcal{S}}$ in the context of a scenario $\mathcal{S}$) if $\mathcal{W} \cup con(\mathcal{S}) \vdash Out(d_i)$. Hence, the definition of triggered defaults rules out the possibility of undercutting defeat.

In contrast, Definition 4.4 covers cases of *rebutting defeat*. A default δ is considered rebutted (or defeated due to rebutting) when there exists another default, prioritized higher, that is triggered by the scenario and entails the opposite conclusion of δ. This reflects the idea that in conflicts between two defaults that are comparable with respect to a priority order the higher-prioritized default should be chosen.

Finally, there is Definition 4.5 of conflicted defaults. Definition 4.5 is similar to Definition 4.4 of defeated defaults, but also covers cases where there is no priority relation between the conflicting defaults. A default δ_i is conflicted in the scenario $\mathcal{S}$ if another default $\delta_j \in \mathcal{S}$ included in the scenario has the opposite conclusion. For instance, if $\delta_i : a \rightarrow b$ and there is another default $\delta_j \in \mathcal{S}$ with $\delta_j : c \rightarrow \neg b$, then δ_i is considered conflicted within the context of scenario $\mathcal{S}$.[4] This can also occur if $\neg b$ is already within $\mathcal{W}$. In one sense, Definition 4.5 is stronger than Definition 4.4 because it excludes not only defaults conflicted with *higher-prioritized* defaults but also defaults in conflict with *any other* default within the scenario, regardless of priority. In another sense, it is less strong because it only considers defaults *included* in the scenario as potential sources of conflict, whereas Definition 4.4 considers all defaults *triggered* by the scenario for the definition of defeated defaults.

Given Definitions 4.3, 4.4, and 4.5, we can define *bound* defaults.

Definition 4.6 (Bound Default)**.** Given a prioritized default theory $\Delta = \langle \mathcal{W}, \mathcal{D}, < \rangle$ and a scenario $\mathcal{S} \subseteq \mathcal{D}$, a default $\delta_i \in \mathcal{D}$ is *bound* by $\mathcal{S}$, iff

[4]Note that according to Definition 4.6 (ii), the default δ_j itself would only be conflicted by the scenario if δ_i (or another default with the conclusion b) were included in the scenario already. If $\mathcal{S}$ is a scenario that does not contain δ_i, then δ_j is not conflicted in the context of $\mathcal{S}$, but δ_i is.

the following three conditions are met: δ_i is

(i) *triggered* by $\mathcal{S}$,

(ii) *not conflicted* in the context of $\mathcal{S}$,

(iii) *not defeated* by $\mathcal{S}$.

Building upon the definition of bound defaults, Horty (2011) goes on to define a *proper scenario* as one that precisely encompasses all its bound defaults.[5] Proper scenarios can be conceptualized as maximally consistent subsets of $\mathcal{D}$ when consistency is defined through the notion of binding. As the following definition illustrates, a scenario may fall short of being a proper scenario in two distinct ways. First, it might *falsely include* defaults that are not bound and thus fail to be proper. Second, it could *fail to include* all defaults that are in fact bound. This second way is what characterizes proper scenarios as *maximal* subsets.

Definition 4.7 (Proper Scenario)**.** Given a (prioritized) default theory $\Delta = \langle \mathcal{W}, \mathcal{D}, < \rangle$, a scenario $\mathcal{S} \subseteq \mathcal{D}$ is proper iff it is stable, i.e., $\mathcal{S} = \mathit{Bound}(\mathcal{S})$.

Once we found the proper scenarios of a default theory, an *extension*[6] of a default theory is then defined as the logical closure of a set consisting of the propositions in $\mathcal{W}$ and the conclusions of the defaults of a proper scenario $\mathcal{S}$ (Horty, 2011, p. 32). Entailment in default logic is entailment with respect to a specific extension. A proposition is considered entailed by an extension if it fulfills classical entailment criteria, meaning it is either part of the extension itself or can be logically inferred from the propositions within the extension. An extension consists of the world descriptions and the conclusions of the acceptable defaults and whatever logically follows from that. Interpreting proper scenarios as sets of defaults the agent can accept, extensions can thus be seen as reasonable belief sets an agent can have.

Definition 4.8 (Extension)**.** Given a (prioritized) default theory $\Delta = \langle \mathcal{W}, \mathcal{D}, < \rangle$, a set $\mathcal{E}^{\mathrm{D}}$ of propositions from $\mathcal{L}$ is an extension of Δ iff

[5] In certain specific examples, the definition for proper scenarios may need to be adjusted, as noted in Horty (2011, p. 221). These cases can involve defaults that are, in a way, "self-triggered" while being not grounded in $\mathcal{W}$ (Horty, 2011, p. 32). However, for the purposes of this chapter, I will adhere to the simpler definition.

[6] Since extensions will occur in argumentation theory as well, we will denote an extension in default logic with an upper index D, i.e., $\mathcal{E}^{\mathrm{D}}$.

there is some proper scenario $\mathcal{S}$ and $\mathcal{E}^{\mathrm{D}} = Th(\mathcal{W} \cup con(\mathcal{S}))$, where Th denotes the logical closure of a set.

For reasons of simplicity, I will, in the following, sometimes leave the Th operator aside and explicitly only mentioning those propositions in $\mathcal{E}^{\mathrm{D}}$ that are relevant to the given example.

4.2.2 Examples

To illustrate the forthcoming examples, I will use the same illustration technique that I employed in Chapter 3, consistent with the approach of Horty (2002).[7] Small letters will be used as abbreviations for propositions. Arrows will be employed to depict arguments, with double arrows signifying deductive, monotonic reasoning, and single arrows indicating defeasible inferences (defaults). If a default arrow is crossed out, it denotes that the conclusion of the default is the *negation* of the proposition the arrow leads to. A double-sided crossed-out arrow signifies a conflict between two propositions, and the symbol **T** represents "truth." Propositions in the knowledge base $\mathcal{W}$ are represented as deductively inferred from **T**.

Let us begin with a basic example of Horty (2011, p. 26) that goes back to the famous bird Tweety, which is known as the standard example in the literature on non-monotonic reasoning, see Figure 4.1.

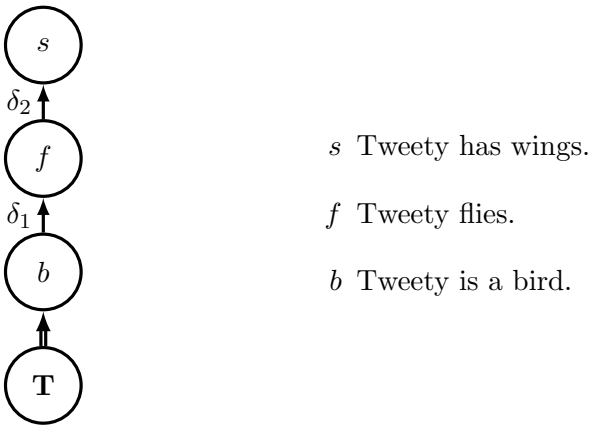
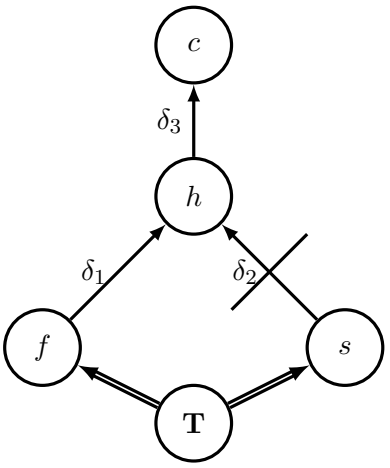


Figure 4.1: Default logic example of Tweety the bird, which serves as a standard example in non-monotonic reasoning.

[7] Horty uses capital letters as abbreviations for propositions. Capital letters abbreviate arguments though in Chapter 5.

In this example, let the priority relation first be empty. Then we have only one proper scenario in this default theory, which is $\mathcal{S} = \{\delta_1, \delta_2\}$. δ_1 is a default that is triggered already from the empty set since its premise is in $\mathcal{W}$. Once δ_1 is included, δ_2 is triggered, too. Both defaults cannot be in conflict or be defeated by any other default within this theory, so the triggered defaults constitute the bound defaults. The extension of this theory is hence $\mathcal{E}^{\mathrm{D}} = \{b, f, s\}$.

The following example, which I will call the *food detector example*, is slightly more complicated. In this chapter, I will revisit this example several times.



c Jam should be served in school cafeterias.

h Jam is healthy.

f Jam contains a lot of fruit.

s Jam contains a lot of sugar.

Figure 4.2: Default logic example of the food detector in an extended version about jam being served in school cafeterias.

In the example of Figure 4.2, the defaults δ_1, δ_2, and δ_3 are included in $\mathcal{D}$. The propositions f and s are included in the knowledge base. Let us first assume that there is no priority ordering among the defaults. The task is to find the proper scenarios of this default theory. In total, there have to be $2^3 = 8$ combinations for possible scenarios.

Analogously to Horty, we can start by considering the empty scenario $\mathcal{S}_1 = \emptyset$. In this scenario both δ_1 and δ_2 are triggered because $W \cup con(\mathcal{S}_1) = \{f, s\} \vdash prem(\delta_1)$ and also $W \cup con(\mathcal{S}_1) \vdash prem(\delta_2)$ (and there is no undercutting going on). So far, neither of the defaults is conflicted by $\mathcal{S}_1$, simply because $\mathcal{S}_1$ does not include any defaults yet. Neither are they defeated. Both defaults are bound by $\mathcal{S}_1$. Hence $\mathcal{S}_1$ is not proper, since it does not include all the defaults bound by it. We can

extend the scenario gradually.

For example, we can consider $\mathcal{S}_2 = \{\delta_2\}$. In this scenario δ_2 is triggered, not conflicted, not defeated, and hence bound. What about the other two defaults? We see that δ_1 is triggered by $\mathcal{S}_2$, but it is also conflicted because $con(\delta_1) = H$ and $W \cup con(\mathcal{S}_2) \vdash \neg H$, hence not bound. The third default δ_3 is not even triggered by the scenario, since its premise is not included in $W \cup con(\mathcal{S}_2)$. Thus, the scenario $\mathcal{S}_2$ contains all of its bound defaults (and only those) and is thereby a proper scenario.

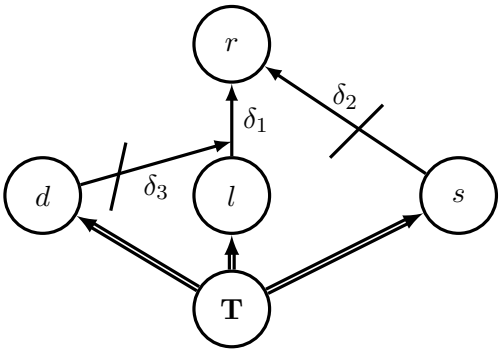
The scenario $\mathcal{S}_3 = \{\delta_1\}$ is not proper as there is the default δ_3 that is not included in $\mathcal{S}_3$, although bound by $\mathcal{S}_3$ (it is triggered by $\mathcal{S}_3$ and not conflicted or defeated by $\mathcal{S}_3$). The scenario $\mathcal{S}_4 = \{\delta_1, \delta_3\}$, in contrast, is proper since both defaults are bound and the only other default δ_2 is not bound because it is conflicted in the context of $\mathcal{S}_4$. $\mathcal{S}_5 = \{\delta_3\}$ is not proper because δ_1 and δ_2 are also bound by it and $\mathcal{S}_6 = \{\delta_1, \delta_2, \delta_3\}$ is not proper because both δ_1 and δ_2 are conflicted and thus not bound.

The two missing scenarios are $\mathcal{S}_7 = \{\delta_1, \delta_2\}$ and $\mathcal{S}_8 = \{\delta_2, \delta_3\}$. $\mathcal{S}_7$ is not proper as both δ_1 and δ_2 are conflicted and thus not bound, and δ_3 is bound but not included. $\mathcal{S}_8$ is not proper as δ_3 is not bound but included.

We end up with two proper scenarios for this default theory: $\mathcal{S}_2 = \{\delta_2\}$ and $\mathcal{S}_4 = \{\delta_1, \delta_3\}$. The extensions of this default theory are respectively $\mathcal{E}_2^{\mathrm{D}} = \{s, f, \neg h\}$ and $\mathcal{E}_4^{\mathrm{D}} = \{s, f, h, c\}$.

If we consider this default theory with a non-empty prioritization, we get a different result. For example, if $\delta_2 < \delta_1$, $\mathcal{S}_2$ is no longer a proper scenario, since δ_2 is no longer bound by $\mathcal{S}_2$. There is a default, δ_1, which is triggered by the scenario $\mathcal{S}_2$ (its premise is included in $\mathcal{W}$), which is prioritized over δ_2, and whose conclusion is the negation of the conclusion of δ_2. Conversely, if $\delta_1 < \delta_2$, $\mathcal{S}_2$ would be the only proper scenario left.

Another example that goes back to Pollock (1995) and shows the possibility of undercutting can be found in Horty (2011, p. 127).



r The object is red.

d I have taken a drug that makes everything look red.

l The object in front of me looks red.

s A reliable source said that the object is not red.

Figure 4.3: Default logic example of an undercutting defeater.

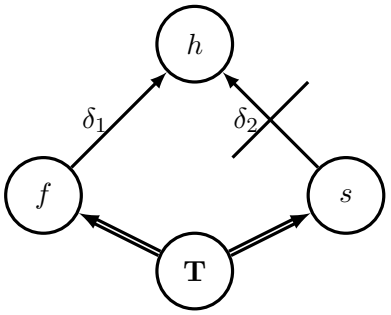
In this example, there is a (rebutting) conflict between the defaults δ_1 and δ_2 since they argue for opposite conclusions. Moreover, we have an undercutting defeat. The default δ_1 states that the object in front of me is red because it looks red. The proposition d states that I have taken a certain drug that makes *everything* look red. This does not defeat the first default in a rebutting way, since we cannot conclude $\neg r$ from d. Just because everything looks red, this does not mean that there cannot be things that actually are red. Rather, the proposition d shows that the connection or inference relation between l and r seems no longer valid. Usually, we can infer that something is red from the fact that it looks red. With the additional information d that everything looks red, this inference is no longer valid. The default δ_3 states that d is a reason for rejecting the default δ_1, i.e., $\delta_3 : d \rightarrow Out(d_1)$.

If we apply the definitions of Horty (2011) to this example, we can see that the only proper scenario is $\mathcal{S}_1 = \{\delta_2, \delta_3\}$. δ_3 is a default that is necessarily included in every scenario, since it is triggered and not possibly conflicted or defeated by any other default.[8] Once δ_3 is in a scenario, δ_1 can no longer be in it, as δ_3 argues for $Out(d_1)$. Hence, δ_1 is excluded and therefore not even triggered. From this, it follows that in all possible scenarios, δ_2 is not conflicted or defeated as δ_1 is not included. Of course, δ_2 is triggered as $s \in \mathcal{W}$ and δ_2 is not excluded. Hence, the only possible extension of this example is $\mathcal{E}^{\mathrm{D}} = \{l, s, d, \neg r, Out(d_1)\}$.

[8]Note that even if there was a priority ordering and both other defaults were higher prioritized than δ_3, it would not be rebutted since the other defaults do not argue for the negation of the conclusion of δ_3.

4.2.3 Conflicts in Default Logic

From the previous examples, it is already clear how conflicts can arise in default theories. The conflicts I will look at in the following are conflicts that arise when one default argues for a conclusion c and another default argues for $\neg c$. The example I will mainly use for this is the food detector. In Figure 4.2, an extended version (with an additional default from h) was introduced. For the purposes of the following investigations, the following short version of the food detector is sufficient.



h Jam is healthy.

f Jam contains a lot of fruit.

s Jam contains a lot of sugar.

Figure 4.4: Default logic example of the food detector in the short version.

In this example, we have two world descriptions, namely that jam contains a lot of sugar, and that jam contains a lot of fruit. Additionally, we have two defaults. One default leads from the premise that jam contains a lot of fruit to the conclusion that jam is healthy. The second default leads from the premise that jam contains a lot of sugar to the contradicting conclusion that jam is not healthy. Thus, we end up with a conflict and two proper scenarios $\mathcal{S}_1 = \{\delta_1\}$ and $\mathcal{S}_2 = \{\delta_2\}$. Although the conflict might not be directly visible at the level of proper scenarios, the extensions following from the two scenarios reveal the conflict clearly. While the extension following from the first proper scenario consists of f, s, h, the extension following from the second one consists of $f, s, \neg h$. One extension contains the proposition h, while another contains its negation, $\neg h$. It is not clear how to deal with a situation in which conflicting results, i.e., h and $\neg h$, occur among the different extensions. This is especially worrisome as the extensions are interpreted as the reasonable positions an agent should adopt when confronted with the default theory.

As demonstrated in Example 4.2, conflicts could be resolved by

assigning higher priority to one of the two defaults resulting in accepting either h or $\neg h$. However, as I will discuss in more detail in Subsection 4.3.1, I will only consider theories without priorities here. In many conflict situations, there appears to be no intuitive prioritization between the conflicting rules, so resolving conflicts through priorities might appear like an artificial, ad-hoc solution. While it may be beneficial to introduce a priority relation between defaults in some cases, my proposal aims to provide a more general solution that does not rely on priorities. Nevertheless, priorities could potentially be incorporated at a later stage. For the purposes of what follows, the priority relations are assumed to be empty sets.

Dealing with Conflicts: State-of-the-art

Default theories with multiple different extensions pose the challenge of how to interpret them and how to continue reasoning. A reasoning agent is supposed to be able to form their doxastic position according to an extension resulting from a proper scenario. However, when there is more than one extension, it is not clear which of these extensions the agent should base their doxastic position on. Horty (2011, p. 34) describes this problem as the problem of finding the consequences (which are supposed to be a set of propositions one should accept) of a given default theory and discusses three possible ways to deal with it.

One possibility, which he calls the *choice-option*, consists in arbitrarily choosing one of the given extensions. The underlying idea here is that, although two proper scenarios might be in conflict with each other, taken alone they both represent a reasonable and justified position an agent can adopt. Choosing one leads to having one possible reasonable position.

This possibility appears highly unsatisfactory. While each distinct scenario might represent a reasonable position, there is no compelling overarching, all-things-considered reason to prefer one scenario (and consequently, one extension) over another. All things considered, these two (or more) extensions seem equally reasonable. Unless the agent has practical (or potentially irrational) motivations for choosing one scenario over the other, such as an intuitive inclination that one default should take precedence over another, or a personal benefit tied to the truth

of one proposition, the choice becomes arbitrary. Within an epistemic interpretation, such a choice is psychologically implausible. If the agent is aware of the multiple extensions and the conflict among them, they will often find it impossible to favor one extension over another. Such a demand would lean towards doxastic voluntarism, asserting that a subject can arbitrarily decide what to believe and disbelieve, which is highly implausible (Verena, 2019).

The second option consists in what Horty calls the *credulous option*. The idea is that a credulous or trustful subject might simply believe any proposition that is supported by at least one proper scenario, i.e., in at least one extension. Each proposition that belongs to an extension is a proposition that is *somehow justified* as the extensions result from scenarios that are proper. Proper scenarios are meant to represent reasonable, justified stances. This option suggests that the consequences of a default theory are found in the *union* of all extensions.

However, this approach will often lead to inconsistent consequences, as exemplified by Example 4.4. Here, the set of the consequences of the default theory according to the credulous approach would be $\{f, s, h, \neg h\}$, containing both the proposition h and its negation. When the subject bases their doxastic attitudes on this union of the extensions, they will end up in a situation of believing both h and $\neg h$, i.e., believing both that jam is healthy and that it is unhealthy. Ending up in contradictory doxastic situations will leave an artificial reasoner with no guidance on what to follow from their beliefs, how to continue reasoning, and how to act based on their beliefs. This cannot be the desired outcome of a default theory.

Horty (2011, p. 37) describes a possible way to circumvent the explicit contradiction within the credulous approach. Instead of taking the propositions of the extensions to be the consequences of the default theory, one might as well take the consequences to be statements *about* these propositions. For this, one can introduce an operator $\mathcal{B}$ on propositions p, which can, for example, be interpreted as believability. The consequences of a default theory would then consist of sentences of the form $\mathcal{B}(p)$ for every proposition p belonging to some extension. This would allow us to somehow include both propositions h and $\neg h$, that Jam is healthy, and that Jam is not healthy, without thereby having an explicit contradiction

of the form $h \wedge \neg h$ in our consequences. Only $\mathcal{B}(h) \wedge \mathcal{B}(\neg h)$, meaning it is *believable* that Jam is healthy, and it is *believable* that Jam is not healthy will be included. Instead of interpreting the operator as "believable that", it is also possible to interpret it as "reasonable that," "there are reasons to believe that," or the like.

This version of the credulous approach has the downside that the consequences of the default theory do not directly tell us anything about the considered propositions anymore. There is no straightforward way in which the consequences of a default theory can determine a subject's doxastic situation, i.e., to determine which propositions p the subject believes or disbelieves. The consequences only tell us on a meta-level which propositions are (at all) believable. However, this does not mean that a subject could simply believe *every* proposition that is marked believable in the consequences. This would directly lead to the problem of contradicting beliefs again. Hence, this option is not perfectly suitable, especially when considering that doxastic attitudes of a subject (that stem from the consequences) should be action-guiding. Still, in the positive account, which I will present in Section 4.3, I will take up the general idea of introducing second-order propositions to store information *about* the propositions.

A third, so-called *skeptical approach* can be seen as the converse approach to the credulous one (Horty, 2011, p. 37). A skeptical subject will believe a proposition only if it follows from *every* proper scenario, i.e., it is in every extension. The consequences of the default theory then consist of the intersection of the resulting extensions. The idea is that a subject can be absolutely sure about a proposition p if p is supported by every proper scenario. Any reasonable position one can take leads to believing p. It is clear that a skeptical approach is not confronted with the drawback of the credulous approach of leading to inconsistent conclusions. If the extensions are consistent, so will the consequences be.

A key area of investigation in skeptical reasoning centers around floating conclusions, a topic that I explored in detail in Chapter 3. The acceptance or rejection of floating conclusions remains an open issue within skeptical reasoning, leading to various versions of the skeptical account depending on the stance taken on this question, as highlighted by Horty (2002). Each

version argues for a uniform treatment of floating conclusions — either accepting all of them or rejecting all of them — an approach that was demonstrated to be highly implausible in Chapter 3.

In skeptical reasoning, only propositions that are supported in every proper scenario are considered as consequences of a default theory. All other propositions that might be supported by some but not all scenarios will simply be "thrown out." If it is the consequences of the default theory that determine the doxastic position of a subject (in an epistemic interpretation), those thrown-out propositions will have the status of not even being considered by the subject. Propositions that were originally involved in certain defaults, which were argued for or against, and which were involved in certain proper scenarios, are simply forgotten or ignored. They are "thrown away" and not represented any longer. In my example, this would mean that the subject would neither have the proposition "Jam is healthy" nor the proposition "Jam is unhealthy" in mind, and we would describe a situation in which the subject never even thought about the healthiness of jam. This is not the situation we want to describe, though. In the example, we are dealing with a situation in which the subject very well thought about the matter, but did not come to a decisive conclusion, as the respective scenarios and extensions are conflicted. Hence, the skeptical approach does not provide us with a fully satisfactory way to deal with conflicting defaults either.

These issues may not appear extremely critical in the given example of the food detector. However, in more high-stakes contexts, relying on these state-of-the-art approaches can result in arbitrary decisions that determine questions with profound and far-reaching consequences for individuals or even entire communities. If certain propositions are randomly chosen, or if conflicting evidence results in the neglect of a proposition (making the situation indistinguishable from scenarios where the proposition never occurred in the first place) the mechanisms by which a system arrives at its current belief states become less transparent. This adds to the challenge that decisions made by artificial devices are often not understandable, leading to further decreasing trust in these systems. As outlined in Chapter 1, the primary objective of this research is to address the trust-related concerns arising from a lack of understanding regarding AI

system decisions. This will be accomplished by transparently representing and communicating uncertainties and conflicts, and thereby preventing arbitrary decisions.

In the following section, I will propose a way to deal with these drawbacks of the current system. I suggest a way of finding the consequences of a default theory that is resistant to the downsides of the choice, the credulous, and the skeptical approach. This enables the appropriate handling and communication of uncertainties and conflicts, intending to enhance trust in the system's responses.

4.3 A Novel Adjustment of Default Logic

4.3.1 Motivation

In the following, I will present my own, alternative proposal for dealing with conflicts in default logic. This proposal involves defining the consequences of a default theory through the extensions of the theory. I denote the consequences with Γ. To accomplish this, I will introduce four distinct second-order propositions that reflect the various statuses a proposition can hold within the (multiple) extensions.

These four second-order propositions, namely believability, certainty, suspension, and ignorance, partly constitute the consequences of a default theory. Nevertheless, the first-order propositions that we aim to infer with certainty are also included in the consequences Γ. Thus, the consequences include both the propositions we consider straightforwardly and the higher-order information about them. There are three primary motivations behind this proposal.

The first motivation is to address the challenge of handling unresolved conflicts and multiple extensions in default logic effectively. As discussed earlier, existing approaches such as credulous, skeptical, or choice reasoning are inadequate. Additionally, relying solely on priorities may not always be suitable either. While in certain cases conflicts can be resolved by assigning higher priority to one default over another, there are situations where defaults appear to have equal priority or are incomparable. Requiring a *strict total* priority ordering to resolve *all* possible conflicts seems unrealistic.

My proposal ensures a *unique* solution for every default theory. The propositions of this unique solution can be viewed as the explicit consequences of a default theory. This prevents the system from encountering errors and allows it to continue the reasoning process without resorting to arbitrary choices regarding the proper consequences.

The second motivation stems from the desire to enhance the conceptual richness of the framework. By introducing second-order propositions into the consequences, I enable the system to engage in higher-order reasoning. This form of reasoning is more sophisticated and is associated with a higher level of intelligence and abstract reasoning abilities, which are desirable attributes for artificial frameworks.

Higher-order reasoning also offers a means to adequately represent certain cognitive processes. Intelligent subjects can think about their reasons, reflect on and evaluate their own beliefs, and assess their rules of reasoning, particularly in situations involving conflicts. This type of thinking *is* higher-order reasoning. "Higher-order reasoning" provides a broad description of what occurs psychologically when individuals reason about their own epistemic state. In fact, one specific instance of higher-order reasoning is found in scenarios where different rules are assigned varying priorities. Some conflicts between rules can be resolved by assigning a higher priority to one rule. What we are doing here is *thinking about* these rules on a second level and evaluating them with the help of a priority order. Another instance of higher-order reasoning emerges when we encounter situations in which we cannot ascertain which rule or proposition should be favored, leading us to enter (or remain) in an undecided stance. These scenarios are also categorized as instances of higher-order reasoning.

The third motivation for this approach is formed by the specific desire to properly represent the indecision of a subject in the framework of default logic. The state-of-the-art framework of default logic does not provide a means to represent suspension or any other form of doxastic neutrality. As argued before, suspension (or indecision, or doxastic neutrality in general) is an important component of an adequate description of epistemic subjects. If we want to use logical frameworks to describe the epistemic household of artificial subjects, it is crucial to represent not

only when a proposition is accepted or rejected, but also when neutrality occurs. Moreover, an extensive representation of epistemic phenomena also includes *different forms of neutrality*. The very basic distinction between suspension and (deep) ignorance is an important first step for this, allowing us to distinguish conflicting situations from situations of mere lack of information.

While this distinction is especially desirable for an epistemic interpretation of default logic, I will show in Subsection 4.3.5 how the distinction between the two second-order attitudes $S(p)$ (suspension towards p) and $I(p)$ (ignorance towards p) finds a reasonable interpretation in a deontic context, too.

4.3.2 Consequences of a Default Theory

In this subsection, I will present my adaption of default logic. The basic framework corresponds to Horty's framework as presented in Section 4.2. All definitions presented in Section 4.2 can be carried over to this subsection. In the following, I will expand the framework by introducing a novel definition of the *consequences* Γ of a default theory. Like Horty, I assume that default theories generally yield multiple extensions. The additional definitions will then provide a way to deal with the multiple extensions. I present a way to merge multiple extension sets into one set of consequences Γ.

As I want to provide a framework that does not depend on priorities, I will use only default theories *without* priorities in the following. However, the framework could easily be extended to include a priority relationship as well.

Definition 4.9 (Consequences)**.** For a default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ and a set of extensions $\{\mathcal{E}_1^{\mathrm{D}}, \mathcal{E}_2^{\mathrm{D}}, \ldots, \mathcal{E}_n^{\mathrm{D}}\}$, the set of consequences Γ is defined via the following rules.
For every proposition p:

(i) $C(p) \in \Gamma$ if $\forall \mathcal{E}_i^{\mathrm{D}} : p \in \mathcal{E}_i^{\mathrm{D}}$

(ii) $B(p) \in \Gamma$ if $\exists \mathcal{E}_i^{\mathrm{D}} : p \in \mathcal{E}_i^{\mathrm{D}}$

(iii) $S(p) \in \Gamma$ if $\exists \mathcal{E}_i^{\mathrm{D}} : p \in \mathcal{E}_i^{\mathrm{D}}$ *and* $\exists \mathcal{E}_j^{\mathrm{D}} : \neg p \in \mathcal{E}_j^{\mathrm{D}}$

(iv) $I(p) \in \Gamma$ if $\nexists \mathcal{E}_i^{\mathrm{D}} : p \in \mathcal{E}_i^{\mathrm{D}}$ *and* $\nexists \mathcal{E}_j^{\mathrm{D}} : \neg p \in \mathcal{E}_j^{\mathrm{D}}$

Additionally:

(v) $p \in \Gamma$ if $C(p) \in \Gamma$

This definition provides us with a way to include one of four second-order propositions into the consequences for a proposition p. Rule (i) of Definition 4.9 describes a situation in which p is included in *every* extension of the default theory. If this is the case, the proposition is certain, which is expressed by the second-order proposition $C(p)$. Since the extensions are always consistent, it follows that $\neg p$ is in no extension. This treatment corresponds to the skeptical way of formulating consequences. Only the propositions that belong to the intersection of all extensions are skeptically accepted. This proposal complies with skeptical reasoning by accepting the skeptically accepted propositions with certainty. Moreover, with Rule (v), $C(p) \in \Gamma \rightarrow p \in \Gamma$, it is ensured that the propositions simpliciter are also included in the consequences if they are certain, i.e., skeptically accepted. Therefore, the set of skeptically accepted propositions is a subset of the consequences defined here.

Rule (ii), in contrast, gives credit to credulous reasoning. It states that all propositions that are included in at least one extension are believable. This is expressed by the second-order proposition $B(p)$ in the consequences Γ. $B(p)$ can be read as p is believable, it is sensible to believe p, it is plausible to believe p, or there is evidence for believing p. Believable propositions are propositions that are accepted by credulous reasoning. The set of believable propositions equals the union of the extensions. However, note that if a proposition is only believable (and not certain), i.e., if a proposition is accepted only credulously but not skeptically, the proposition simpliciter will not be included in Γ. Thereby, it is avoided that Γ becomes inconsistent. It is very well possible, though, that for a proposition p, both $B(p)$ and $B(\neg p)$ are included in Γ. This corresponds to the proposal of Horty (2011, p. 36) about the adaption of credulous reasoning.

The situation in which both p as well as its negation $\neg p$ are believable is described in Rule (iii). This can be expressed by $B(p)$ and $B(\neg p)$ or by the existence of one extension $\mathcal{E}_i^{\mathrm{D}}$ that includes p, and another extension $\mathcal{E}_j^{\mathrm{D}}$ that includes $\neg p$. Note that necessarily it is $i \neq j$ since the extensions are

always consistent sets if $\mathcal{W}$ is consistent (Reiter, 1980). This rule expresses conflicting situations like the one I described in the food detector example. There is one extension that speaks for p and one extension that speaks for $\neg p$. In the present terminology, this means that both p and $\neg p$ are believable, that there is evidence for both p and $\neg p$, or that it is plausible to believe p, but also plausible to believe $\neg p$. The present proposal deals with such situations by taking p to be suspended. Hence, $S(p)$ must be read as: It is suspended whether or not p. From this, it directly follows that also $\neg p$ is suspended.

Finally, Rule (iv) covers the remaining case. When there is no extension in which p is included, but neither an extension in which $\neg p$ is included, $I(p)$ is added to the consequences Γ. This represents ignorance towards p.[9] We have a situation in which neither p nor its negation is believable or plausible to believe. We have no evidence for or against p. Hence, it is natural to describe this as a state of deep ignorance (or mere non-belief) towards p, see Chapter 2. p is in a sense disregarded or not considered. It is easy to see that ignorance towards p transfers to ignorance towards $\neg p$, too.

It is important to note that in the philosophical context, many scholars would not classify (deep) ignorance as an attitude. Especially since ignorance towards p often describes those cases in which one has not even considered p, ignorance can in a sense be characterized by not having *any attitude* towards p. Ignorance, as the less sophisticated form of neutrality, is rather the *lack* of any of the three doxastic attitudes.

Still, in the formal setting considered here, we can deduce the information that p is ignored (according to the formal definition above) from a default theory. This information is valuable and should be saved and represented in some manner. As we will see, my framework allows to capture certain information that might otherwise be lost. This is particularly valuable in situations involving undercut defaults. Hence,

[9] I will use the term (deep) ignorance to refer to propositions that are not considered and are, therefore, *ignored*. These cases are comparable to the scenarios discussed in Chapter 2, where one lacks cognitive contact with the relevant proposition. Throughout this chapter, I will use both the terms "deep ignorance" and simply "ignorance" to describe these phenomena. I will not make a finer differentiation between different ignorance cases, i.e., cases in which there is no understanding of the concepts involved versus cases in which the proposition is simply not considered, as I presented in Part 2.3.2.b of Subsection 2.3.2.

in order to maintain completeness and consistency with the other three situations, this framework includes the fourth possibility, ignorance, through a second-order attitude, too.

Some Examples

Before we present the logical principles governing my proposal, I will show how the consequences are defined for the food detector example and how they are defined for cases of undercutting.

Considering the simple food detector again, we can offer a new solution for the consequences of this example. As already described, the proper scenarios of this theory displayed in Figure 4.4, are $\mathcal{S}_1 = \{\delta_1\}$ and $\mathcal{S}_2 = \{\delta_2\}$. The extensions are accordingly $\mathcal{E}_1^{\mathrm{D}} = \{f, s, h\}$ and $\mathcal{E}_2^{\mathrm{D}} = \{f, s, \neg h\}$. According to Definition 4.9, the consequences of the default theory are $\Gamma = Th(\{C(s), C(f), B(h), B(\neg h), S(h), f, s\})$.[10] One can see that the propositions f and s, which are included in the knowledge base, are included with certainty.[11] As they are certain, the propositions simpliciter are also included in Γ. The proposition h, for which there is conflicting evidence for and against, is included as a "to be suspended" proposition. Both h and $\neg h$ are believable. This gives us the desired result for a proposition for which we have conflicting evidence for and against. According to my proposal, the proposition is neither forgotten (as in the skeptical approach) nor is a contradiction included in Γ. Rather, the information that the proposition is suspended (which is the most plausible situation) is stored in the consequences.

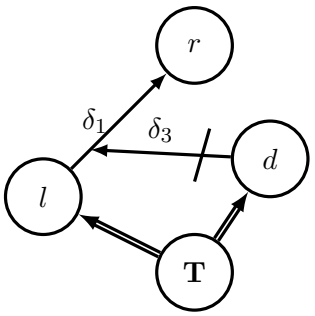
Next, I will consider what the consequences look like for cases of undercutting. When we again consider the example of Figure 4.3 on page 102, we find that the only possible extension, in this case, is $\mathcal{E}^{\mathrm{D}} = \{l, s, d, \neg r, Out(d_1)\}$. Hence, the consequences of the theory are

[10] The logical closure *within* the second-order propositions follows from the logical closure of the extensions. This means that if, for example, $C(a)$ and $C(b)$, then a and b are included in every extension. Since the extensions are additionally logically closed, $a \wedge b$ will also be in every extension and we have $C(a \wedge b)$. As already mentioned, the closure of Γ itself further guarantees that if $C(a)$ and $C(b)$ are included in Γ, then $C(a) \wedge C(b)$ is also included. For the sake of readability, I will sometimes skip writing Th in front of each consequence set.

[11] They are, of course, also believable. For reasons of readability, I will only include this information when necessary.

$\Gamma = \{l, s, d, \neg r, Out(d_1), C(l), C(s), C(d), C(\neg r), C(Out(d_1))\}$. This is not surprising, as this example only has one extension. Every proposition from this unique extension has then to be accepted with certainty. As shown in Figure 4.3, the statement $\neg r$ is supported by an additional argument $(\delta_2 : s \rightarrow \neg r)$, which is not conflicted anymore once δ_1 is undercut.

The example becomes more interesting if we consider it in an adjusted version where we do not have the additional support for the proposition that the object is not red. Here, we do not have the proposition s stating that a reliable source said that the object is not red. Neither do we have the default δ_2.



r The object is red.

d I have taken a drug that makes everything look red.

l The object in front of me looks red.

Figure 4.5: Default logic example of an undercutting defeater in a restricted version without additional information for or against r.

In this example, the undercutting of default δ_1 leaves us without any information, either supporting or opposing the proposition r, as there are no further arguments concerning r. Consequently, we have only one proper scenario consisting solely of δ_3 and only one extension with l, d, and $Out(d_3)$. In this single extension, neither r nor $\neg r$ is included. This is also reflected in the consequences $\Gamma = \{l, d, Out(d_3), C(l), C(d), C(Out(d_3)), I(r)\}$. In cases where we only have one default arguing for r, which is then undercut, we are left with ignorance regarding the proposition r.

At first glance, this might appear unconvincing, as we are treating r as if it were a proposition we have never considered and about which we have no information at all. This situation may seem akin to skeptical approaches, where information about r is effectively "thrown away" and forgotten. Nevertheless, the representation of propositions that are conclusions of undercut arguments as ignored aligns precisely with Horty's

conception of undercutting arguments. Horty regards undercut defaults as erased defaults, effectively rendering them non-existent. This is because when a default is undercut, the inferential relationship between the premise and the conclusion is cut, and there is no longer any connection between the two. Consequently, the premise of an undercut default no longer serves as a "reason" for its conclusion. As Horty defines a reason in terms of the premise of a default, an undercut default is excluded from consideration. In Horty's system, undercutting attacks take precedence. The definition of excluded, undercut defaults in Horty (2011, p. 124) highlights the clear precedence of undercutting defeats over rebutting defeats. Undercut defaults are already excluded from the set of triggered defaults, and only the triggered defaults are assessed for conflicts or rebutting defeats. In a way, there is a two-step process: First, the default theory is "cleaned" by removing all undercut defaults, and then the remaining defaults from the cleaned default theory are evaluated and potentially included in proper scenarios.

While more sophisticated accounts of undercutting exist (see, for example, Knoks, 2021), I will adhere to Horty's interpretation here, which leads to the situation that conclusions of undercut defaults are treated as ignored.

4.3.3 Logical Principles

Connections between Attitudes

The various second-order propositions I introduced are not independent; they are interconnected. The following theorem illustrates the logical relationships between these second-order propositions and the corresponding proposition simpliciter.

Theorem 4.1. The different second-order propositions on p and the proposition p simpliciter are logically connected in the consequences Γ in the following way:

1. $S(p) \not\leftrightarrow C(p)$.
2. $S(p) \rightarrow B(p)$.
3. $C(p) \rightarrow B(p)$.

4. $C(p) \leftrightarrow p$.

5. $B(p) \not\Leftrightarrow I(p)$.

The connections from Theorem 4.1 can also be illustrated in Figure 4.6.

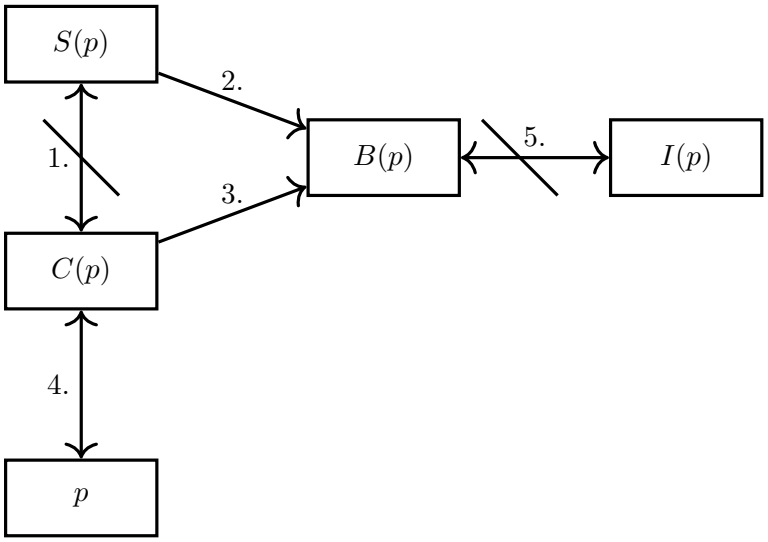


Figure 4.6: Graphical representation of the logical connections of the different second-order propositions in the consequences Γ.

Proof.

1. To show that for all p we have $S(p) \not\Leftrightarrow C(p)$, which means $(S(p) \wedge C(p)) \notin \Gamma$, let first $S(p) \in \Gamma$. This means that there is one extension $\mathcal{E}_i^{\mathrm{D}}$ such that $\neg p \in \mathcal{E}_i^{\mathrm{D}}$. Since extensions are consistent, it is $p \notin \mathcal{E}_i^{\mathrm{D}}$. Hence, p is not in every extension, so $C(p) \notin \Gamma$. Secondly, if $C(p) \in \Gamma$, then $p \in \mathcal{E}_i^{\mathrm{D}}$ for all i. Due to the consistency of the extensions, it is for all i: $\neg p \notin \mathcal{E}_i^{\mathrm{D}}$, which means that $B(\neg p) \notin \Gamma$ and $S(p) \notin \Gamma$.

2. To show that $S(p) \rightarrow B(p)$, assume $S(p) \in \Gamma$. Then, there is an extension $\mathcal{E}_i^{\mathrm{D}}$, such that $p \in \mathcal{E}_i^{\mathrm{D}}$. Thus, the consequences contain $B(p)$.

3. To show that $C(p) \rightarrow B(p)$, assume $C(p) \in \Gamma$. Then $p \in \mathcal{E}_i^{\mathrm{D}}$ for all i. Because $W \neq \emptyset$ and $W \subseteq \mathcal{E}_i^{\mathrm{D}}$ for all i, there is always at least one non-empty extension $\mathcal{E}_i^{\mathrm{D}}$ and $p \in \mathcal{E}_i^{\mathrm{D}}$. Hence, $B(p) \in \Gamma$.

4. $C(p) \rightarrow p$ is established by definition. Also, for a proposition $p \in \mathcal{W}$, p will be in every extension (since $W \subseteq \mathcal{E}_i^{\mathrm{D}}$ for all i) and hence, $C(p)$ and thereby p will also be in the consequences Γ. Since the definition provides the sole means by which propositions simpliciter can be included in Γ, the proposition p will only be included once $C(p)$ is included and thereby we also have $C(p) \leftarrow p$.

5. To show that $B(p) \not\Rightarrow I(p)$, first assume $B(p) \in \Gamma$. Then, there is an extension $\mathcal{E}_i^{\mathrm{D}}$ with $p \in \mathcal{E}_i^{\mathrm{D}}$. Hence, $I(p)$ cannot be in Γ. If $I(p) \in \Gamma$, then for all extensions $\mathcal{E}_i^{\mathrm{D}}$ it is $p \notin \mathcal{E}_i^{\mathrm{D}}$. Hence, $B(p)$ *cannot* be in Γ.

□

Given that the extensions themselves are logically closed, we can establish certain theorems concerning the behavior of different second-order attitudes about first-order propositions connected by the Boolean operators $\neg, \wedge, \vee$.

I will present a table for each of these logical connectives, investigating whether we also have a second-order proposition about the Boolean connectives when there is a second-order proposition about the atomic proposition(s). The tables are supposed to answer questions like: Assuming $B(p)$, does a second-order attitude towards $\neg p$ exist, and if yes, which one(s)? Similarly: Assuming $B(p)$ and $S(q)$, is there a second-order proposition about $p \wedge q$ or $p \vee q$, and if yes, which one(s)? Before exploring all possible combinations for all three connectives $\neg, \wedge$, and $\vee$ in Tables 4.1, 4.2, and 4.3, I will give a brief guide of how to read them: The first column and row specify the second-order attitude supposed to hold for the atomic propositions p and q. In the case of negation, there is only p, and only the first *row* shows the supposed second-order attitude on p. The other cells denote the second-order attitude for the respective Boolean connective, e.g., which second-order attitude exists for $p \wedge q$, assuming $B(p)$ and $S(q)$.

It is crucial to note that not all positions in the table will be filled. For certain positions, it is evident that the respective Boolean connective will not appear in the consequences through any second-order proposition. For specific assignments of p and q, it will be the case that neither $C(p \wedge q)$, nor $B(p \wedge q)$, nor $S(p \wedge q)$, nor $I(p \wedge q)$ hold. In such instances, a horizontal line is used to denote the absence of these outcomes.[12]

[12]Hence, I will not explicitly indicate which second-order propositions *do not* apply. For instance, in the case of negation, if $C(p)$ holds, it implies $\neg C(\neg p)$, $\neg B(\neg p)$, and so forth.

Additionally, if $B(p)$ is present in the first column or the first row, it signifies that *only* $B(p)$ is supposed to hold, without the additional presence of $C(p)$ or $S(p)$. Cases where these additional attitudes are supposed to hold are covered in other columns or rows. Furthermore, for some connectives, there may be multiple potential outcomes, depending on the specific distribution of the extensions. If one possibility is weaker or implied by the others, *only the minimal attitude* is inserted in the table. For instance, if both $B(p \vee q)$ and $S(p \vee q)$ are possible (depending on the exact example), only the minimal setting "min. $B(p \vee q)$" is inserted in the table. In cases where two *incompatible* attitudes are possible, both options are presented in the table.

I will start by introducing the tables for negation and conjunction, followed by a discussion including some examples and proof of selected entries in the conjunction table. Subsequently, I will proceed with the table for disjunction, along with comments and proofs for some entries in the disjunction table.

Negation

	$C(p)$	only $B(p)$	$S(p)$	$I(p)$
$\neg$	—	—	$S(\neg p)$	$I(\neg p)$

Table 4.1: Second-order attitudes towards the negation of an atomic proposition p (second row) based on the supposed attitudes towards the original atomic proposition p.

It is straightforward to see why suspension and ignorance are both closed under negation. Additionally, if there is only $B(p)$, it is *not* $S(p)$. Hence, $B(\neg p)$ is not in Γ. It is easy to see that $I(\neg p)$ can neither be in Γ. Similarly, if there is $C(p)$, no attitude towards $\neg p$ is possible.

These negative results will be represented by the horizontal line.

Conjunction

$\wedge$	$C(p)$	only $B(p)$	$S(p)$	$I(p)$
$C(q)$	$C(p \wedge q)$	$B(p \wedge q)$	$S(p \wedge q)$	$I(p \wedge q)$
only $B(q)$	$B(p \wedge q)$	$B(p \wedge q)$ or $I(p \wedge q)$	$S(p \wedge q)$ or —	$I(p \wedge q)$
$S(q)$	$S(p \wedge q)$	$S(p \wedge q)$ or —	$S(p \wedge q)$ or —	—
$I(q)$	$I(p \wedge q)$	$I(p \wedge q)$	—	$I(p \wedge q)$

Table 4.2: Second-order attitudes towards the conjunction $p \wedge q$ based on the supposed attitudes towards the original atomic proposition p (first row) and q (first column).

When one proposition is certain, such as $C(p)$, the attitude of the second proposition is transferred to the conjunction. This can be explained by the fact that if p is certain, it implies that p is present in every extension. Consequently, what happens to the conjunction $p \wedge q$ is solely determined by the status of the second proposition q. In every extension where q is included, the conjunction will also be included. Moreover, it is evident why I is closed under conjunction. If there is no extension that includes either p or its negation and there is no extension that includes either q or its negation, then no extension will include $p \wedge q$. Correspondingly, no extension will include $\neg(p \wedge q)$. Further interesting results are elaborated in the following theorems and examples.

Theorem 4.2. If $S(p)$ and $I(q)$ then there is no attitude towards $p \wedge q$.

Proof. Let $S(p)$ and $I(q)$. As it is $I(q)$, there is no extension that contains q. Hence, there is no extension that contains $p \wedge q$. This means that $B(p \wedge q)$ and the stronger attitudes $S(p \wedge q)$ and $C(p \wedge q)$ are excluded. The only possible attitude left is $I(p \wedge q)$. However, for this to be the case, also the negation, $\neg(p \wedge q)$ must not be in any extension. However, due to $S(p)$,

there is an extension $\mathcal{E}_i^{\mathrm{D}}$ such that $\neg p \in \mathcal{E}_i^{\mathrm{D}}$. From this, it follows that $(\neg p \vee \neg q) \in \mathcal{E}_i^{\mathrm{D}}$ which is equivalent to $\neg(p \wedge q) \in \mathcal{E}_i^{\mathrm{D}}$. □

Theorem 4.3. If $S(p)$ and $S(q)$, then either $S(p \wedge q)$ or there is no attitude towards $p \wedge q$. (Likewise, if $B(p)$ and $S(q)$ or if $S(p)$ and $B(q)$, it is either $S(p \wedge q)$ or there is no attitude towards $p \wedge q$).

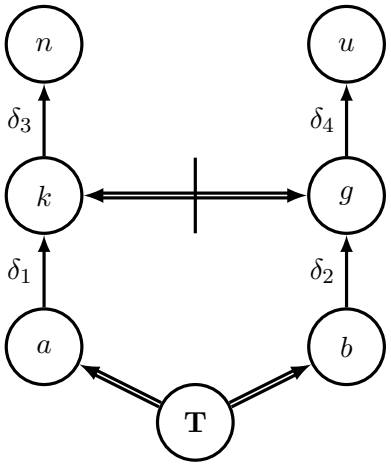
Proof. Let $S(p)$ and $S(q)$. Then there are extensions $\mathcal{E}_i^{\mathrm{D}}, \mathcal{E}_j^{\mathrm{D}}, \mathcal{E}_m^{\mathrm{D}}, \mathcal{E}_n^{\mathrm{D}}$ such that $p \in \mathcal{E}_i^{\mathrm{D}}$, $\neg p \in \mathcal{E}_j^{\mathrm{D}}$, $q \in \mathcal{E}_m^{\mathrm{D}}$, and $\neg q \in \mathcal{E}_n^{\mathrm{D}}$. Due to $\mathcal{E}_j^{\mathrm{D}}$, it is clear that $B(\neg p \vee \neg q)$, i.e., $B(\neg(p \wedge q))$.[13] Due to consistency, it is $i \neq j$ and $m \neq n$. We distinguish two cases. In the first case, there is an extension $\mathcal{E}_k^{\mathrm{D}}$ (possibly $k = i = m$) with $p \in \mathcal{E}_k^{\mathrm{D}}$ *and* $q \in \mathcal{E}_k^{\mathrm{D}}$. In the second case, there is no such extension.

- If there is an extension $\mathcal{E}_k^{\mathrm{D}}$ with $p \in \mathcal{E}_k^{\mathrm{D}}$ *and* $q \in \mathcal{E}_k^{\mathrm{D}}$, then $(p \wedge q) \in \mathcal{E}_k^{\mathrm{D}}$ and hence $B(p \wedge q)$. With $B(\neg(p \wedge q))$, it follows that $S(p \wedge q)$.
- If there is no extension $\mathcal{E}_k^{\mathrm{D}}$ with $p \in \mathcal{E}_k^{\mathrm{D}}$ *and* $q \in \mathcal{E}_k^{\mathrm{D}}$, then $\neg B(p \wedge q)$. With this, it follows that $\neg S(p \wedge q)$ and not $\neg C(p \wedge q)$. With $B(\neg(p \wedge q))$, we exclude the last possible attitude, hence, $\neg I(p \wedge q)$. In this case, there is no attitude towards $p \wedge q$.

A similar reasoning can be made for $B(p)$ and $S(q)$, or vice versa. □

The two possible outcomes that are described in Theorem 4.3 can be visualized by the following two examples, see Figure 4.7 and Figure 4.8.

[13] In a special case of this, there is even $C(\neg(p \wedge q))$.



n Peter has a knife.

u Peter has a gun.

k Peter killed the victim with a knife.

g Peter killed the victim with a gun.

a Witness A says that Peter killed the victim with a knife.

b Witness B says that Peter killed the victim with a gun.

Figure 4.7: Adapted version of a floating conclusion example from the Chapter 3 with no floating conclusion but two distinct conclusions for the different conflicting propositions k and g. $\mathcal{W} = \{a, b, \neg(k \wedge g)\}$.

It is straightforward to identify that the only proper scenarios of Figure 4.7 are $\mathcal{S}_1 = \{\delta_1, \delta_3\}$ and $\mathcal{S}_2 = \{\delta_2, \delta_4\}$. The extensions are accordingly $\mathcal{E}_1^{\mathrm{D}} = \{a, b, k, n, \neg(k \wedge g), \neg g\}$ and $\mathcal{E}_2^{\mathrm{D}} = \{a, b, g, u, \neg(k \wedge g), \neg k\}$. Therefore, the (relevant part) of the consequences are: $\Gamma = \{C(a), C(b), S(k), S(g), B(n), B(u)\}$.

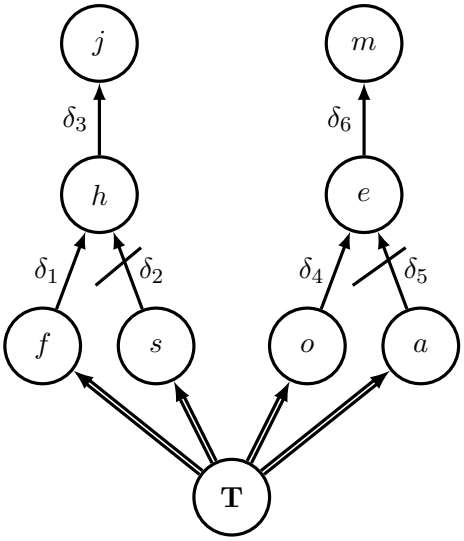
Now, if we consider the logical closure of these consequences, we encounter an instance of Theorem 4.3. Specifically, we have both $S(k)$ and $S(g)$. As the example shows, this does not necessarily imply that we also have $S(k \wedge g)$. In this case, there is no extension in which both k and g are included. Thus, while the negation of the conjunction is believable, i.e., $B(\neg(k \wedge g))$, which is equivalent to $B(\neg k \vee \neg g)$, the conjunction itself is not believable.[14] Even though we have evidence for both k and g, the evidence is not independent of each other, and therefore we cannot combine it to form evidence for $k \wedge g$.

In this particular case, it is even certain that $\neg(k \wedge g)$ because this is already included in the knowledge base. Examples like the one in Figure 4.7 serve as counterexamples to the claim that suspension is always closed under conjunction. It might very well be that we suspend about two

[14] Since the negation is believable, i.e., $B(\neg(k \wedge g))$, it follows that we cannot have $I(k \wedge g)$ either.

propositions, but we know that not both can be true simultaneously.

However, as demonstrated in Theorem 4.3, there are also cases where the conjunction of two suspended propositions is indeed also suspended. This can be seen in the following example in Figure 4.8.



j Jam should be served in school cafeterias.

m Mackerel should be served in school cafeterias.

h Jam is healthy.

e Mackerel is healthy.

f Jam contains a lot of fruit.

s Jam contains a lot of sugar.

o Mackerel contains a lot of Omega-3.

a Mackerel contains a lot of fat.

Figure 4.8: Default logic example of the food detector in an extended version considering the healthiness of both jam and mackerel.

In this extended example, we have two independent chains of arguments that constitute one default theory. The proper scenarios of this theory are $\mathcal{S}_1 = \{\delta_1, \delta_3, \delta_4, \delta_6\}$, $\mathcal{S}_2 = \{\delta_2, \delta_4, \delta_6\}$, $\mathcal{S}_3 = \{\delta_1, \delta_3, \delta_5\}$, and $\mathcal{S}_4 = \{\delta_2, \delta_5\}$. The extensions are accordingly $\mathcal{E}_1^{\mathrm{D}} = \{f, s, o, a, h, j, e, m\}$, $\mathcal{E}_2^{\mathrm{D}} = \{f, s, o, a, \neg h, e, m\}$, $\mathcal{E}_3^{\mathrm{D}} = \{f, s, o, a, h, j, \neg e\}$, and $\mathcal{E}_4^{\mathrm{D}} = \{f, s, o, a, \neg h, \neg e\}$. Hence, the second-order attitudes towards the atomic propositions are $\{C(f), C(s), C(o), C(a), S(h), S(e), B(j), B(m)\}$. The situation is similar to the one in Figure 4.7 in the sense that there are also two propositions e and h that are both suspended in the consequences. However, the difference is that the conjunction is suspended, too, i.e., it is $S(h \wedge e)$. This is the case here because there is an extension, $\mathcal{E}_1^{\mathrm{D}}$, in which both e and h are true, which leads to $h \wedge e$ being believable. This is plausible for the example in Figure 4.8 since the propositions h and e seem to be somehow independent. Whether or not jam is healthy does not seem to influence whether or not

mackerel is healthy and vice versa. At least there are no defaults in this theory that represent a connection between the two propositions. We have evidence for and against h and we have evidence for and against e, and since the evidence is independent, we can somehow combine the evidence. Thus, we end up with both evidence for and against $h \wedge e$, which in turn leads to suspension about $h \wedge e$.

Theorem 4.4. If $I(q)$ and only $B(p)$ (and not $S(p)$ or $C(p)$) hold then $I(p \wedge q)$.

Proof. Due to $I(q)$, no extension includes q. Hence, no extension includes $p \wedge q$. Then, it remains to be shown that no extension includes the negation, $\neg(p \wedge q)$, i.e., $\neg p \vee \neg q$. Due to $I(q)$, no extension contains $\neg q$ and because we assume that only $B(p)$ (and not $S(p)$), no extension contains $\neg p$. Hence, no extension includes the disjunction $\neg p \vee \neg q$, which is equivalent to $\neg(p \wedge q)$, and $I(p \wedge q)$ follows. □

Theorem 4.5. If only $B(p)$ and only $B(q)$ (and not $S(p)$, $S(q)$, $C(p)$, or $C(q)$) hold then $B(p \wedge q)$ or $I(p \wedge q)$.

Proof. Let $B(p)$ and $B(q)$. Then, there exist $\mathcal{E}_i^{\mathrm{D}}$ and $\mathcal{E}_j^{\mathrm{D}}$ with $p \in \mathcal{E}_i^{\mathrm{D}}$ and $q \in \mathcal{E}_j^{\mathrm{D}}$. We distinguish two cases. In the first case, there is no extension $\mathcal{E}_k^{\mathrm{D}}$ with $p \in \mathcal{E}_k^{\mathrm{D}}$ *and* $q \in \mathcal{E}_k^{\mathrm{D}}$. In the second case, there is such an extension (possibly $k = i = j$).

- If there is no extension that includes p *and* q, there is no extension that includes $p \wedge q$. Hence, $\neg B(p \wedge q)$ holds. Moreover, since we assume that only $B(p)$ and $B(q)$ (and not also $S(p)$ and $S(q)$), no extension includes $\neg p$ and no extension includes $\neg q$. Hence, $\neg p \vee \neg q$ holds, which is equivalent to $\neg(p \wedge q)$, is not included in any extension. Hence, $\neg B(\neg(p \wedge q))$ and thereby $I(p \wedge q)$ hold.
- If there is an extension $\mathcal{E}_k^{\mathrm{D}}$ with $p \in \mathcal{E}_k^{\mathrm{D}}$ and $q \in \mathcal{E}_k^{\mathrm{D}}$, i.e., $(p \wedge q) \in \mathcal{E}_k^{\mathrm{D}}$, then $B(p \wedge q)$. With the reasoning from the other case, it follows that $\neg B(\neg(p \wedge q))$, which leads to $\neg S(\neg(p \wedge q))$. Moreover, since *only* $B(p)$ and $B(q)$, there must be extensions $\mathcal{E}_m^{\mathrm{D}}$ and $\mathcal{E}_n^{\mathrm{D}}$ (possibly $m = n$), with $p \notin \mathcal{E}_m^{\mathrm{D}}$ and $q \notin \mathcal{E}_n^{\mathrm{D}}$, because otherwise it would be the case that $C(p)$ or $C(q)$. Hence, it is not the case that $C(p \wedge q)$ and only the case that $B(p \wedge q)$.

□

In the following, I will present the table for the logic of the disjunction and add some proofs for certain entries.

Disjunction

$\vee$	$C(p)$	only $B(p)$	$S(p)$	$I(p)$
$C(q)$	$C(p \vee q)$	$C(p \vee q)$	$C(p \vee q)$	$C(p \vee q)$
only $B(q)$	$C(p \vee q)$	min. $B(p \vee q)$	min. $B(p \vee q)$	$B(p \vee q)$
$S(q)$	$C(p \vee q)$	min. $B(p \vee q)$	min. $B(p \vee q)$	$B(p \vee q)$
$I(q)$	$C(p \vee q)$	$B(p \vee q)$	$B(p \vee q)$	$I(p \vee q)$

Table 4.3: Second-order attitudes towards the disjunction $p \vee q$ based on the supposed attitudes towards the original atomic proposition p (first row) and q (first column).

When examining the behavior of the disjunctive connective in relation to the second-order propositions, several observations can be made. First, it is evident that the certainty of one proposition transfers to the certainty of the disjunction. This is unsurprising, as the disjunction is true in an extension if at least one of the disjuncts is true, which is always the case if one of them is certain. Second, in a similar manner, the believability of one proposition transfers to the disjunction. If one of the disjuncts is believable, then the disjunction is also believable. Third, if both propositions are ignored, then the disjunction is ignored as well. This is a straightforward outcome because ignoring both disjuncts implies ignoring the disjunction. Finally, there are two cases that require further explanation: when both propositions are suspended, or when one is suspended, and the other is ignored. In both cases, the disjunction will be (at least) believable.

Theorem 4.6. If $S(p)$ and $S(q)$ then at least $B(p \vee q)$.

Proof. Let $S(p)$ and $S(q)$. Then, there are extensions $\mathcal{E}_i^{\mathrm{D}}, \mathcal{E}_j^{\mathrm{D}}, \mathcal{E}_m^{\mathrm{D}}, \mathcal{E}_n^{\mathrm{D}}$ such that $p \in \mathcal{E}_i^{\mathrm{D}}$, $\neg p \in \mathcal{E}_j^{\mathrm{D}}$, $q \in \mathcal{E}_m^{\mathrm{D}}$, and $\neg q \in \mathcal{E}_n^{\mathrm{D}}$. It is $p \vee q \in \mathcal{E}_i^{\mathrm{D}}$ and $p \vee q \in \mathcal{E}_m^{\mathrm{D}}$. Hence, $B(p \vee q)$. Additionally, we can make the following distinction:

- If there is an extension $\mathcal{E}_k^{\mathrm{D}}$ (possibly $k = j = n$) with $\neg p \wedge \neg q \in \mathcal{E}_k^{\mathrm{D}}$, then it is $B(\neg(p \vee q))$ and thereby also $S(p \vee q)$. This situation we see in the example illustrated in Figure 4.8.
- If there is no such extension, it is only $B(p \vee q)$, or possibly even $C(p \vee q)$ which is the case in the example illustrated in Figure 4.7. In the case of certainty, we must have for each $\mathcal{E}_j^{\mathrm{D}}$ with $\neg p \in \mathcal{E}_j^{\mathrm{D}}$ that $q \in \mathcal{E}_j^{\mathrm{D}}$ *and* for every $\mathcal{E}_n^{\mathrm{D}}$ for which $\neg q \in \mathcal{E}_n^{\mathrm{D}}$ that $p \in \mathcal{E}_n^{\mathrm{D}}$.

□

This theorem captures what I described in Chapter 3 about floating conclusions. In these cases, we have two suspended propositions that are contrary to each other. To define the appropriate reaction to the inference (the floating conclusion) of the two propositions one has to determine whether we have more evidence in favor of $p \vee q$ (as in the second case of the proof of Theorem 4.6), or whether there is some evidence for $\neg(p \vee q)$, i.e., $\neg p \wedge \neg q$, which is captured by the first case of the proof.

Theorem 4.7. If $S(p)$ and $I(q)$ then $B(p \vee q)$.

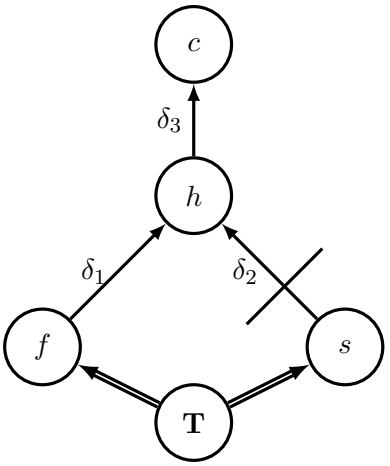
Proof. Let $S(p)$. Then, there is some extension $\mathcal{E}_i^{\mathrm{D}}$ with $p \in \mathcal{E}_i^{\mathrm{D}}$ and thereby it is $p \vee q \in \mathcal{E}_i^{\mathrm{D}}$, which means that $B(p \vee q)$. Let $I(q)$. Then, for every extension $\mathcal{E}_j^{\mathrm{D}}$ it is, in particular, $\neg q \notin \mathcal{E}_j^{\mathrm{D}}$. Since there is no extension that includes $\neg q$, there is no extension that includes $\neg p \wedge \neg q$, which is equivalent to $\neg(p \vee q)$. Hence, the negation of the disjunction is not believable, and the disjunction is not suspended but only believable. □

4.3.4 Inferences from Suspension and Floating Conclusions

As already mentioned in Chapter 3 on floating conclusions, it is interesting to see what my proposal suggests for claims that follow from suspended propositions.[15] Floating conclusions thereby only represent one such kind of situation. Here, I want to distinguish different cases and generally investigate the status of conclusions that follow from premises that are suspended.

4.3.4.a Inferences from Suspension

First, let's consider the following example that is also an extension of the food detector from Figure 4.4.



c Jam is not turquoise.

h Jam is healthy.

f Jam contains a lot of fruit.

s Jam contains a lot of sugar.

Figure 4.9: Default logic example of the food detector in an extended version with the proposition that jam is not turquoise.

In this example, from the suspended proposition that jam is healthy, it follows by another default that jam is not turquoise. (Per default healthy food is not turquoise.) Proposition c, which states that jam is not turquoise, is included in the consequences as follows: Given the two proper scenarios for this extended example, $\mathcal{S}_1 = \{\delta_1, \delta_3\}$ and $\mathcal{S}_2 = \{\delta_2\}$, the extensions are accordingly $\mathcal{E}_1^{\mathrm{D}} = \{f, s, h, c\}$ and $\mathcal{E}_2^{\mathrm{D}} = \{f, s, \neg h\}$. Hence, following the guidelines outlined in Definition 4.9, $B(c)$ will be included in the consequences Γ.

[15] Note that here "follow" has to be read in a non-monotonic way. The conditional $a \to b$ is not to be interpreted as a material conditional that is equivalent to $\neg a \vee b$, but as a non-monotonic default rule from a to b.

It is interesting to see that in this case, a proposition that follows from a suspended proposition is not suspended itself. Rather, the framework tells us that the proposition that follows is believable, and this is very plausible. The example of Figure 4.9 displays a situation in which we have evidence for h and evidence for $\neg h$. Having no priority among the evidence, it makes sense that one suspends about h. The situation is different for the proposition c, though. There is only positive evidence for c captured in the default $h \rightarrow c$. It would be a logical fallacy to conclude $\neg h \rightarrow \neg c$ from this. We have no evidence for $\neg c$. There is no default from something being not healthy to something being turquoise. Hence, the situation is not balanced as it is for h. Still, the positive evidence for c is not enough to actually conclude c. We have an argument (a default) for c but the premise of the argument is suspended. If it turned out that the premise h is in fact true, then we should conclude c. Whereas, if it turned out that h is false, then we have no evidence whatsoever for or against c. This unsymmetrical situation is captured by c being labeled believable (though not certain) without labeling $\neg c$ believable.

Of course, in other situations, propositions that follow from suspended propositions still should be (and in fact are) suspended as the following example in Figure 4.10 shows.

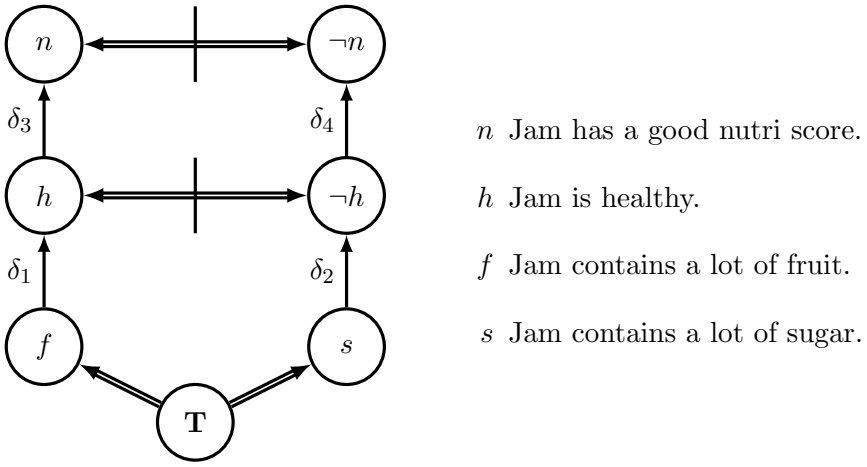


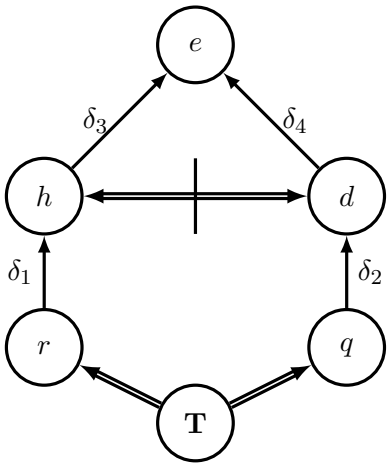
Figure 4.10: Default logic example of the food detector in an extended version with the proposition that jam has a good nutri score.

Example 4.10 differs from the previous one in that the proposition that follows from the suspended proposition h is that jam has a good nutri score

(n). Moreover, we have explicitly included the default $\neg h \rightarrow \neg n$. This makes the situation for n symmetric again. We have (conditional) evidence for and against n, which makes it reasonable to suspend about n, too. $S(n)$ is in the consequences of this example.

4.3.4.b Floating Conclusions

Another special kind of situation in which inferences from suspended propositions are investigated is situations of floating conclusions that were analyzed precisely in Chapter 3. To test what my proposal yields for floating conclusions, let us consider the Nixon case that was presented in the previous chapter, see Ginsberg (1993) and Horty (2002).



e Nixon is politically extreme.

h Nixon is a hawk.

d Nixon is a dove.

r Nixon is a Republican.

q Nixon is a Quaker.

Figure 4.11: Default logic example of the Nixon case, which is a floating conclusion example from Chapter 3.

Recall that in this example, we have two lines of reasoning. One starts from the proposition r that Nixon is a Republican, which is a reason to believe h, Nixon is a hawk, which again is a reason to believe e, Nixon is politically extreme. The second line of reasoning starts from the fact q, Nixon is a Quaker, which is a reason for d, that Nixon is a dove, which again is a reason for e, Nixon is politically extreme. These two lines of reasoning contradict each other, since Nixon cannot be both a hawk and a dove. The propositions h and d are incompatible. Still, both lines of reasoning end up with the same conclusion, the floating conclusion e, that Nixon is politically extreme.

What are the consequences of this example? We find two proper

scenarios: $S_1 = \{\delta_1, \delta_3\}$ and $S_2 = \{\delta_2, \delta_4\}$. Thus, the extensions are $\mathcal{E}_1^{\mathrm{D}} = \{r, q, h, e, \neg(h \wedge d), \neg d\}$ and $\mathcal{E}_2^{\mathrm{D}} = \{r, q, d, e, \neg(h \wedge d), \neg h\}$. Thus, the consequences Γ is the closure of the set $\{C(r), C(q), S(d), S(h), C(e), C(\neg(h \wedge d))\}$. We see that while the propositions that are in conflict with each other are suspended, the floating conclusion is *included with certainty* in the consequences.

This appears to be an undesirable feature of this proposal. As we have seen in the previous chapter, there are many examples in which the floating conclusion seems unacceptable (in the specific context). It seems unsatisfactory that floating conclusions are accepted with certainty in my approach, whereby they have the same status as propositions from the knowledge base. Floating conclusions show us cases where propositions inferred from suspended propositions are accepted with certainty. Where does the certainty of floating conclusions come from, if the premises of the arguments for them are suspended?

The certainty of floating conclusions can be attributed to the closed-world assumption that underlies systems like default logic (see Section 4.1). Recall that the closed-world assumption states in principle that all propositions that are not known to be true are false. While this property is already in principle opposed to the concept of doxastic neutrality, its problems show up even more clearly in cases like floating conclusions. For cases of floating conclusions, the closed world assumption undermines the possibility of hidden defaults, which can be blamed for making floating conclusions appear unacceptable, as described by Prakken (2002) and in the previous chapter. In the example illustrated in Figure 4.11, there is an argument for h and there is an argument for d. Although they are incompatible, both deductively support the disjunction $h \vee d$. There is no argument against the disjunction in the theory, and hence, no argument for $\neg h \wedge \neg d$ (although there might be a hidden default for this). Hence, the negation of the disjunction is assumed to be false and the disjunction itself is assumed to be true in this closed world. According to the previous chapter, this means that the third option (i.e., $\neg h \wedge \neg d$), which would possibly provide a reason to reject the floating conclusion, is disregarded. For the particular case of the Nixon example, the reason to deviate from the acceptance of the floating conclusion would be a possible

compromising proposition between h and d, from which the floating conclusion does not follow. This proposition is not represented in the default theory, and thus assumed to be false according to the closed world assumption.

Nevertheless, assuming the correctness of the closed-world assumption (which is considered a useful feature of default logic after all) my approach produces the desired result. Considering examples like the Nixon case under the closed-world assumption, i.e., under the assumption that the information provided is complete, it is perfectly plausible to conclude with certainty that Nixon is politically extreme.

Cases of floating conclusions, though, shed light on the limitations of systems like default logic. They demonstrate that the closed-world assumption is questionable. Hence, my proposed modification of default logic is only well-suited for describing a *static* epistemic scenario where a subject represents a closed and comprehensive situation containing all relevant pieces of information. In this sense, it provides a framework for representing suspension of judgment as the third doxastic attitude. Suspension is here characterized, on the one hand, by its neutrality regarding the truth of a proposition and, on the other hand, by being a *response* to a question or an outcome of an inquiry or deliberation. This interpretation of suspension aligns with the use of suspension in this thesis, as elaborated in detail in Chapter 2.

As discussed in Subsection 2.4.2, there is another facet of suspension characterized by its *zetetic* nature. This form of suspension signals open-mindedness and serves as the initiation of inquiry, deliberation, or the re-opening of a question. This inquiring attitude enables us to question the completeness of available information and to seek out new evidence. For instance, in the Nixon case, one might zetetically suspend judgment regarding Nixon's political stance and actively search for additional evidence, such as his voting history on military deployment. However, this dynamic, inquisitive, and suspending attitude is incompatible with the closed-world assumption in systems like default logic. Consequently, the framework presented does not accommodate the representation of such a dynamic and inquiring form of suspension. When deriving the entailments of a default theory, the resulting consequences are perceived as the *outcome*

of the inquiry, given the information provided by the world descriptions and the defaults.

4.3.5 Deontic Interpretation

Since Horty's default logic is originally supposed to be a system for practical, deontic reasoning, it makes sense to give a deontic interpretation to the adjusted proposal presented here. In Horty's terminology, premises of defaults represent *reasons for acting* in a certain way. Hence, the consequences of a default theory can be represented as *guidelines for acting* that are supported by some reasons. The following table shows how the second-order attitudes introduced in Section 4.3 can be interpreted in those terms.

Attitude	Rule	Epistemic Interpretation	Deontic Interpretation
$C(p)$	$\forall \mathcal{E}_i^{\mathrm{D}} : p \in \mathcal{E}_i^{\mathrm{D}}$	p is certain	p is obligatory
$B(p)$	$\exists \mathcal{E}_i^{\mathrm{D}} : p \in \mathcal{E}_i^{\mathrm{D}}$	p is believable	one ought to p (prima facie)
$S(p)$	$\exists \mathcal{E}_i^{\mathrm{D}} : p \in \mathcal{E}_i^{\mathrm{D}}$ *and* $\exists \mathcal{E}_j^{\mathrm{D}} : \neg p \in \mathcal{E}_j^{\mathrm{D}}$ $\Leftrightarrow B(p) \wedge B(\neg p)$	p is suspended	one ought to p and one ought to $\neg p$
$I(p)$	$\nexists \mathcal{E}_i^{\mathrm{D}} : p \in \mathcal{E}_i^{\mathrm{D}}$ *and* $\nexists \mathcal{E}_j^{\mathrm{D}} : \neg p \in \mathcal{E}_j^{\mathrm{D}}$ $\Leftrightarrow \neg B(p) \wedge \neg B(\neg p)$	p is ignored	one neither ought to p nor ought to $\neg p$

Table 4.4: Epistemic and deontic interpretation of the consequences of a default theory.

First, it should be noted that I am agnostic about personal and impersonal ways of spelling out obligations, permissions, etc. As noted by Horty (2011, p. 68), for this discussion, one can switch between a personal reading like "I (or one) ought to do p" and an impersonal reading like "It is obligatory that p" as in both cases conflicts can arise.

When interpreting the second-order attitudes in deontic terms, I can basically distinguish between things that one ought to do (represented by $B(p)$), and things that one must do ($C(p)$) (see Mullins, 2021, p. 572–573). The things one must do are the obligatory things and they are obligatory *all things considered*. The things one ought to do are *prima facie* oughts, i.e., things one has reason to do.

One might consider that if $C(p)$ is interpreted as "p is obligatory," it would be suitable to interpret $B(p)$ as "p is permissible." However, two reasons make this interpretation untenable. First, taking defaults to be reasons, it is evident that when $B(p)$ holds, one has *reason to do* p (or, in the epistemic interpretation, reason to believe p). p being permissible is distinct from having a reason to do p. Just because a certain action is permissible does not imply that one has a reason to actually do it.

Secondly, applying this definition to $S(p)$ and $I(p)$ leads to obscure outcomes. On the one hand, $S(p)$ would be a situation in which *both* p and $\neg p$ are permissible. This means that we can do whatever we want since every option (regarding p) is permissible. On the other hand, $I(p)$ would be a situation in which *neither* p nor $\neg p$ is permissible. This would correspond to a situation of conflict in which the subject would be left unable to act.

Not only would this interpretation not correspond to my epistemic interpretation, but it also would not accurately describe the respective examples. With $S(p)$ we want to describe conflicting situations. In epistemic terms, p is suspended, and it is suspended because conflicting defaults are arguing for p and $\neg p$. There really is a conflict in the respective situations. The deontic interpretation just sketched of "both p and $\neg p$ are permissible" does not represent this conflict at all. Rather, in this interpretation, the conflict occurs in $I(p)$, as $I(p)$ would be interpreted as "neither p nor $\neg p$ is permissible." This represents a practical dilemma in which we are left unable to act. In the epistemic interpretation though, $I(p)$ describes situations in which we do not have any information or reasons for or against p and are ignorant about it. In the examples, this is represented by having no (triggered) default inferring either p or $\neg p$. In a deontic interpretation, the respective examples describe situations in which we have neither a reason to do p nor a reason to do $\neg p$, and it seems wrong to describe those situations with an impermissibility towards either.

Since there is a correspondence between epistemic and deontic conflicts and dilemmas as noticed in Wagner (2021), the deontic interpretation should do justice to the character of the epistemic one. $S(p)$ should be used to describe a kind of dilemma, while $I(p)$ should rather be used to describe some "anything goes" situation. In my proposed interpretation in the table above, I do justice to this requirement. The interpretation of $S(p)$ as "I ought to p and I ought to $\neg p$" fits the conflicting character of this situation. Moreover, the interpretation of $I(p)$ as "I neither ought to p nor to $\neg p$" fits the character of "anything goes," since I have no reasons or oughts for either side.

At this point, it is interesting to note again that propositions for which $B(p)$ holds are credulously accepted in my framework while propositions for which $C(p)$ holds are skeptically accepted in conventional terms. Here, my difference between the obligations and the oughts corresponds to the distinction of Horty (2011, p. 73) between different deontic operators. He introduces a conflict account of ought (what is $B(p)$ here) and a disjunctive account for ought (what is $C(p)$ here). Both accounts were combined in the theory of Mullins (2021). He aims to provide a framework in which the relationships between reasons, oughts, requirements, and permissibility are properly spelled out. Similar to my account, Mullins (2021) defines the deontic "must" and "ought" in terms of a universal and existential quantifier applied to the extensions. What Mullins (2021) calls "must," I call obligatory here.

Interestingly, Mullins (2021) also defines permissibility (as it is done in standard deontic, modal logic) via: p is permissible if and only it is not the case that one must $\neg p$. In my framework, p would be permissible if $\neg C(\neg p)$, i.e., if p is not obligatory. For example, in the case of $I(p)$, we would get that both p and $\neg p$ are permissible, since neither is obligatory. This describes the situation of "anything goes" properly because there are no reasons concerning p. However, also in a case in which $B(p)$ (but not $S(p)$ or $C(p)$) is the case, we would get that one ought to p, but $\neg p$ is permissible. This is a rather wide notion of permissibility. If I have reason to do p (so a default leads to p) and *no reason not to do* it, why should it

be permissible to do $\neg p$?

An alternative, stricter version of permissibility could consist in denoting p to be permissible if and only if it is not the case that one ought to $\neg p$ (i.e., $\neg B(\neg p)$). Then, if $I(p)$ is the case, both p and $\neg p$ would be permissible, paralleling the nice feature of the approach of Mullins (2021). Additionally, if *only* $B(p)$ is present, indicating a reason for p but no reason for $\neg p$, p would be permissible while $\neg p$ would not be permissible. This interpretation of permissibility appears slightly stricter compared to the suggestion by Mullins (2021).[16]

4.4 Conclusion

In this chapter, I investigated suspension within the framework of default logic. We saw that, in the typical scenario, a default theory will involve multiple scenarios and thereby multiple extensions. Often, these sets are in conflict with each other. The investigated framework of default logic, as introduced by Horty (2011), offers various methods (choice, skeptical, credulous) to unify the different extensions and determine what the consequences of the default theory are. However, I have demonstrated that these state-of-the-art solutions are unsatisfactory. Furthermore, the option to demand a strict total ordering of priorities between the defaults as a workaround does not adequately represent our general reasoning processes.

Instead, I have introduced a definition that yields a unique solution for the consequences of any default theory. This solution incorporates the propositions themselves as well as four second-order attitudes about these propositions. It effectively combines the fundamental concepts from both skeptical and credulous reasoning. The four second-order attitudes represent certainty, believability, suspension, and (deep) ignorance. As a result, the consequences, as defined in this proposal, not only guide

[16]However, in both versions of permissibility, one can also critique the feature whereby, when $I(p)$ — that is, when there is no explicit reason for p or for $\neg p$ — both p and $\neg p$ are always deemed permissible. This feature can to some extent be attributed to the closed-world assumption discussed in Section 4.1. In practical, deontic reasoning scenarios, where defaults denote reasons for actual action, this might appear even more problematic than in the epistemic context. The system delineates a static scenario of what is allowed and forbidden. However, when we aim to provide guidance for action, a dynamic approach appears more promising.

us on which propositions we should accept but also inform us about the appropriate doxastic responses to these propositions.

The proposed solution offers an appropriate way to handle conflicting situations. In cases of unresolved conflicts, the involved propositions are suspended, and this suspension is indicated by the inclusion of the second-order attitude $\mathcal{S}(p)$ in the consequences. Hence, we are able to introduce the concept of suspension into default logic through this enhancement. The adjusted framework not only allows the representation of neutrality in the form of suspension but also distinguishes it from a different form of neutrality, i.e., ignorance. The incorporation of these four second-order propositions enriches the conceptual framework of default logic. It permits the reasoning system to construct meta-beliefs, signifying a capacity for self-insight and self-reflective reasoning. This potentiality forms the foundation for enabling a system to provide transparent and comprehensible explanations for its decisions and can possibly increase the trust in the system.

Furthermore, I examined the logical principles underpinning my definition of the consequences. Importantly, I reaffirmed that the logic of suspension is more intricate than the logic of belief. The applicability of the second-order attitude to many Boolean combinations depends on the exact circumstances, even with the second-order attitudes of the atomic propositions being fixed. Through my investigations, I have outlined a complex logical profile of suspension. For instance, I demonstrated that propositions that follow from suspended propositions are not necessarily suspended themselves. Instead, I showed that they are always deemed to be (at least) believable.

While the proposed adjustment of default logic appropriately addresses numerous situations, such as the treatment of undercutting defaults, I have demonstrated that on its own it only offers an imperfect solution for cases involving floating conclusions. The framework accepts floating conclusions with certainty, while I have illustrated in the previous chapter that their acceptance heavily relies on the context. This highlights the framework's limitations, given its closed-world assumption, as it cannot represent dynamic reasoning processes needed for incorporating inquiring states of mind. Such processes would possibly enable a more suitable treatment of

floating conclusions.

The presented framework and the logical investigations provide some initial guidelines for further research in this direction. Further research could involve expanding the framework to accommodate not only defaults between propositions alone but also defaults whose premise or conclusion are already a second-order attitude. To do this, we would need to extend the background language to include $\mathcal{L}$ as well as all possible combinations of the four second-order attitudes along with a proposition of $\mathcal{L}$. Once we can construct defaults with second-order propositions, we could establish rules that introduce specific second-order defaults in the presence of particular situations. For instance, in the case of floating conclusions, a rule might be introduced stating that if a proposition f follows exclusively from defaults whose premises are suspended, then f cannot be certain. In this context, priorities could be added to the framework to ensure that these second-order rules always override first-order rules. This might lead to floating conclusions being only believable and not certain in the consequences.

In the next chapter (Chapter 5), I will delve into the concept of suspension within a different framework of logic-based AI. I will explain the framework of *abstract argumentation theory* and examine suspension not about propositions but about entire arguments themselves. At the end of this chapter, in Subsection 5.3.2, I will demonstrate how default logic can be integrated into abstract argumentation theory. Through this translation, it will once again become apparent that the unadapted default logic presents a binary perspective, leaving no space for doxastic neutrality.

4.4.1 Answers to the Research Questions

1. Does the considered framework allow for a way to deal with conflicting or uncertain information?

   State-of-the-art default logic lacks a mechanism to effectively deal with conflicts. The existing options, namely choice, credulous, and skeptical reasoning, are primarily designed to bypass conflicts and facilitate uninterrupted reasoning rather than to explicitly handle

or communicate conflicting situations. This limitation has been addressed through my proposal for an adapted default theory, which adeptly stores conflicting information in the resulting consequences.

2. Is there something in the light of suspension of judgment present in the framework?

   I deliberately modified the framework to explicitly incorporate suspension of judgment, represented as a second-order attitude, within the consequences. This attitude is used to signify propositions for which we possess conflicting evidence. Representing suspension as a second-order attitude effectively captures the demanding nature of qualified suspension, as described in Chapter 2. The representation aligns well with the meta-cognitive accounts of suspension.

3. Can we find and distinguish different forms and epistemological norms of doxastic neutrality in the framework?

   In my adapted version of the default theory framework, the definition of the consequences allows for the representation of both a suspension attitude and a distinct attitude of (deep) ignorance. Both these attitudes indicate neutrality towards a proposition. Suspension comes into play when there is conflicting evidence (both positive and negative) for a proposition, while ignorance encompasses cases where we lack evidence for and evidence against a proposition. Ignorance also includes propositions that are the conclusions of undercut defaults.
   The Balance Norm is particularly relevant for the cases of suspension. The Absence Norm is more applicable to cases of ignorance. Positive justification of suspension can only be said to emerge in undercutting cases. The premise of the default that undercuts can be seen as a positive reason for adopting a neutral position (in this case, ignorance instead of suspension) towards the conclusion of the undercut default. However, it is worth noting that representing ignorance using a second-order attitude does not align perfectly with the philosophical description of this state.

Chapter 5

Argumentation Theory

This chapter is to a large extent based on and taken over from my paper (Schuster, 2021).

Contents

5.1 Introduction

In this chapter, I will delve into the framework of formal argumentation theory. Argumentation theory can be seen as a component of non-monotonic reasoning, which again is a crucial subfield of logic-based artificial intelligence. As outlined in the Introduction (Chapter 1), systems equipped with non-monotonic reasoning capabilities can effectively simulate human everyday reasoning, allowing for the inference of new knowledge beyond the initially provided knowledge base.

My primary focus in this chapter will be on abstract argumentation theory. Abstract argumentation theory revolves around the idea that non-monotonic reasoning can be aptly described by modeling arguments and their relations. Unlike other types of arguments, such as mathematical proofs, the arguments described in abstract argumentation theory are defeasible, meaning they can be attacked or defeated by opposing arguments (Baroni et al., 2011).

The key objective of abstract argumentation theory is to establish a framework for modeling and evaluating these arguments and their relationships. This is done via an argumentation framework. It is important to note that while abstract argumentation theory addresses the modeling and evaluation of arguments, this is just one part of the entire non-monotonic reasoning process, as outlined by Baroni et al. (2011). The complete process involves creating a knowledge base from which inferences can be drawn, visualizing these inferences as arguments that can attack each other, and evaluating which arguments should be accepted. This last part is what is covered by abstract argumentation theory. A final step involves defining the consequences of the evaluated and accepted arguments.

Beyond the *modeling* of arguments and their relationships, the second main objective of abstract argumentation theory is to *evaluate* the arguments. Given an argument framework, the overarching goal is to determine which arguments from a given set should be accepted and which should be rejected. This yields an extension – a set of accepted arguments – and an antiextension – a set of rejected arguments. Additionally, it is possible to distill a third set of arguments, which is in certain approaches to abstract argumentation theory not characterized explicitly. This is the set of

arguments one is undecided about. The involvement of indecision is made explicit in the so-called labeling-based approach of abstract argumentation theory, where arguments are labeled according to three different labels: `in`, `out`, and `undecided`.

Very often indecision is only seen as a quite useful but rather unimportant byproduct when characterizing the acceptance and rejection of arguments. This thesis, though, aims to put indecision at the center of investigation. From a philosophical point of view, indecision (or suspension) represents the third main doxastic response, besides acceptance and rejection, and it should be taken seriously. Philosophical investigations from both epistemology and philosophy of mind can be useful when applied to argumentation theory. Investigating the notion of indecision that is used in argumentation theory more precisely and comparing it to the philosophical concepts of doxastic neutrality can help us to observe the different options of how to use indecision as a tool to describe uncertain, doubtful, or conflicting information. This will allow us to find ways to improve the representation of the given knowledge and to make certain decisions appear less bold and more understandable and trustworthy.

In this chapter, I will transfer the philosophical considerations of Chapter 2 to argumentation theory and thereby reveal important parallels. I want to illustrate by what means the various semantics of abstract argumentation theory treat indecision differently and how this relates to the epistemological debate about rationality norms, which I investigated in Section 2.2. Next, I will shed some light on the different forms of indecision that can be found in abstract argumentation theory and on how they correspond to particular philosophical phenomena that I described in Section 2.3. Moreover, some considerations from the overlapping field between epistemology and philosophy of mind from Section 2.4 will be applied, too.

In the context of argumentation theory, the term “indecision” is explicitly used. Therefore, throughout this chapter, I will use the term “indecision” when discussing concepts within argumentation theory. However, when addressing philosophical concepts, I will encompass the broader spectrum of all doxastically neutral states. Specifically, I will use the term “suspension,” as introduced in Chapter 2, when explicitly

referring to philosophical phenomena. Additionally, I will employ other terms such as "non-belief" or "ignorance," as introduced in Chapter 2, whenever they are applicable.

In a second step, I will delve into the more structured level of statement-labeling, by considering not only arguments as such but also their conclusions. Analyzing the conclusions of arguments represents the concluding step in the non-monotonic reasoning process. I will briefly examine various approaches from argumentation theory to label statements that serve as the conclusions of arguments. This investigation will uncover how certain philosophical norms and forms of indecision, as discussed in Chapter 2, manifest not only at the level of arguments but also in the statements themselves.

While philosophical ideas may align more smoothly with statements than with arguments, this chapter's primary focus remains on abstract argumentation theory. I have already discussed the statement-based framework of default logic in the previous chapter. When we abstract away from propositions in default logic and only consider defaults, it is possible to translate a default theory into an abstract argumentation framework (Dung, 1995). As such, default logic can be seen as a variant of "structured argumentation." Since I have extensively examined the potential for suspension at the statement level in the context of default theory, this chapter will shift its focus to the more abstract realm of assessing arguments. Only in the last part of the chapter will I address approaches related to labeling statements and reveal more precisely how a default theory can be translated into an argumentation framework.

The chapter is structured in the following way. In Section 5.2 possible transfers from the philosophical investigations presented in Chapter 2 to the area of argumentation theory will be suggested. Parallels will be drawn between suspension in philosophy and indecision in abstract argumentation theory. Here, I am operating on the level of arguments. For this, some formal backgrounds will be introduced (Subsection 5.2.1) and the philosophical considerations will be applied (Subsection 5.2.2). In a second step, in Section 5.3, I will consider parallels between suspension in philosophy and indecision on the statement level of argumentation theory

and show how default logic can be translated into argumentation theory.

5.2 Indecision on the Level of Arguments

5.2.1 Argumentation Theory Background

Before I can shed some light on the parallels between indecision in abstract argumentation theory and suspension in philosophy, some basic notions of abstract argumentation theory have to be introduced. Still, this subsection does not aim to provide a complete overview of the definitions and theorems of abstract argumentation theory. Only the parts that are directly relevant for the considerations concerning indecision are presented. The subsequent definitions are conventional within the domain of abstract argumentation theory, which traces back to the ideas of Dung (1995), and can for example be found in Baroni et al. (2011).

Abstract argumentation theory focuses on modeling arguments and their relation of attack on an abstract level. This is done by introducing argumentation frameworks, which allow for an illustration of arguments (nodes of a graph) and their attack (edges of the graph). More precise features, such as the internal structure of the arguments and how they attack each other, are not representable at this abstract level.

Definition 5.1 (Argumentation Framework)**.** An Abstract Argumentation Framework is a pair (Ar, att), where Ar is a set of arguments and $att \subseteq Ar \times Ar$ is a relation of attack between the arguments.

Example 5.1. The argumentation framework (Ar, att) with $Ar = \{A, B, C, D\}$ and $att = \{(A, B), (D, A)\}$ can be represented as shown in Figure 5.1.

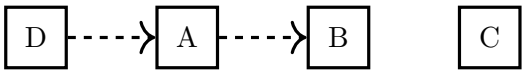


Figure 5.1: Graph of argumentation framework for Example 5.1 with $Ar = \{A, B, C, D\}$ and $att = \{(A, B), (D, A)\}$.

Besides the modeling challenge, abstract argumentation theory also deals with evaluating arguments in order to choose a proper subset of acceptable

arguments among the modeled ones.
In general, there are two approaches to this: the extension-based approach and the labeling-based approach. The labeling-based approach provides a function that maps each argument to a label. The three options for a label are `in`, `out`, or `undec` (undecided).

Definition 5.2 (Labeling)**.** Given an argumentation framework (Ar, att) and the set of labels, {`in, out, undec`}, a labeling is a function $Lab : Ar \rightarrow$ {`in, out, undec`} which maps each argument to one of the three possible labels.

In contrast to this, there is the extension-based approach. The extension-based approach simply yields a subset of the considered arguments (the extension),[1] which consists of the accepted arguments.

Definition 5.3 (Extension)**.** Given an argumentation framework (Ar, att), an extension $\mathcal{E}^{\mathrm{A}}$ is a subset of the arguments, $\mathcal{E}^{\mathrm{A}} \subseteq Ar$ that are taken to be accepted.

Unlike the terminology employed in Chapter 4 regarding default logic, in abstract argumentation theory, the term "extension" denotes a set of arguments, specifically the set of accepted arguments. In default logic, an extension is a set of propositions that *follow* from a set of accepted defaults. A set of accepted defaults (or arguments) in default logic is termed a "scenario." As we will see in Definitions 5.4-5.8, akin to default logic, an argumentation framework in argumentation theory typically gives rise to more than one possible extension.

The extension-based approach and the labeling-based approach are convertible into each other (Baroni et al., 2011) via

$$\mathcal{E}^{\mathrm{A}} = \mathtt{in}(Lab)$$

for a given argumentation framework. With the notation $\mathtt{in}(Lab)$, I refer to the arguments labeled as `in` according to the labeling Lab. This could also be denoted as $Lab^{-1}(\mathtt{in})$ (and similarly for `out` and `undec`). I use the notation $\mathtt{in}(Lab)$, following the definitions of Baroni et al. (2011, Def. 7), which is based on the correspondence between the extension-based and the

[1] Since extensions occur both in default logic and argumentation theory, I will refer to the extensions of argumentation theory with $\mathcal{E}^{\mathrm{A}}$.

labeling-based approach.

This means that the arguments that are labeled `in` are exactly those that are included in the extension. The arguments that are rejected or undecided are then defined in terms of the extension. Rejected arguments (`out`-labeled arguments) are those that are attacked by (at least) one member of the extension and undecided arguments (`undec`-labeled arguments) are those that are neither included in the extension nor attacked by it. Although the approaches and the subsequent definitions and results can be formulated both in extension-based and in labeling-based terms, (Baroni et al., 2011), I will focus on the labeling-based approach in this chapter, since the labeling-based approach provides a straightforward way to distinguish the three possible states of an argument and includes an undecided evaluation explicitly.

Semantics

The challenge for argumentation theory is to determine which label each argument should receive. This decision is not arbitrary but must follow specific rules, which are provided by different *semantics*. Starting from the set of all possible labelings, i.e., all possible combinations of labels for the arguments in the argumentation framework, these rules lead to a subset of possible labelings that meet the requirements of the respective semantics. The semantics I will present in the following are well-established and can be found in works such as Baroni et al. (2011). While some semantics are more general, allowing for more possible labelings, others are more stringent, resulting in only a few or possibly even a unique labeling. The most general semantics is the conflict-free semantics.

Definition 5.4 (Conflict-free Semantics)**.** Let (Ar, att) be an argumentation framework. A labeling Lab is called a conflict-free labeling iff the following two conditions hold for every argument $A \in Ar$

(i) if A is labeled `in`, then there exists no $B \in Ar$ that attacks A and is also labeled `in`,

(ii) if A is labeled `out`, then there is at least one argument $B \in Ar$ that attacks A and is labeled `in`.

The idea of conflict-freeness can be captured even better in the extension-based approach. Here, an extension is said to be conflict-free if every argument in the extension does not have an attacker in the extension. The extension set is, so to speak, coherent.[2] Slightly more advanced but still rather basic rules are presented by admissible semantics.

Definition 5.5 (Admissible Semantics)**.** Let (Ar, att) be an argumentation framework. A labeling Lab is called an admissible labeling (or a labeling according to admissible semantics), iff the following two conditions hold:

(i) every `in`-labeled argument $A \in Ar$ is legally `in`, i.e., $\forall B \in Ar$: if $(B, A) \in att$ then $Lab(B) =$ `out`,

(ii) every `out`-labeled argument $A \in Ar$ is legally `out`, i.e., $\exists B \in Ar$, such that $(B, A) \in att$ and $Lab(B) =$ `in`.

Admissible semantics demand that only arguments that are *legally* `in` or `out`, should get the respective label. An argument is legally `in` if all its attackers (if there even are any) are labeled `out`. An argument is legally `out` if it has at least one attacker that is labeled `in`. It can be noted that Condition (ii) of legally `out` arguments is already met by a conflict-free labeling. Condition (i) about which arguments can be legally `in` is weaker in the conflict-free labeling. The difference is that for an argument to be labeled `in` according to admissible semantics, *all* of its attackers have to be labeled `out`. In conflict-free semantics, *none* of its attackers can be labeled `in`, i.e., they can be labeled `out` or they can be labeled `undec`.

Another semantics, that I will introduce here is complete semantics. It differs from admissible semantics in some important manner, and I will discuss these differences from a philosophical point of view in Part 5.2.2.a.

Definition 5.6 (Complete Semantics)**.** Let (Ar, att) be an argumentation framework. A labeling Lab is called a complete labeling (or a labeling according to complete semantics), iff the following three conditions hold:

(i) every `in`-labeled argument $A \in Ar$ is legally `in`, i.e., $\forall B \in Ar$: if $(B, A) \in att$ then $Lab(B) =$ `out`,

[2] This rules out self-attacking arguments.

(ii) every `out`-labeled argument $A \in Ar$ is legally `out`, i.e., $\exists B \in Ar$, such that $(B, A) \in att$ and $Lab(B) = \texttt{in}$,

(iii) every `undec`-labeled argument $A \in Ar$ is legally `undec`, i.e., $\exists B \in Ar$, such that $(B, A) \in att$ and $Lab(B) \neq \texttt{out}$ and $\forall B \in Ar$: if $(B, A) \in att$ then $Lab(B) \neq \texttt{in}$.

The set of complete labelings is a subset of the set of admissible labelings. Complete semantics fulfills the requirements of admissible semantics and fulfills on top of that the requirement that every `undec`-labeled argument has to be also legally `undec`. An argument is legally `undec`, iff it is neither legally `in` nor legally `out`. This means that an argument is legally `undec` iff *none* of its attackers are labeled `in` and it has at least one attacker that is *not* labeled `out`.

Many other semantics build on complete semantics and add more conditions. I will introduce two refinements of complete semantics, grounded semantics and preferred semantics.

Definition 5.7 (Grounded Semantics)**.** Let (Ar, att) be an argumentation framework. A labeling Lab is called a grounded labeling (or a labeling according to grounded semantics), iff the following conditions hold:

(i) Lab is a complete labeling,

(ii) the set $\texttt{in}(Lab)$ is minimal with respect to set inclusion among all complete labelings.

Condition (ii) of this definition is equivalent to the set $\texttt{out}(Lab)$ being minimal and equivalent to the set $\texttt{undec}(Lab)$ being maximal with respect to set inclusion. This means that in the grounded labeling, only those arguments are labeled `in` that are labeled `in` in every complete labeling. In this sense, grounded semantics are the most cautious semantics. Moreover, the grounded labeling is always unique (Baroni et al., 2011, p. 11).

The second refinement of complete labelings, which is in a sense contrary to the idea of grounded semantics, is preferred semantics.

Definition 5.8 (Preferred Semantics)**.** Let (Ar, att) be an argumentation framework. A labeling Lab is called a preferred labeling (or a labeling according to preferred semantics), iff the following conditions hold:

(i) Lab is a complete labeling,

(ii) the set $\texttt{in}(Lab)$ is maximal with respect to set inclusion among all complete labelings.

The idea of preferred labeling is to accept as many arguments as possible (in extension-based terms: to maximize the extension set with respect to set inclusion). This usually yields many possible preferred labelings. Since only complete labelings are considered, the sets are in particular conflict-free. Condition (ii) of the definition, the set $\texttt{in}(Lab)$ being maximal, is equivalent to $\texttt{out}(Lab)$ being maximal.

One refinement of preferred semantics that is interesting from the prospective of indecision is stable semantics.

Definition 5.9 (Stable Semantics)**.** Let (Ar, att) be an argumentation framework. A labeling Lab is called a stable labeling (or a labeling according to stable semantics), iff the following conditions hold:

(i) Lab is a complete labeling.[3]

(ii) the set $\texttt{undec}(Lab) = \emptyset$.

Admissible, complete, grounded, preferred, and stable semantics are only five of the many possible semantics that have been investigated in argumentation theory. An overview of the most important semantics is presented in Figure 5.2, reproduced from Baroni et al. (2011, p. 24):

[3]Note that Lab being a preferred labeling would yield an equivalent definition.

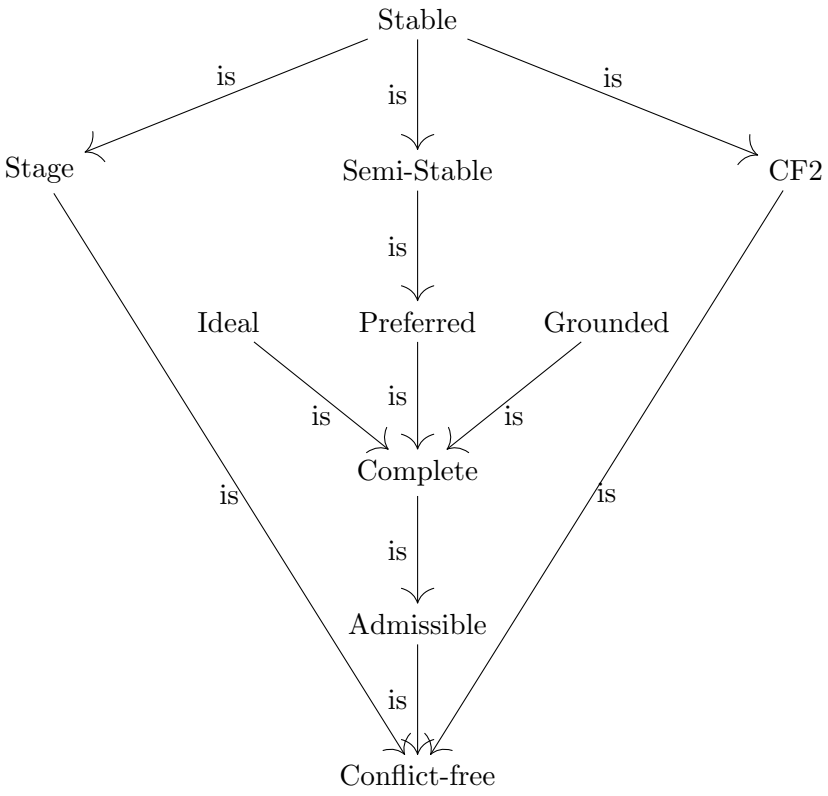


Figure 5.2: Relations of different semantics in abstract argumentation theory.

Some of these semantics (like conflict-free, preferred, grounded, or stable) were already defined in the original paper of Dung (1995). We can observe that the conflict-free labelings are the most fundamental, followed by admissible labelings. Furthermore, it is worth noting that all the other semantics I have explicitly defined above are a subclass of admissible (and complete) labelings.
To illustrate that all grounded and preferred labels are complete, that all complete ones are admissible, and that they all ultimately trace back to being conflict-free, we can make use of an example.

Example 5.2. This example is (in a slightly different version) presented in (Wu et al., 2010, p. 21). I will later use it again, in an extended form, to illustrate subsequent definitions. Imagine you are introduced to a group of friends, and you are trying to evaluate the different people and their relationships. You get the following information:

E: Gina hates Harold according to Emily.
F: Gina does not know Harold according to Fred.
G: Gina says that Harold is not reliable.
H: Harold is generally considered trustworthy.

The argumentation graph of the argument is be visualized in Figure 5.3.

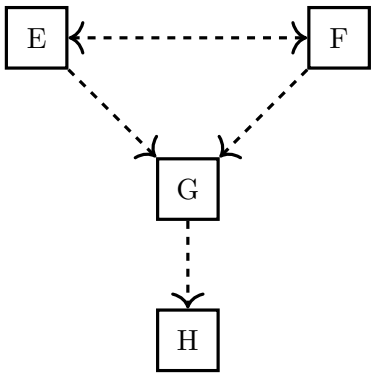


Figure 5.3: Argumentation Graph of Example 5.2.

The following tables show the labelings that are only conflict-free, the labelings that are conflict-free and admissible, and the labelings that are conflict-free, admissible, and complete. In the complete labelings, we can also see two labelings that are also preferred and stable and one labeling that is also grounded.

E	*F*	*G*	*H*
undec	in	out	undec
undec	in	out	in
undec	in	undec	undec
undec	in	undec	in
in	undec	out	undec
in	undec	out	in
in	undec	undec	undec
in	undec	undec	in
undec	undec	in	undec
undec	undec	in	out
undec	undec	undec	in
in	out	undec	in
out	in	undec	in

Table 5.1: Labelings of Example 5.2 that are *only* conflict-free, i.e., not admissible.

E	F	G	H
`in`	`out`	`undec`	`undec`
`in`	`out`	`out`	`undec`
`out`	`in`	`undec`	`undec`
`out`	`in`	`out`	`undec`

Table 5.2: Labelings of Example 5.2 that are conflict-free and admissible, but not complete.

E	F	G	H
`in`	`out`	`out`	`in`
`out`	`in`	`out`	`in`
`undec`	`undec`	`undec`	`undec`

Table 5.3: Labelings of Example 5.2 that are conflict-free, admissible, and complete. The first two rows show labelings that are also preferred and stable, and the last row shows the only labeling that is also grounded.

We see that we have three complete labelings in this example from which two are preferred and one is grounded. Seven labelings are admissible and the requirements of conflict-free semantics are fulfilled by 20 labelings. A not conflict-free labeling would, for example, be a labeling in which all arguments are labeled `in`, in which all arguments are labeled `out`, or in which argument E is labeled `out` although F is not labeled `in` or vice versa.

All semantics make demands on the labelings of the arguments. Some semantics (e.g., grounded semantics) only yield one allowed labeling.

Most semantics, like admissible, complete, preferred, or stable semantics, however, usually allow for a plurality of possible labelings. In such situations, the question arises of how the different possible labelings can be synthesized to yield a unique solution of the argumentation framework.[4]

Justification Status

This question can be framed as the question about the *justification status* of an argument. If an argument is, for example, labeled `in` according to one complete labeling and labeled `out` in another one, how should we evaluate that argument after all?
Analogous to the situation in default logic, the two simplest variants to synthesize the multiple results are to form either the union or the intersection of the resulting sets. According to the proposals in default logic, one speaks of a credulous (forming the union) and a skeptical (forming the intersection) procedure. The same terms are used in the realm of argumentation theory. In the extension-based picture, the skeptical approach consists of forming the intersection of all extensions. Therefore, only arguments that are accepted in all extensions (i.e., extensions that are labeled `in` according to all labelings) are in turn labeled `in` according to the skeptical synthesis. Analogously, only those arguments are labeled `out` that are attacked by every extension (i.e., attacked by the intersection). All other arguments are `undec`. According to the credulous approach, the union of all extensions is used. In labeling terminology: All arguments that are labeled `in` according to at least one labeling are labeled `in` according to the credulous policy. Whether an argument is labeled `out` (which is in this sense often referred to as a strong form of reject) or labeled `undec` (which is in this sense referred to as a weak reject) will be determined by whether it is attacked by an `in`-labeled argument or not.[5]

[4] The same question appeared in the context of default logic when different, conflicting extensions of propositions resulted. I defined the consequences Γ of a default theory in Subsection 4.3.2 for this purpose.

[5] It is essential to note that in the domain of argumentation theory, the term "credulous" carries different usages. Credulous reasoning is frequently used *within* a specific semantics. For instance, credulous preferred reasoning might imply selecting a single preferred labeling at random. In alternative contexts, credulous reasoning refers to a multi-labeling approach where the various resulting labels are not further consolidated or synthesized.

It quickly becomes apparent that the two approaches group many arguments under the same justification status, eliminating the distinctions between arguments that play distinct roles in their respective frameworks and extensions. In Example 5.2, considering only admissible labelings, the skeptical approach would yield the following $Lab_S(E) = Lab_S(F) = Lab_S(G) = Lab_S(H) =$ `undec`, the credulous approach yields $Lab_C(E) = Lab_C(F) = Lab_C(H) =$ `in` and $Lab_C(G) =$ `out`. Besides other problems (like not conflict-free results for the credulous approach), one can see that a lot of information is lost by reducing the labels in this way. For example, in the skeptical synthesis, argument E and H have the same status, although H is not labeled `out` in any labeling, but E is labeled `out` according to three labelings.

A more detailed analysis of different justification statuses is provided in Baroni et al. (2004) and Wu et al. (2010). Baroni et al. (2004) have the aim to compare the different semantics according to their degree of skepticism. For this, they define two different skepticism relations between different semantics. The relations are defined with the help of the justification status that a particular argument is assigned according to a semantics. Their basic idea is that a semantics S_1 is less skeptical than a semantics S_2 if the justification status of every argument in S_1 is "more committing" than the one in S_2. For this, they define seven different justification statuses and order them on a semi-lattice. Similarly, Wu et al. (2010) define different justification statuses. Wu et al. (2010) explicitly consider complete semantics and define the possible justification status an argument can have in complete semantics. They suggest a function that maps each argument to its justification status, which they take to be the set of all possible labels each argument gets from the different complete labelings. Since the representation of Wu et al. (2010) is clearer, I will define the justification status according to their terminology in the following.

Definition 5.10 (Justification Status)**.** Let (Ar, att) be an argumentation framework. For $A \in Ar$, let $J(A)$ be the justification status of A, given by the function $J : Ar \to \mathcal{P}(\{$`in, out, undec`$\})$ (Wu et al., 2010, p. 16).

J maps every $A \in Ar$ to a subset of {`in, out, undec`}, consisting of all the labels A gets by one or more complete labeling. Considering complete

semantics, in Wu et al. (2010, p. 16), 6 possible justification statuses are obtained: {`in, out, undec`}, {`in, undec`}, {`out, undec`}, {`in`}, {`out`}, {`undec`}. Note that the remaining option {`in, out`} is not considered in Wu et al. (2010), as they consider only complete semantics which do have the characteristics of being "abstention allowing," meaning that if there is a complete labeling that labels A `in` and there is another complete labeling that labels A `out`, there has to be a third one that labels A `undec` (Baroni et al., 2011, p. 27). As Baroni et al. (2004) do not restrict their definition to complete semantics, the status {`in, out`} is also considered. Otherwise, the justification statuses defined in Wu et al. (2010) correspond to the ones from Baroni et al. (2004). The empty set is not considered in either approach, since only non-empty labeling sets are taken into account.

If we extend Example 5.2 slightly, we can see how the different justification statuses come into play.

Example 5.3. This example is an extended version of Example 5.2.
A: Alice says that Carole is a Liar.
C: Carole says that David is a Liar.
D: David says that Alice is a Liar.
B: Everybody agrees that Bob is really nice.
E: Gina hates Harold according to Emily.
F: Gina does not know Harold according to Fred.
G: Gina says that Harold is not reliable.
H: Harold is generally considered trustworthy.

The argumentation graph of the argument can be visualized like this:

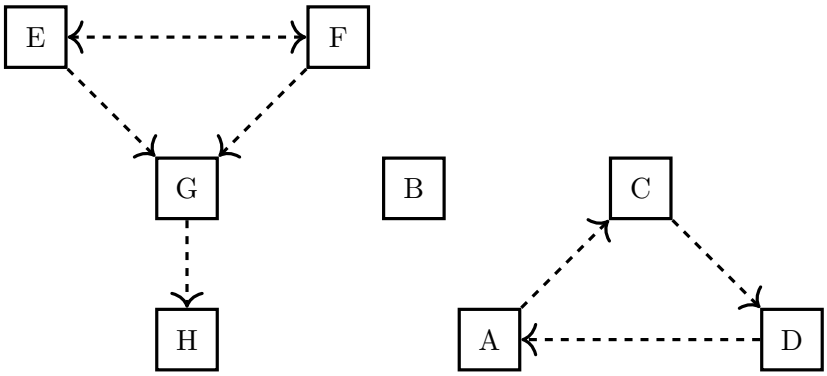


Figure 5.4: Argumentation Graph of Example 5.3.

The justification statuses of the arguments (according to complete semantics) are:

$$
\begin{aligned}
J(A) &= \{\texttt{undec}\},\\
J(B) &= \{\texttt{in}\},\\
J(C) &= \{\texttt{undec}\},\\
J(D) &= \{\texttt{undec}\},\\
J(E) &= \{\texttt{in, out, undec}\},\\
J(F) &= \{\texttt{in, out, undec}\},\\
J(G) &= \{\texttt{out, undec}\},\\
J(H) &= \{\texttt{in, undec}\}.
\end{aligned}
$$

If there were an argument that was attacked by B, then this argument would get the justification status $\{\texttt{out}\}$.

According to Baroni et al. (2004), the different justification statuses can be ordered in the following semi-lattice (translated into the terminology of Wu et al. (2010)), see Figure 5.5.

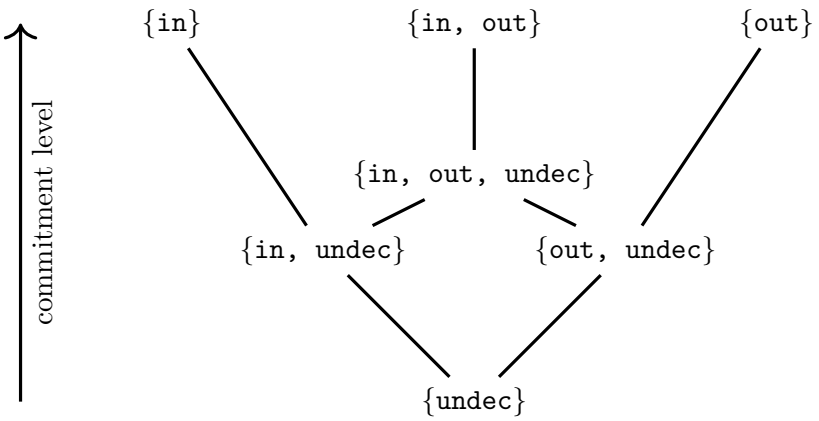


Figure 5.5: Semi-lattice of justification statuses by Baroni et al. (2004).

This semi-lattice is meant to represent different levels of commitment, with the highest commitment at the top of the figure and the lowest commitment at the bottom.

Non-standard approaches

The definitions provided above can all be characterized as belonging to the standard approach of abstract argumentation theory.[6] In the following, I also want to introduce two non-standard ways of labeling arguments. The first approach from Jakobovits and Vermeir (1999) allows for partial labeling functions, leaving some arguments "unlabeled." The second approach from Arieli (2016) enriches the possible labelings by allowing four instead of only three possible labels.

Definition 5.11 (Partial labeling)**.** A labeling *Lab* is called partial on an argumentation framework (Ar, att), if there exist some arguments $A \in Ar$, for which $Lab(A)$, is not defined.

Alternatively, a partial labeling can be defined by introducing a fourth possible label $\emptyset$, which is assigned to those arguments that do not get one of the three ordinary labels. A partial labeling function, as introduced in Jakobovits and Vermeir (1999), does not necessarily map each argument to one of the three labels: `in`, `out`, `undec`. This is motivated by the idea

[6] Although the approach of providing a justification status can be seen as an advanced alternative to the more simple credulous/skeptical synthesis approaches.

that sometimes, it is not necessary or desirable to label all arguments of a given argumentation framework, since only parts of the framework might be of interest. There are different motivations for allowing this option, from saving computational resources to the virtue of modesty, which leads to providing a clear opinion only if it is really required (Baroni et al., 2015, p. 268). If in a partial labeling an argument $A \in Ar$ is not labeled, this can be represented by writing $Lab(A) = \emptyset$. The empty set is introduced as an additional "don't care" label that applies to arguments for which the reasoner does not care to provide a label for (Baroni et al., 2015).

In Example 5.3, various possible partial labelings can be provided. Without delving into details, it should be noted that there are also legality restrictions for when an argument can be labeled as "don't care." In the approach of Jakobovits and Vermeir (1999) the restrictions intuitively demand that if you don't care about an argument, you should also not care about the arguments that are affecting it or are affected by it (Baroni et al., 2015). It is always possible to abstain from labeling *any* argument or abstaining from all arguments that are involved in an isolated sub-framework. For example, it would be possible to abstain on A, C, and D while labeling all other arguments from the framework in an ordinary way.

A different variant of the standard labeling-based approach consists in providing a labeling function that, albeit being a complete function, maps the arguments to four different labels.

Definition 5.12 (Four-Valued Labeling)**.** A four-valued labeling is a labeling function $Lab : Ar \rightarrow \{\texttt{in}, \texttt{out}, \texttt{none}, \texttt{both}\}$ that maps each argument $A \in Ar$ to one of the four possible labels.

The four-valued labeling introduced in Arieli (2016) is somewhat analogous to the partial labeling presented in Definition 5.11. The partial labeling can also be viewed as mapping each argument to one of four labels (considering "don't care" as the fourth label). However, there are distinctions in the general concept and the rules about how the labels are assigned. While the partial labeling aims to exclude parts of the graph entirely, the four-valued labeling seeks to provide further differentiation based on different situations of attacks by different arguments.

Furthermore, the four-valued labeling of Arieli (2016) is a *paraconsistent* approach. When translated into the extension-based framework, an argument A is labeled

- `both`, if A is in the extension $\mathcal{E}^A$ and an attacker B of A is also in $\mathcal{E}^A$,
- `none`, if neither A nor any of its attackers are in $\mathcal{E}^A$,
- `in`, if $A \in \mathcal{E}^A$ and none of its attackers are in $\mathcal{E}^A$, and
- `out`, if A is not in $\mathcal{E}^A$ but at least one of its attackers is in $\mathcal{E}^A$.

Due to the definition of `both`, the extension is not conflict-free anymore, see Arieli (2016, p. 5) or Baroni et al. (2015, p. 274).

As in classical labeling semantics, there are also different options to provide semantics for the four-valued labelings. According to the four-valued complete semantics, provided by Arieli (2016, p. 7), an argument is

- `in` if all of its attackers are `out`,
- `out` if there is either an attacker that is `in`, or there is one attacker that is `both` and one attacker that is `none`,
- `both` if all its attackers are either `out` or `both` and at least one is `both`, and
- `none` if all its attackers are either `out` or `none` and at least one is `none`.

The complete four-valued labelings of the Example 5.2 are:

E	*F*	*G*	*H*
`none`	`none`	`none`	`none`
`both`	`both`	`both`	`both`
`in`	`out`	`out`	`in`
`out`	`in`	`out`	`in`

Table 5.4: Complete four-valued labelings of Example 5.2.

In the extended version of Example 5.3, the argument B is labeled `in` according to every complete four-valued labeling and the arguments C, D, and A of the odd attacking cycle are either all `none` or all `both` in the different complete four-valued labelings. The four-valued labeling, alongside the partial labeling and the standard approach in abstract argumentation, will be relevant in Subsection 5.2.2 as I apply the philosophical findings and explore the various forms and norms of indecision present in these approaches.

5.2.2 Philosophical Interpretation

As stated in the introduction, the main goal of this chapter is to link the usage of indecision in argumentation theory with the philosophical concept of suspension and possibly other concepts of doxastic neutrality, which I introduced in Chapter 2. When we examine this within abstract argumentation theory, we are focusing on arguments only. This framework solely concentrates on arguments and how they relate to each other. It does not delve into the details of arguments involving specific statements or propositions. The terms "acceptance," "rejection," and "indecision" relate to arguments. In contrast, when we talk about suspension in philosophy, suspension revolves around a proposition p.[7] An argument and a proposition are structurally different. Hence, the various ways

[7]Or, as discussed in Chapter 2, it could be about a question whose possible answers can be expressed as propositions.

and norms tied to doxastic neutrality that philosophers explore cannot easily be applied to arguments in the realm of argumentation theory. Nevertheless, when we abstract and move away from specific objects, the mental states characterized by neutrality that philosophy describes exhibit certain similarities with the labels we assign to arguments in argumentation theory, so that it is fruitful to apply the considerations from Chapter 2 to the field of abstract argumentation theory.

In the forthcoming parts of this subsection, I will maintain the overarching structure of Chapter 2. Firstly, in Part 5.2.2.a, I will delve into the *epistemological* and *normative* dimensions, applying the findings from Section 2.2 to arguments within argumentation theory. Secondly, in Part 5.2.2.b, I will employ insights from *philosophy of mind*, as discussed in Section 2.3, to uncover the various forms of doxastic neutrality which are present in argumentation theory. Lastly, akin to the philosophical discourse, I will address *overlapping* considerations in Part 5.2.2.c, exploring aspects that do not exclusively align with epistemology or philosophy of mind as introduced in Section 2.4.

5.2.2.a Norms of Indecision on the Argument Level

In this section, I will explore the norms of suspension found for indecision in argumentation theory.[8] Before delving into specific norms, it is useful to state a general observation about the normative status of indecision in argumentation theory. Interestingly, the various semantics of argumentation theory, which delineate the rules for labeling arguments, take divergent perspectives on the normative profile of indecision. In some cases, indecision appears to be always permissible, while in others, it is constrained by certain normative requirements.

First, one can note that admissible semantics only formulates rules (or necessary conditions) about when an argument can be accepted (labeled `in`) or rejected (labeled `out`). There is no such requirement for when an

[8] As previously mentioned in Chapter 1, I predominantly use the term "suspension" in the philosophical domain, referring to the norms for suspension elaborated in Section 2.2. Within argumentation theory, I adhere to the term "indecision," as it is explicitly used in this context.

argument is labeled `undec`. A labeling that simply labels all arguments `undec` always fulfills the requirements of admissible semantics. This means that admissible semantics only provide norms (in the form of restrictions) for acceptance and rejection, but no norms about when an argument should or should not be regarded as undecided. Hence, the label `undec` is always suitable for any argument.

In contrast, complete semantics expand admissible semantics (as they adopt the necessary conditions on when an argument is legally `in` or `out`) but also provide restrictions on when an argument can be labeled `undec`. The treatment of the undecided label is what distinguishes complete semantics from admissible semantics. While in admissible semantics, the set of undecided arguments is basically only a collection of whatever is left over (because it does not suffice to be labeled `in` or `out`), in complete semantics, there are rules about when an argument can legally be called undecided.

The difference is also visible in Example 5.2. While the labeling in which every argument is labeled `undec` is admissible, this labeling is not complete. In particular, complete labelings forbid labeling the argument B `undec`, since it is not attacked by any other argument.

Hence, complete semantics demand that when there is good evidence for an argument, one should *not be undecided* but accept it, and when there is good evidence against an argument, one should not be undecided either but reject it.

Moreover, grounded semantics and preferred semantics manifest distinct attitudes towards the concept of indecision, too. While both are subtypes of complete semantics (allowing the labeling of arguments as `undec` only when warranted) grounded semantics is marked by its favor of indecision, whereas preferred semantics tries to avoid it. The cautious attitude expressed by grounded semantics employs indecision whenever conflicts concerning the acceptance or rejection of an argument come to the fore. In preferred semantics, conversely, indecision performs the role of a "last resort" label, to be avoided whenever possible.

An escalation of this strategy to avoid indecision can be observed in stable semantics. In this semantics, indecision is outright prohibited. Only labelings where each argument is either labeled `in` or `out` are considered stable. Therefore, for some argumentation frameworks, no stable labeling

exists. Baroni et al. (2011, p. 13) describe stable semantics as a semantics "in which there is no room for neutrality or shades of gray and everything is just black or white."

In conclusion, it is evident that various semantics in argumentation theory exhibit different attitudes towards the normative status of indecision per se. To some extent, a similar divergence in perspectives on suspension can also be observed in the philosophical literature. In Section 2.2, I discussed how epistemology neglected to explicitly consider suspension and its norms for a long time. Instead, suspension was often viewed simply as a lack of belief or disbelief, a default position for propositions that either failed to meet the criteria for belief and disbelief or had not yet been thoroughly investigated. This perspective shares, for example, some similarities with the treatment of indecision in admissible semantics. Only in recent years have epistemologists begun to explore specific norms regarding suspension, leading to the emergence of explicit criteria for when suspension is justified.

The different attitudes in the different semantics are summarized in Table 5.5.

Semantics	Perspective on indecision
Admissible	Indecision as a fallback
Complete	Indecision to be justified
Preferred	Indecision to be avoided
Grounded	Indecision appreciated

Table 5.5: Different semantics in argumentation theory and their normative perspective on indecision.

Beyond these general observations, additional insights for argumentation theory can be made when considering the explicit epistemological norms governing suspension, as outlined in Section 2.2. The two most prominent

norms for suspension that I explained in Chapter 2 are the Absence Norm and the Balance Norm. When exploring the various reasons for adopting an undecided position in argumentation theory, it is reasonable to consider approaches that accommodate different versions of an undecided label. The two prototypical norms for suspensions can be found when considering the four-valued labeling function, presented in Arieli (2016). Here, the author motivates the introduction of a four-valued labeling, amongst other things, with the idea that there are two very different situations in which an argument is assigned the label `undec` in standard argumentation theory. Thereof, "one case is that the reasoner abstains from having an opinion about an argument because there are no indications whether this argument should be accepted or rejected. Another case that may cause a neutral opinion is that there are simultaneous considerations for and against accepting a certain argument. These two cases should be distinguishable" (Arieli, 2016, p. 5). This finds a quite straightforward counterpart in the two norms considered in epistemology.

The Absence Norm states that one should suspend if one has no relevant information about a proposition. This is the case when one did not even think about a proposition, or one did think about it but lacks any relevant information to answer the question of whether the proposition is true. This norm is reflected in the `none` label of the four-valued labeling. An argument A is assigned this label if neither A itself nor any attackers of A are included in the extension.

In contrast, in the situations covered by the Balance Norm, one does have information that is relevant for deciding whether the proposition is true or not, but this information does not provide a clear recommendation. Here, suspension stems from the conflicting and (more or less) equally balanced evidence one has towards a proposition. This is comparable to the case when arguments are labeled `both` in the four-valued labeling approach. The indecision at issue here stems from a contradiction since an argument A is labeled `both` if A is in the extension but is also attacked by another argument from the extension. In Baroni et al. (2015, p. 274), the authors argue that in the case of a `none`-labeled argument, there is ignorance,[9] and

[9]I will come back in Part 5.2.2.b to how far this label corresponds to what was called "ignorance" in the debate within philosophy of mind.

the corresponding attitude of the reasoner is correctly described by saying "I don't know." Whereas, in the case of a `both`-labeled argument, the indecision "is due to contradiction rather than to ignorance. [...] Your reply should now be "I know too much" since you got an excess of (inconsistent) information."

The observation concerning the `both` and `none` labels in the four-valued labeling is illustrated in Table 5.6. It is important to note that the table is intended to offer a rough representation of a potential philosophical interpretation of the distinction and does not include the detailed explanations provided in the text above.

Label of Four-Valued Labeling	Philosophical Interpretation
`none` label	Absence Norm
`both` label	Balance Norm

Table 5.6: Four-valued labeling of argumentation theory and potential correspondence to different philosophical norms.

Both these norms provide us with what was called a *privative* justification for suspension in Subsection 2.2.1. In both cases, the evidential situation is such that neither belief nor disbelief, or in the picture of argumentation theory, neither acceptance nor rejection, is justified. What is not taken into consideration in the four-valued labeling are cases in which we have *positive evidence* for indecision. This is also reflected in the discussion of Baroni et al. (2015) about what they call epistemological undecidedness, and I will call *positive indecision*. The authors notice that the underlying idea behind the four-valued labeling approach (which is tightly connected to the Belnap-Dunn-logic (Belnap, 1977)) is that an agent is generally supposed to provide a binary and definite judgment. Only when aggregated, different forms of indecision result in a second step in the form of the two labels `both` and `none`.[10] This picture suggests that belief or acceptance and disbelief or rejection are somehow more fundamental than indecision and that every

[10]Similarly, justification statuses are supposed to aggregate the different labels.

proposition (or argument in the case of abstract argumentation theory) is supposed to be eventually rejected or accepted. The picture, however, does not accommodate the situations in which being undecided is *really the right thing to do* – not only because one has conflicting or no evidence for or against something, but because the evidence explicitly points towards indecision.

This more fundamental kind of indecision is called epistemological undecidedness in Baroni et al. (2015, p. 277). I find the term "epistemological undecidedness" to be misleading as it implies that indecision arises solely from epistemological deficiencies, such as a lack of evidence. However, this type of indecision encompasses cases where indecision arises from the nature of the object itself, rather than from the evidence available to the subject evaluating the object. Since this form of indecision is closely linked to positive justification, I will use the term "positive indecision" instead of "epistemological undecidedness," as suggested by Baroni et al. (2015).

To give an example of when such indecision is at work, the authors consider a situation in which you are looking for a weather forecast for a location "with a specially complex geography, such that no existing weather forecast model is applicable. Then, you are undecided about whether tomorrow will be sunny (indeed you have good fundamental reasons to be so) and this indecision is rather different from the ones considered above. First, it clearly does not arise from contradictory information: you can not certainly say 'I know too much'. Second, even if it bears some superficial resemblance with the case 'I don't know' represented by the [value `none`], it is really different" (Baroni et al., 2015, p. 277). They argue that the `none`-label represents cases for which there is no information at all. Whereas, in cases of positive indecision, as presented in the example, there is some information, but the information clearly indicates towards indecision.

Similarly, in the philosophical discussion, the prototypical cases for positively justified suspension are cases in which your suspending attitude is, in some sense, due to the constitution of the proposition itself (Zinke, 2021b; Feldman and Conee, 2018; Ferrari and Incurvati, 2022). If you, for example, think that a cup is a borderline case between being blue

and green, you have positive evidence that supports you in suspending towards the proposition that the cup is blue. If you are sure that the throwing of a coin is a fair chance process, you will be positively justified for being undecided about the proposition "This coin will land on heads." Suspension, here, seems to be the *correct attitude* towards the relevant proposition, independent of the subject and its evidential state. Other examples of this sort might include liar-like sentences or self-defeating arguments, like attacking loops in argumentation theory.

As I have argued in Section 2.2, this type of justification for suspension is fundamentally different from privative justification for suspension, which can be found in the Absence Norm or the Balance Norm. However, it is not possible to represent such a positive justification for indecision in argumentation theory. Baroni et al. (2015) argue that this additional form of indecision introduces a new kind of attack[11] and should be represented by introducing a new *basic label* $U!$ to the framework. This demand stems from the observation that the labels `both` and `none` arise, following the Belnap-Dunn-logic, through a combination of either both asserting the basic truth values T and F or asserting none of them. The idea is that the labels `both` and `none` are not basic but derived, whereas the label $U!$ is supposed to represent the fundamental form of intrinsic undecidedness.

Hence, to give justice to this form of intrinsic undecidedness, one would have to adjust the basic framework of argumentation theory. In those cases, one would want to represent not only attack and counterattack of the arguments, but also positive arguments for the $U!$ label. Such a framework would then be different from ordinary abstract argumentation frameworks where only attack is representable. A similar adjustment must be made when one wishes to properly represent cases of undercutting defeat (Pollock, 1995).

5.2.2.b Forms of Indecision on the Argument Level

In this Section, I will try to uncover how the different usages of indecision that are present in argumentation theory represent different forms of doxastic neutrality. For this, I will refer to the philosophical considerations in Section 2.3.

[11] This form of attack is in their view related to Pollock's undercutting defeater, see Pollock (1995).

The first observation of varying forms of indecision in argumentation theory emerges from the four-valued labeling system, concerning the labels `both` and `none`. While this distinction was already relevant for illustrating different norms of suspension in the previous section, it also influences the potential types of indecision. As outlined in Chapter 2, completely disentangling these two perspectives is not always feasible. Often, a particular normative reason for suspension closely aligns with a specific form of doxastic neutrality. Conversely, a particular form of doxastic neutrality may conflict with a specific normative reason.

In the four-valued labeling, it can be observed that the label `both` can only describe forms of indecision that have a certain level of commitment or engagement already. It is *not* possible that I am deeply ignorant about or in the state of mere non-belief towards an argument that is labeled `both`, as this label describes a situation in which I have evidence (in fact too much evidence) both for and against the argument. In contrast, the label `none` fits a very uncommitted state of neutrality and even the state of deep ignorance well, since having no evidence for or against an argument is perfectly compatible with not having considered it and being deeply ignorant about the argument.[12] This is visualized in Table 5.7. Note again that these tables are meant to give a rough overview and only suggests one possible philosophical interpretation. It is essential to consider them along the preceding discussions for a comprehensive understanding.

[12]Note that in Part 5.2.2.a, I have associated the label `none` with the Absence Norm. Concerning forms of neutrality, I here posit that the label `none` aligns with an uncommitted form of neutrality (akin to ignorance). However, it is crucial to recognize that the uncommitted neutrality (or especially the state of ignorance) and justification via the Absence Norm do not always coincide. If a subject is justified in being in a state of ignorance, the Absence Norm might be the sole suitable justification, as the Balance Norm typically cannot provide justification here. Still, being justified via the Absence Norm does not guarantee that the subject is in the state of ignorance. A subject might be justified in their *suspending attitude* via the Absence Norm, since they genuinely lack evidence for the proposition (or the argument).

Label in Four-Valued Labeling	Philosophical Interpretation
`none` label	Uncommitted neutrality
`both` label	Committed neutrality

Table 5.7: Four-valued labeling of argumentation theory and potential correspondence to different forms of doxastic neutrality from philosophy.

Besides these different forms in the four-valued labeling, some observations about the form of indecision in the standard approach can be made, too. It can be observed that the use of the indecision label in the standard approach already signifies a relatively engaged and committed form of neutrality within the spectrum presented in Figure 2.2 in Part 2.3.2.b. When we employ the `undec` label in argumentation theory, we are already making a choice to categorize this specific argument as undecided. In doing so, we *commit* to the state of indecision. This is particularly apparent when considering the regular `undec` label within the context of the partial labeling approach.

The idea behind partial labelings is that, beyond the possibility to label an argument either `in`, `out`, or `undec`, it is possible to label an argument not at all. This is interpreted as *abstaining from labeling* the argument or giving the argument a "don't care" label (Baroni et al., 2015, p. 269) and (Jakobovits and Vermeir, 1999, p. 6). In Jakobovits and Vermeir (1999), the authors present acceptable semantics for partial labelings. They provide rules for when a reasoner can abstain from an argument. These rules always allow for "total abstaining," but restrict the possibilities when abstaining is done only on some of the involved arguments (Baroni et al., 2015, p. 270).

In my view, there are two forms of neutrality involved in this picture. The "normal" indecision arises when an argument is assigned the label `undec`, and the additional neutrality, which is due to the abstaining or "don't care"-label option. When considering Example 5.3 on page 154 and the sub-frame consisting of arguments A, C, and D, it is possible, for example, to label all three arguments `undec`. Additionally, it is also possible to stay

neutral by not labeling the three arguments at all, because one might, for example, just not be interested in Alice, Carol, or David.

Within the spectrum of different forms of doxastic neutrality presented in Part 2.3.2.b, it is evident that the assignment of the "don't care" status to an argument would represent a less sophisticated and less engaged form of neutrality, while assigning the (regular) `undec` label signifies a more sophisticated form of neutrality (i.e., suspension), as it is a more committed position.

In some sense, this aligns with the expectations of scholars in philosophy regarding suspension. Labeling an argument as `undec` implies making a decision about the label for that argument and can be viewed as adopting a *positive attitude* towards the argument. This is also in line with the demands of philosophers like Friedman (2013c) and Wagner (2022) regarding a person who is said to be suspending judgment about a proposition.

In contrast, refraining from labeling an argument or assigning it the label "don't care" can be regarded as being in a state of mere non-belief or deep ignorance regarding that argument. The argument is not considered, akin to how A Example 2.1 in Part 2.3.2.b does not even consider the proposition concerning the vitamin C content of the guava fruit. A never even entertained a thought about it. In this sense, A "doesn't care" about ascribing a suitable truth-value to the proposition. Similarly, one might not care to provide a suitable label for a certain argument when using a partial labeling approach.[13] In contrast to the `none` label, it becomes even more apparent that we are encountering a genuine case of deep ignorance or mere non-belief in the partial labeling. This is emphasized by the fact that both mere non-belief and the "don't care" label are characterized purely negatively. Non-belief is the *absence* of believing and disbelieving. An argument gets the label "don't care" if it *does not* get any (ordinary) label. This is visualized in Table 5.8. Note again that these tables are

[13] Arguably, in order to care or not care about a proposition, one should at least be able to grasp the proposition. So the "don't care" label does not perfectly align with the examples of deep ignorance that I described in Part 2.3.2.a. In those examples, the subjects were deeply ignorant because they lacked either an understanding of the proposition or cognitive contact. The partial label does not symbolize a lack of understanding but still signifies that the respective argument is *ignored*, which can indeed be seen as not considering it and, therefore, lacking cognitive contact.

meant to give a rough overview and only suggests one possible philosophical interpretation.

Label in partial labeling	Philosophical Interpretation
no label	Ignorance
`undec` label	Committed neutrality

Table 5.8: Four-valued labeling of argumentation theory and potential correspondence to different forms of doxastic neutrality from philosophy.

It might be argued that abstaining on arguments, as introduced in Jakobovits and Vermeir (1999), does not correspond to any form of indecision at all. In some situations, the label of an argument A might be the same in all possible *completions* of the partial labeling. This means that every extension of the partial labeling to a full labeling would e.g., lead to labeling the argument A `in`, and still, it would be possible to abstain and label this argument as "don't care." Hence, the acceptance status of the argument can be regarded as being to some extent "determined," which does not fit our idea of being *undecided* (Baroni et al., 2015, p. 270). However, in this thesis, I introduced a very broad notion of doxastic neutrality towards a proposition (or an argument) which also includes neutrality due to carelessness. Although a reasoner might easily find out how to label an argument; if they do not care, they will remain neutral. In the same way, a proposition that a person is ignorant about might be easily decidable if the person just considered *any* evidence, but they can still be ignorant and in that sense neutral about it. The philosophical considerations on ignorance show situations in which an individual would form a belief if they ceased ignoring a proposition and started considering it. This corresponds to the idea of the "don't care" assignments, where arguments, too, can be assigned a regular label once they come under consideration.

5.2.2.c Overlapping Considerations between Forms and Norms

Interesting insights into the roles of indecision in argumentation theory can also be found when considering not only the different labels but the *different justification statuses* the labels can be synthesized to (Baroni et al., 2015; Wu et al., 2010). Wu et al. (2010) interpret arguments with justification status {`out`} to be clearly (or strongly) rejected and arguments with the justification status {`in`} to be strongly accepted. In comparison to that, they interpret arguments with justification status {`in, undec`} to be only weakly accepted and arguments with justification status {`out, undec`} only weakly rejected. Two of the six possible statuses represent (two forms of) acceptance and another two represent rejection. Finally, the remaining two justification statuses {`in, out, undec`} and {`undec`} represent indecision. In Wu et al. (2010), {`in, out, undec`} is called an *undetermined borderline case*, while {`undec`} is taken to be a d*etermined borderline case*.

The concept of justification statuses neatly bridges the overlap between normative and descriptive considerations. An argument's justification status emerges from the synthesis of various labels, depicting distinct forms of indecision. However, the term "justification status" inherently implies that the primary differentiation among these statuses lies on the normative side.

The difference between determined and undetermined borderline also resembles the distinction between "indeterminacy suspension" and "epistemic suspension," which I considered in Section 2.4 and which traces back to Ferrari and Incurvati (2022).[14] This distinction involves distinct attitudes regarding whether the considered question under discussion is answerable in general or not. As I have employed the term, epistemic suspension represents a neutral stance towards a proposition (or question) that is characterized by one's evidential situation being too deficient to determine whether p or $\neg p$. When one suspends epistemically, one still holds the belief that there is, ontologically, a correct and definite answer to

[14]One could argue that this distinction also corresponds to the two labels in the four-valued labeling. This is not surprising, as the four-valued labeling, to some extent, synthesizes a bivalent labeling, just as the justification status synthesizes a trivalent labeling.

the question of whether p is true or false.[15] In contrast, the most extreme instances of indeterminacy suspension pertain to cases of mathematical indeterminacy, where a subject can conclude that the proposition is neither true nor false but is ontologically indeterminate.

In the determined borderline case of the justification status {`undec`}, it is decided that the argument has to be labeled `undec`. In all possible labelings,[16] the argument is labeled `undec`. The only plausible label in this case is `undec`. In this sense, the label `undec` is the right position towards the argument, independent of any evidential circumstances. It seems like the constitution of the argumentative situation is such that no other label than `undec` is possible. The argument is an indeterminate case.

In comparison to this, in the case of the justification status {`in, out, undec`}, all options are still available, i.e., the argument may be labeled `in`, it may be labeled `out`, and it may be labeled `undec`. In the undetermined borderline case of the justification status {`in, out, undec`}, the different labelings disagree on how to label the argument. The different labelings can be interpreted as different reasoners (with different positions but all justified by the respective semantics) or different voters. In the situation of the justification status {`in, out, undec`}, there would be some reasoners voting for `in`, some for `out`, and some for `undec`. It is also possible to interpret this situation within one reasoning subject. The different voters would then correspond to different points of evidence that *one* subject has and that are pointing in different directions, i.e., towards `in`, `out`, or `undec`. Both interpretations take the question of what label the argument should get to be not settled, i.e., to be still an open question. The form of indecision that is presented here aligns well with what I described as epistemic suspension. The evaluation of the argument appears to depend on evidential standpoints. Epistemic suspension indicates a shortage of evidential resources necessary to adopt an alternative stance. This notion is effectively encapsulated by the justification status where all options (labels) remain open for consideration.

[15] Recall that Ferrari and Incurvati (2022) use the term somewhat differently, requiring a pessimistic stance, while in my usage the subject can be optimistic that further evidence might settle the question.

[16] This has to read as "possible according to a particular semantics," here complete semantics.

The terminology used by Wu et al. (2010) also suggests this picture. The term "undetermined borderline" signals that the the borderline status is not determined, and all options remain open, while the term "determined borderline" for the justification status {`undec`} fits the indeterminacy picture.[17] These findings are sketched in Table 5.9. As for the other tables, Table 5.9 is also meant to only give a rough overview and suggest one possible philosophical interpretation. It is essential to consider these tables along the preceding discussions for a comprehensive understanding.

Justification status	Philosophical Interpretation
{`undec`}	Indeterminacy suspension
{`in, out, undec`}	Epistemic suspension

Table 5.9: Four-valued labeling of argumentation theory and potential correspondence to different forms of doxastic neutrality from philosophy.

Evaluating the borderline justification statuses regarding their *commitment* to indecision, it can be noted that the {`undec`} status is accompanied by a certain level of commitment regarding the indecision. Indecision seems to be the correct stance.

The observation that at least some forms of neutrality come with a certain level of commitment has been largely accepted in the philosophical literature, as described in Section 2.3. Interestingly, the way the different justification statuses are described in Baroni et al. (2015) is not in line to this philosophical finding. Baroni et al. (2015) claim that the seven different justification statutes ({`in, out, undec`}, {`in, undec`}, {`out, undec`},

[17]Note that in complete semantics the justification status {`in, out, undec`} does only appear in even-length cycles of attack, whereas odd-length attacking cycles will lead to a justification status of {`undec`} for the involved arguments. This was observed by Pollock (2001, p. 242) and also revisited later, e.g., by Baroni et al. (2011, p. 21). Many scholars regard the unequal treatment of odd and even cycles in semantics like complete semantics as problematic. Therefore, other not-admissible based semantics have been developed that allow for an equal treatment of the cycles (Baroni et al., 2005). However, Horty (2023) dismisses this as a misplaced concern. He argues that while odd cycles are necessarily paradoxical, even attacking cycles are pathological but not paradoxical.

{in}, {out}, {undec}, {in, out}) are accompanied by different levels of commitment. In their semi-lattice presented in Figure 5.5 in Subsection 5.2.1, it is easy to see that an increase in commitment goes necessarily hand in hand with a decrease in undecidedness. The more indecision is "involved" in the justification statutes, the less commitment the respective justification status is assigned. Primarily, this might look reasonable, as indecision reflects some sort of openness towards the truth value of a proposition or the acceptance status of an argument. However, as I explained in Chapter 2, some forms of doxastic neutrality reflect a very committed state of mind. In particular, cases of indeterminacy suspension seem to be cases where a subject can be fully committed to their own indecision. Hence, it seems incorrect that Baroni et al. (2015) take the justification status {undec} to be *less committing* than the justification status {in, undec}. Furthermore, the status allowing all three labels, {in, out, undec}, is supposed to be even more committing than {in, undec} in the view of Baroni et al. (2015). This is implausible. If the philosophical observations on suspension are taken seriously, the justification status {undec} should either be at the same level of maximal commitment as the status {in} and {out} or should at least be more committing than any other status involving more than one label option.[18]

In summary, it can be concluded that various philosophical norms regarding suspension, the epistemological development of the debate, and different forms of doxastic neutrality all have some correspondence at the argument level of argumentation theory. Next, I intend to shift the focus to statements instead of arguments and uncover parallels with philosophical investigations at this level as well.

5.3 Indecision on the Level of Statements

5.3.1 Statement Labelings in Argumentation Theory

Thus far, I have examined argumentation theory only with respect to evaluating arguments themselves. However, one can argue that it is not

[18] Wu et al. (2010) introduce a distinct lattice for various justification statuses that spans from acceptance to rejection, without addressing commitment. In this framework, the borderline cases are both positioned in the middle between acceptance and rejection.

the arguments per se, but rather sentences or propositions that constitute the doxastic situation of an agent, forming the basis for their actions. To uncover this doxastic situation, one must delve deeper and extract the inner structure of arguments, i.e., the premises and conclusions involved. This brings us to the level of statements.

As I previously mentioned, I have already introduced a method for incorporating suspension when considering statements or propositions using the default logic framework. Therefore, I will not examine any structured argumentation frameworks where statements are treated as the fundamental elements. Nevertheless, there are also intriguing investigations stemming from abstract argumentation theory, wherein the foundational elements are abstract arguments from which (in a second step) labels for the argument conclusions (which are statements) can be inferred.

It is quite surprising that assigning justification statuses to statements has gotten only little attention, in comparison to arguments. Just recently, Baroni and Riveret (2019) systematically describe possibilities to transfer the justification of arguments to the justification of statements by evaluating the conclusions of the arguments in question.[19] In general, they distinguish two approaches, which they call "argument-focused" and "statement-focused," to determine the status of statements. In both approaches, the status of a statement is eventually determined by the statuses of the arguments speaking for or against that statement. In argument-focused approaches like ASPIC+ (Modgil and Prakken, 2014), labels of arguments are first synthesized to justification of arguments. Then, the justification of the arguments is translated to the justification of their conclusions. In statement-focused approaches, like the approach of Wu et al. (2010), labels of arguments are directly translated to their conclusions from which the justification of the conclusions can be synthesized. For the purpose of this section, the different approaches are not directly relevant. Here, I will rather focus on the different *types* of statement labelings that evolve in Baroni and Riveret (2019). They distinguish between three types of statement labelings.

[19] In fact Wu et al. (2010) consider also how to transfer their justification statuses to statements. Their approach can be subsumed under the statement-focused approaches of Baroni and Riveret (2019).

1. Bivalent Labelings: Bivalent statement labelings only allow for two possible labels a statement can obtain. Baroni and Riveret (2019, p. 839) define the possible labels to be {`yes, no`}.

2. Doubt-Tolerant Labelings: Doubt-tolerant labelings allow for a third label between the definite answers of `yes` and `no` of the bivalent labelings. This third, intermediate label expresses some sort of doubt. Baroni and Riveret (2019, p. 848) call the three labels {`yes, fal, ni`}, where `yes` means that a statement is accepted (or verified), `fal` means that a statement is falsified, and `ni` means that there is doubt about the status of the statement.

3. Ignorance-Aware Labelings: Ignorance-aware labelings further divide the group of intermediate or undetermined statements. Besides `yes` and `fal`, there are two intermediate labels: `unk` and `ni`, yielding the set {`yes, fal, unk, ni`}, where `unk` stands for unknown statements and is meant to capture statements for which there is no evidence or lack of knowledge, and `ni` captures the statements for which the evidence indicates indecision.

Transferring this distinction to the argument level of abstract argumentation, the usually used and above-discussed labelings are doubt-tolerant according to this terminology. The non-standard approach of the four-valued labeling of Definition 5.12 could count as ignorance-aware.

The differences of the three labeling types can be understood best when one looks at the following example from Baroni and Riveret (2019), which was originally introduced in Baroni et al. (2016, p. 489):
"Suppose that Dr. Smith says to you: 'Given your clinical data I conclude you are affected by disease D1'. Suppose then that another equally competent physician Dr. Jones says to you: 'Given your clinical data I conclude you are not affected by disease D1'. Your view on the justification of the statements s_1='I am affected by disease D1' and $\neg s_1$='I am not affected by disease D1' may become quite uncertain. In a different situation, at home, you use an off-the-shelf test kit suggesting you have caught disease D2. You then undertake a serious and reliable clinical test, which excludes disease D2. Would you consider the same status for the statement s_2='I am affected by disease D2' and the statement s_1? [....] Consider [as well the] statement s_3='I am affected by D3', where D3 is a poorly studied

and initially asymptomatic disease you only know by name." (Baroni and Riveret, 2019, p. 793–794)

Baroni and Riveret (2019) follow the intuition that there should be a different justification status for the statement s_1 and the statement s_2, although in both cases, there are arguments for *and* arguments against the statement. Moreover, we want to say that the status of statement s_3 should be different from the statuses of s_1 and s_2, too. s_3 should intuitively get a justification status of "full ignorance," as there is no evidence that concerns the third disease (and thereby statement s_3).
In fact, the evaluation of the different labelings by means of the example shows that only the ignorance-aware labeling can account for this intuition.[20] The different labelings provide the results displayed in Table 5.10.

	s_1	$\neg s_1$	s_2	$\neg s_2$	s_3
Bivalent labelings	`no`	`no`	`no`	`yes`	`no`
Doubt-Tolerant labelings	`ni`	`ni`	`fal`	`yes`	`ni`
Ignorance-Aware labelings	`ni`	`ni`	`fal`	`yes`	`unk`

Table 5.10: Bivalent, doubt-tolerant and ignorance-aware labelings for the example of Baroni and Riveret (2019) for statement labeling.

One can see that the bivalent labelings cannot even illustrate the different intuitive statuses of statements s_1 and s_2. Only when there is a clearly stronger argument, as in the case of $\neg s_2$, which strictly *defeats* the argument in favor of s_2, a statement ($\neg s_2$) will be labeled `yes`. In all

[20]The authors initially explore various formalisms of structured argumentation and how different types of labelings can be implemented within these formalisms. They subsequently analyze the outcomes produced by different labelings for the example provided. In some formalisms, a distinction is made between a skeptical and a credulous approach. However, this distinction only affects the treatment of statement s_1 and its negation, while remaining silent on the other statements. For the sake of simplicity, I will focus on the results of the skeptical labelings, noting that the argument in favor of the fourfold differentiation remains valid for the credulous approach as well.

other cases, all statements are rejected and labeled `no`. The doubt-tolerant labelings already do a better job in recognizing that the question about s_1 is, in comparison to s_2, not decided, and hence, taking both s_1 and its negation to be `ni`, while s_2 is labeled `fal` and its negation is labeled `yes`. However, there is no way to distinguish the justification status of s_1 (and $\neg s_1$) from the justification status of s_3. This is only achievable in the ignorance-aware labelings. Recall that ignorance-aware labelings distinguish between two forms of the middle, undetermined status. The label `ni` represents "conflicting support," while `unk` represents the "absence of support," (Baroni and Riveret, 2019, p. 848). This fits our intuitions of the example, as s_3 is a statement for which there is no support at all, while s_1 is a statement for which the support is conflicting.

In the example and within the ignorance-aware labelings, different norms of suspension and forms of doxastic neutrality can be observed. With the two different labels from ignorance-aware labelings that represent indecision, Baroni and Riveret (2019) want to distinguish cases where evidence is absent from cases where the evidence is equally balanced. This distinction is exactly reflected in the two basic norms for suspension from epistemology, the Absence Norm and the Balance Norm.
Moreover, the different forms of neutrality found in the spectrum of neutrality presented in Figure 2.2 are represented in the above example and covered (at least to some extent) by the ignorance-aware labelings. As we have seen, there is no evidence at all speaking for or against s_3, therefore, s_3 is labeled `unk`. The subject has not even considered s_3, and hence it is comparable to the example from Part 2.3.2.b, "The guava fruit has a lot of vitamin C," which A has never considered. In the range of neutrality, s_3 similarly has to be placed quite at the beginning of the axis of engagement and seems to present a case of ignorance. It is clear that we have some form of unsophisticated neutrality, as the subject is not concerned with the statement at all. In contrast, s_1 is a statement the subject already collected evidence for. One physician is arguing for s_1 and another equally competent physician is arguing against s_1. The subject has engaged with the statement in question and concluded that they cannot tell whether they have the disease. This is a form of more sophisticated neutrality, as the subject not only considered the statement s_1 but collected quite some

evidence for (and against) it. This can be compared to the example of A wondering in 2021 whether the COVID pandemic will be over in 2022 from Chapter 2. A collected evidence for and against the proposition, too (and in fact, new evidence got through to them every day), but they still could not decide, because the evidence seemed more or less balanced.

The ignorance-aware labelings allow the distinction of two different forms of doxastic neutrality. However, expanding this to represent even more than two forms of neutrality could be beneficial. Reflecting on the spectrum of neutrality illustrated in Figure 2.2, we identify at least four distinct forms of doxastic neutrality. As previously mentioned, the scenario where A remains neutral about the nutritional value of the Guava fruit parallels the indecision regarding statement s_3, which asserts that the subject has an unfamiliar disease, D3. Both cases reflect a passive and unengaged form of neutrality, indicating a lack of consideration for the respective statement or proposition. In both instances, the subject lacks cognitive contact.

In the philosophical example, the caveman's case illustrates another form, positioned even further down on the axis and thus even less engaging. Another instance in this direction is described by the situation of A being neutral about whether Selenocysteine has the EC number 808-428-7. In these most "extreme" cases of deep ignorance, the subject (A or the caveman) does not understand the respective proposition.

In argumentation theory, it can also be useful to make a more fine-grained distinction between such unsophisticated forms. Although s_3 is a statement that has not been considered yet, the reasoning subject or system at least understands the statement. However, there can be cases in which the system cannot grasp or process the statement, when, for example, the statement contains words or phrases which do not belong to the language of the respective argumentation theory system. Such cases would correspond to the caveman being neutral about whether quarks exist. In such a case, the system should be able to reply with a different form of neutrality than in the case of s_3. While s_3 is a statement for which the system is not equipped with any argument for or against, it still could in principle evaluate the statement with a decisive label if, for example, at some later point new arguments would come into play. With a statement

that is not even included in the system's language, this is different. The system here should give a reply that shows the deep ignorance, a "syntax error," and the lack of understanding about what to do with that statement.

On the other side of the spectrum, there is room for finer distinctions, too. Philosophers distinguish between rather open, but yet reflective forms of neutrality (the COVID case) from settled forms of neutrality (the proposition about God's existence). As we have seen, the example of s_1 and $\neg s_1$ represents a similar situation as the COVID case, since the subject will be undecided whether they have disease D1 but will still look for further evidence, possibly changing the label of the statement with the help of further arguments. However, there might be other examples, where the argumentation framework is built in a way in which a statement *cannot* have a different label than that of being undecided. As we have seen on the level of arguments, it might be useful to ask for a label that represents some kind of intrinsic undecidedness (which was called "epistemological undecidedness" by Baroni et al. (2015) and I called "positive indecision") for labeling statements that are intrinsically not decidable like cases of vagueness, liar-like sentences, or other self-refuting statements. To distinguish these two forms with two different labels can be helpful for a system to recognize when there is no need for further deliberation (or further arguments) concerning a certain statement, and when it is worth allowing further arguments in order to reevaluate.[21]

One possibility to provide a richer framework in both directions consists in trying to build on the non-standard ideas from Arieli (2016), Baroni et al. (2015), Jakobovits and Vermeir (1999) about partial labelings, four-valued labelings, and so-called positive indecision and transmitting these to the statement level. With the idea of positive indecision, one could, for example, describe cases of clear and settled neutrality, such as liar-like situations or the neutrality about the existence of God from Part 2.3.2.b. Conversely, with the abstaining "don't care" label, one could try to capture cases that are similar to cases of *deep* ignorance.

[21] These cases also offer a distinct normative profile. Often, what I have termed "positive justification" for indecision applies in these instances. Another way to characterize these cases is through the concept of "pessimistic suspension," outlined by Ferrari and Incurvati (2022), describing situations in which further inquiry is not expected to definitively resolve the matter in either direction.

5.3.2 Relation between Default Logic and Argumentation Theory

Recall that I put the focus of this chapter on *abstract argumentation* instead of considering structured argumentation frameworks like ASPIC+ (Modgil and Prakken, 2014) or ABA (Toni, 2014). This decision was motivated by the fact that I had already provided a detailed investigation of default logic (Horty, 2011) in Chapter 4. Default logic is a framework that is akin to these structured argumentation frameworks, as both default logic and structured argumentation frameworks operate at the level of propositions (or statements) from which default rules or arguments can be derived.

Unlike frameworks like ASPIC+, however, default logic was not designed with the purpose to serve as a model for structured argumentation. Thus, it is not embedded in abstract argumentation. This is especially evident when considering the prioritized default logic. While there has been research on integrating prioritized default logic (Horty, 2011) into abstract argumentation theory (Pardo and Straßer, 2022), these approaches rely on specific lifting principles (such as the Last-Link or the Weakest-Link Principle (Beirlaen et al., 2018)), indicating the absence of a straightforward translation method. This remains an ongoing area of research.

Nevertheless, we can set aside considerations of priorities for the purposes of this work and examine Horty's default logic without them. In this context, the logic aligns with the original default logic proposed by Reiter (1980). This unprioritized default logic can be elevated to the level of abstract argumentation. Furthermore, once a default theory is elevated to abstract argumentation theory, it becomes possible to translate the outcomes of the abstract argumentation framework back down to the domain of default logic.

Starting with a default theory Δ, I will demonstrate how one can derive the extensions of the default logic ($\mathcal{E}_i^{\mathrm{D}}$) using either the conventional approach of default logic or taking a detour through a translation to abstract argumentation. To accomplish this, I will employ methods outlined in Dung's seminal paper on abstract argumentation (Dung, 1995), as well as insights from Horty (2023) and Wu et al. (2010). The overall

process is illustrated in the following graph in Figure 5.6.

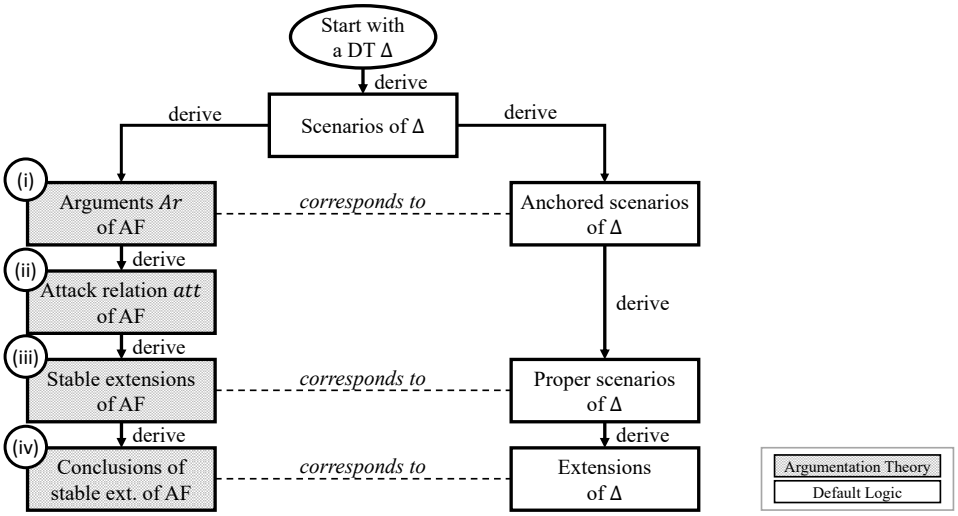


Figure 5.6: Process to derive extensions of a default theory via the canonical way in default logic (right-hand side) or via uplifting to an abstract argumentation framework (left-hand side).

The steps for default logic on the right-hand side of this graph follow the conventional approach within default theory, as described in Chapter 4. Hence, I put the focus on the detour within abstract argumentation in the following, explaining the correspondence between the two. Here, several steps are necessary.

(i) Firstly, an argumentation framework $AF = (Ar, att)$ must be established from a default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$. To achieve this, we begin by defining the set of arguments Ar as which is identified with the set of the *anchored scenarios* of the default theory Δ. A scenario $\mathcal{S}$ is anchored, according to Horty (2023), iff there exists a sequence $\langle \delta_1, \delta_2, \ldots, \delta_n \rangle$ of defaults in $\mathcal{S}$ such that for each default δ_k, it holds that

$$\mathcal{W} \cup \left\{ \bigwedge_{i=1}^{k-1} con(\delta_i) \right\} \vdash prem(\delta_k).$$

In essence, a scenario is anchored iff all its defaults are anchored by the scenario, meaning that the defaults are triggered by $\mathcal{W}$, or by

any other default that is already anchored and therefore preceding in the sequence. These anchored scenarios form the arguments of the argumentation framework and are denoted as A_i in the following.

(ii) Secondly, we define the attack relation *att* within the argumentation framework AF. An argument $A_1 \in Ar$ attacks argument $A_2 \in Ar$ if and only if

$$\mathcal{W} \cup con(A_1) \vdash \neg \bigwedge \{con(\delta) : \delta \in A_2\}$$

or

$$\mathcal{W} \cup con(A_1) \vdash \neg \bigvee \{out(\delta) : \delta \in A_2\},$$

(Horty, 2023).

(iii) With the defined argumentation framework AF, we derive the stable extensions in the usual manner, see Section 5.2. Note that a *stable extension* of an argumentation framework contains the arguments labeled `in` in the respective *stable labeling*. A stable extension $\mathcal{E}^{\mathrm{A}}$ corresponds to a proper scenario $\mathcal{S}$ of the default logic as follows:

$$\mathcal{S} = \bigcup \{A_i : A_i \in \mathcal{E}^{\mathrm{A}}\},$$

(Dung, 1995; Horty, 2023). Keep in mind that arguments are defined as anchored scenarios, i.e., sets of defaults. Consequently, an argumentation theory extension $\mathcal{E}^{\mathrm{A}}$ is defined as a set of arguments, i.e., a *set of sets of defaults*. Proper scenarios are simply *sets of defaults*. Therefore, to establish equality, we form the union of the elements in the argumentation extension $\mathcal{E}^{\mathrm{A}}$.

(iv) From the stable extensions $\mathcal{E}^{\mathrm{A}}$ in the argumentation framework, statement extensions can be derived.[22] This can be accomplished via an adapted method inspired by Wu et al. (2007). Given a stable extension $\mathcal{E}^{\mathrm{A}}$ of AF, a statement s is included in the statement extension iff there exists an argument A in $\mathcal{E}^{\mathrm{A}}$ with $s \in conc(A)$.[23]

[22] It is, of course, also possible to revert to default logic at this stage, deriving the proper scenarios and the default extensions from them.

[23] Wu et al. (2010) utilized the labeling approach in their work, but establishing correspondence to the extension approach here is straightforward. Furthermore, I had to adjust the method from Wu et al. (2010) slightly since they based their idea on structured argumentation frameworks like ASPIC+ (Modgil and Prakken, 2014), where

> Once unified with the set $\mathcal{W}$, these derived statement extensions become equivalent to the extensions $\mathcal{E}^{\mathrm{D}}$ of the default theory Δ.

I will illustrate the procedure by revisiting the example from Figure 4.2 in Chapter 4:

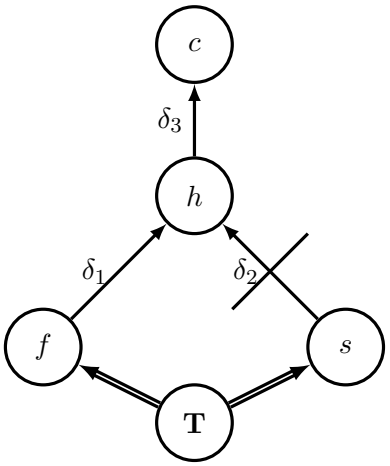


Figure 5.7: A default logic example of the food detector that was illustrated in Figure 4.2 in Chapter 4.

Here, the scenarios of this default theory Δ are $\mathcal{S}_1 = \emptyset$, $\mathcal{S}_2 = \{\delta_2\}$, $\mathcal{S}_3 = \{\delta_1\}$, $\mathcal{S}_4 = \{\delta_1, \delta_3\}$, $\mathcal{S}_5 = \{\delta_3\}$, $\mathcal{S}_6 = \{\delta_1, \delta_2, \delta_3\}$, $\mathcal{S}_7 = \{\delta_1, \delta_2\}$, and $\mathcal{S}_8 = \{\delta_2, \delta_3\}$.

With these scenarios, the arguments for the argumentation graph, i.e., the anchored scenarios of Δ, can be derived as described in (i). Scenario $\mathcal{S}_8$ is not anchored, since the premise of δ_3 is neither in W nor the conclusion of δ_2. Likewise, $\mathcal{S}_5$ is not anchored. All other scenarios are anchored. Thus, we end up with six anchored scenarios $\mathcal{S}_i$, accordingly with six arguments $A_i \in Ar$ for our argumentation framework: $A_1 = \emptyset$, $A_2 = \{\delta_2\}$, $A_3 = \{\delta_1\}$, $A_4 = \{\delta_1, \delta_3\}$, $A_6 = \{\delta_1, \delta_2, \delta_3\}$, $A_7 = \{\delta_1, \delta_2\}$.

The attack relation *att* can be derived according to step (ii). It is straightforward to confirm that each scenario that contains δ_1 attacks each scenario that contains δ_2 and vice versa. Hence, we find the following

each argument has one *unique conclusion*. However, an argument A in an argumentation framework defined *via a default theory* can have *multiple conclusions*. Arguments are defined as scenarios, and a scenario contains (possibly multiple) defaults. The conclusions of a scenario are then defined as the set of conclusions of the (multiple) defaults in the scenario.

attacks: A_1 attacks no other argument and is not attacked by any. A_2 attacks all of A_3, A_4, A_6, and A_7. A_3 attacks A_2, A_6, and A_7. Likewise, A_4 attacks A_2, A_6, and A_7. A_6 attacks itself and all of A_2, A_3, A_4, and A_7. Likewise, A_7 attacks itself and all of A_2, A_3, A_4, and A_6.

According to step (iii), the stable extensions of this argumentation graph are $\mathcal{E}_1^{\mathrm{A}} = \{A_1, A_2\}$ and $\mathcal{E}_2^{\mathrm{A}} = \{A_1, A_3, A_4\}$.

The sets of conclusions of the stable extensions (iv) are identical to the extensions of the default theory Δ of this example, which can be directly seen when deriving them in the canonical way shown on the left-hand side of Figure 5.6 and as described in Subsection 4.3.2. They are $\mathcal{E}_1^{\mathrm{D}} = \{f, s, \neg h\}$ and $\mathcal{E}_2^{\mathrm{D}} = \{f, s, h, c\}$.

Given this translation, it is apparent how my proposal to integrate suspension and ignorance into default theory via second-order attitudes in the consequences Γ can be applied in structured argumentation. Starting with an argumentation framework, stable argumentation theory extensions need to be extracted from which the conclusions can be derived. The sets of conclusions from the various stable AT-extensions can be equated with the default logic extensions, from which the consequences Γ can be derived as I outlined in Chapter 4. This correspondence between abstract argumentation frameworks and default logic yields several interesting observations concerning the picture on suspension.

First, it can be noted that the alignment between Horty's proper scenarios and *stable extensions* underscores the absence of indecision in default logic. Stable semantics prohibit indecision, presenting a pure binary perspective. This is particularly apparent in the labeling approach, where a stable labeling forbids assigning the `undec` label to any argument. When deriving statement labelings or extensions from stable semantics, this binary perspective extends to statement evaluation. Through the lens of different statement labelings, as discussed in Subsection 5.3.1, default logic offers only a binary labeling scheme. Consequently, it becomes impossible to differentiate between statements that are rejected and those that are undecided.

The alignment between default logic's proper scenarios and the strict stable semantics highlights that default logic does not offer a possibility

to incorporate indecision as present in other semantics of abstract argumentation. All other considered semantics, such as admissible, complete, preferred, or grounded, at least permit labeling arguments as `undec`. This transfers to the possibility of labeling statements `undec`, allowing for the triad of acceptance, rejection, and indecision that we find in philosophy.

This observation underscores the significance of my adaptation of the default logic framework, providing not only a means to represent indecision itself but also to distinguish between suspended and ignored propositions through second-order propositions in the consequences Γ of the default theory.

Secondly, once the translation from a default theory into an argumentation framework is made, it is intriguing to examine whether my proposal of the four different attitudes of believability ($B(p)$), certainty ($C(p)$), suspension ($S(p)$), and ignorance ($I(p)$) finds correspondence when considering more liberal semantics than the stable semantics in the argumentation framework. To compare this, it is reasonable to employ the proposal of Wu et al. (2010), which suggests not only to provide different labelings for the statements but also to synthesize the different statement labelings into justification statuses of statements, parallel to the justification statuses of arguments that I described in Subsection 5.2.1. This comparison is akin to what I do when defining the consequences Γ of a default theory in Chapter 4, where I synthesize the different default theory extensions into one set of consequences.

Interestingly, there appears to be no clear correspondence between my approach and the justification statuses of statements derived via semantics other than stable semantics. This can be observed, for instance, in the example of Figure 5.7. In my approach, the proposition h is suspended, while the proposition c, which follows from the suspended proposition h, is deemed believable (see Section 4.3.2 of the default logic chapter). I argued that this unequal treatment of the two propositions is justified because there is evidence both for and against proposition h, whereas for proposition c, we only have some (weak) evidence in its favor and no evidence against it.

When we distill the complete labelings[24] of the argumentation framework (derived from this default theory) and distill the labels for the conclusions, we find that h and c have the same labels in all the statement labelings.[25] Synthesizing to the justification status as done in Wu et al. (2010), we find that both propositions receive the same justification status: $\{\texttt{in}, \texttt{out}, \texttt{undec}\}$. Therefore, my proposal and the synthesis of complete labelings produce differing outcomes. I perceive the ability to differentiate the statuses of the two propositions h and c as an advantage of my approach.

Nonetheless, there is a high potential for further investigation into the correspondence between my proposal for suspension in default theories and indecision in argumentation frameworks derived from default theories. For instance, extracting justification statuses of statements from complete labelings as proposed by Wu et al. (2010) yields more intuitive results in handling floating conclusions compared to my approach. While my approach treats floating conclusions as accepted with certainty, Wu et al. (2010) assigns a justification status of $\{\texttt{in}, \texttt{undec}\}$ to floating conclusions, considering them only *weakly accepted*.

[24] Applying admissible labelings does not alter this situation.

[25] The reader can verify the admissible and complete labelings of this example themselves. In a labeling setting, the derivation of the labels from the arguments to the conclusions is done as follows: The labeling of the statement h is defined as the "maximum" labeling of the arguments with conclusion h (Wu et al., 2010). If there is an argument with conclusion h that is `in`, h will also get the label `in`; if there is no such `in`-labeled argument but an argument with conclusion h that is labeled `undec`, h will also get the label `undec`. Only if all arguments with conclusion h are labeled `out` is h labeled `out`.

5.4 Conclusion

In this chapter, I explored the intricacies of suspension within the realm of argumentation theory. I showed that a lot of the philosophical investigations about suspension and doxastic neutrality find some correspondence when evaluating arguments in abstract argumentation. This is intriguing because indecision can serve as a valuable tool for signaling unclear or critical situations. As elaborated in Chapter 1, when a system designed to represent artificial reasoning is equipped with an effective means of employing the tools of doxastic neutrality, the occurrence of critical and unclear situations diminishes (or at the very least, they are communicated appropriately). This can enhance trust in these systems. Argumentation theory uses indecision in different ways. However, there has been no conceptual understanding of *how* these different uses differ and whether they are even referring to the same phenomenon. Philosophical investigations can help to clarify the different uses and their connection to the respective concepts. Thereby, they can reveal the respective motivation for the use of indecision in various approaches.

While my primary focus was on abstract arguments (Section 5.2), I briefly sketched some findings in structured argumentation as well (Subsection 5.3.1) and showed how the framework of default logic (which I investigated in Chapter 4) can be incorporated into the framework of argumentation theory (Subsection 5.3.2). Regarding abstract arguments, I initially delved into the fundamental framework of abstract argumentation theory (Subsection 5.2.1). This involved introducing and explaining the standard labeling approach, various semantics, and justification statuses. Additionally, I introduced two non-standard approaches: partial and four-valued labeling. Subsequently, in Subsection 5.2.2, I uncovered philosophical analogs within these approaches.

In Part 5.2.2.a, for instance, I demonstrated that the diverse semantics of abstract argumentation yield distinct treatments of indecision. Notably, I highlighted how complete labelings more accurately reflect the philosophical picture of suspension compared to admissible labelings, which view indecision merely as a fallback option. Furthermore, grounded semantics have the most positive perspective on indecision, whereas

preferred semantics suggest that indecision should be avoided whenever possible. My investigations suggest that non-standard approaches often offer advantages over standard ones, especially in distinguishing between different forms of doxastic neutrality and between different norms. This is, for instance, the case for the four-valued labeling, which enables the integration of both the Absence Norm and the Balance Norm into the framework. Additionally, the four-valued labeling allows for the differentiation between two labels that can be interpreted to signify suspension and ignorance, a capability lacking in standard labelings. In the standard approach, only via the justification statuses are we able to identify two distinct borderline statuses, which can provide a more fine-grained picture of indecision.

Still, I found that some aspects of suspension and doxastic neutrality are not properly representable in all these approaches to argumentation theory. For example, what I called "positive indecision," a kind of neutrality that does not arise due to an evidentially deficient situation, but from the constitution of the proposition or argument itself, cannot be represented in the four-valued labeling, either. Connected to this, it is hardly possible to represent positive evidence for indecision. Moreover, I argued against the view in argumentation theory that indecision is always accompanied by non-commitment and that the more indecision there is, the less commitment there is, which was reflected in the discussion of justification statuses. In the philosophical discussion, it becomes evident that some forms of neutrality, especially suspension, do in fact inherit a high degree of commitment.

Also, on the level of statements, I found a mixed picture. While I argued that Ignorance-Aware labelings (in contrast to other available statement labelings) manage to differentiate between different norms and forms of indecision, the picture was claimed to be not fine-grained enough. An argumentation system should be capable of distinguishing cases in which it cannot deal with the input from cases in which it has no evidence for or against a statement. It should distinguish cases for which the evidential situation seems balanced, so that further arguments might be required, from cases that clearly cannot be decided. Applying the ideas from non-standard approaches in abstract argumentation theory, such as

the approach of Arieli (2016), Baroni et al. (2015), and Jakobovits and Vermeir (1999), to the statement level seems like a good starting point for this. Another challenge in the field of statement evaluation concerns the incorporation of the premises of the arguments instead of focusing only on conclusions. This is especially relevant when we want to consider different forms of defeat, such as undercutting versus rebutting defeat (Pollock, 1995).

Yet, I discovered that the opportunities to label statements in structured argumentation with an Ignorance-Aware labeling are significantly more expressive than what is feasible in default logic. The translation from default logic to abstract argumentation revealed that default logic provides no space for indecision. It only allows binary labels for statements.

In analogy to the findings in default logic, there remains the open challenge of exploring how alternative interpretations of suspension, such as those rooted in the zetetic perspective (where suspension initiates inquiry), can be integrated into argumentation theory. Given the focus of my work on the static framework, a shift towards a more dynamic perspective that encompasses the possible representation of inquiry holds promise for future investigations. To this end, investigating different processes of sequentially labeling arguments could serve as a valuable starting point.

5.4.1 Answers to the Research Questions

1. Does the considered framework allow for a way to deal with conflicting or uncertain information?

   Argumentation theory, as a framework, is already well-prepared to address conflicting information. Especially within the labeling-based approach, there is an explicit consideration of arguments that are categorized as "undecided." Thus, the framework departs from a binary assessment of only accepting or rejecting arguments and facilitates the adoption of a clear neutral stance.

2. Is there something in the light of suspension of judgment present in the framework?

   The terminology in argumentation theory primarily suggests

that the concept of indecision rather than that of suspension is present. However, one can identify elements of what we would term "suspension" within this framework. The use of indecision implies that it is a label actively ascribed to an argument. Therefore, indecision in argumentation theory already denotes a committed and settled form of neutrality that closely resembles suspension.

3. Can we find and distinguish different forms and epistemological norms of doxastic neutrality in the framework?

   Both at the level of arguments and at the level of statements, we encounter various forms of doxastic neutrality and norms governing suspension within argumentation theory. In the standard framework of abstract argumentation theory, the use of "indecision" already implies a somewhat engaging, committed interpretation. Additionally, non-standard approaches allow for a finer distinction and the identification of different forms of neutrality. Here, a distinction can be made between indecision in argumentation theory, which more closely resembles ignorance, and a form of indecision, which aligns with what I typically described as suspension. Various degrees of committed and sophisticated neutrality can also be found in ignorance-aware labeling at the statement level. The differentiation between different forms of suspension, such as epistemic vs. indeterminacy suspension, can be found at the argument level when synthesizing the different labelings into different justification statuses. From the perspective of argumentation theory (at least as presented by Baroni et al. (2015)), indecision is always associated with a lack of commitment. This contrasts with the philosophical viewpoint, where qualified indecision in the form of suspension is considered highly committed.

   Within the normative perspective, we encounter both the Absence Norm and the Balance Norm within the non-standard approach of a four-valued labeling. Both represent cases of privative justification. As also pointed out by Baroni et al. (2015), there is no representation for positive justification within argumentation theory. Similarly, at the statement level, we find the Absence Norm and the Balance Norm

but no representation for positive justification.

In the standard approach of abstract argumentation theory, it is challenging to identify distinct norms for suspension. In a way, it becomes difficult to clarify what an Absence Norm could even consist in, as abstract argumentation solely models the relationship of attack, while lacking a representation for the concept of support. Consequently, if an argument is presented without an attack (which might be interpreted as an absence of evidence), it is accepted rather than considered undecided.

Chapter 6

Machine Learning

Contents

6.1 Introduction

This chapter investigates neutral behavior in machine learning (ML). In particular, I investigate so-called *Abstaining Machine Learning* (AML) systems (Campagner et al., 2019), sometimes also referred to as *ML with a reject option* (Hendrickx et al., 2021), and draw parallels to the philosophical use of suspension of judgment. While in philosophy, I mostly employed the term "suspension," in the context of machine learning, I will refer to the neutral behavior with the term "abstention" following the standard terminology within this field.

For exploring the parallels between suspension in philosophy and abstention in machine learning, it is beneficial to view both as neutral behaviors towards certain questions as characterized in Subsection 2.3.1. This unification allows for explaining important similarities and differences between suspension and abstention. We consider questions like: "Which dog breed is displayed in this image", "Is this tumor malignant or benign?" or "Is this person creditworthy?", which have a finite set of well-defined, full answers A. This set consists of all the *defined* possible answers to the question. For Q_1 = "Is this tumor malignant or benign?", A_1 = {*malignant, benign*}. For the question Q_2 = "Which dog breed is displayed in the image?", possibly A_2 = {*Husky, Labrador, Dachshund, Retriver*}. And for propositional questions like "Is this person creditworthy?" the set can simply be {*yes, no*}. In the context of Machine Learning, the answers are typically identified with *outputs*. To indicate the use of a term as an output, I will employ a typewriter font, i.e., `malignant` and `Labrador`, and so on.

In this work, I focus on those situations in which *none of the answers* from the answer set is selected. Instead, the question is addressed with a response that expresses neutrality, uncertainty, or indecision about the correct answer.

As elaborated on in Chapter 2, suspension of judgment in philosophy is usually characterized as a doxastic, mental stance whose counterparts are belief and disbelief. While belief and disbelief express those doxastic positions that are accompanied by some certainty or decisiveness about a question Q and its correct answer, suspension expresses neutrality and

indecision about Q.[1]

In machine learning, neutral outputs are described with the term "abstention." Traditionally, for an ML algorithm tasked with answering a question Q of the above type, the set of possible outputs is equal to the set of the defined answers A.[2] For the question about the dog breed, the algorithm could output `Husky`, and for the question about the tumor, the algorithm could output `benign`. Abstaining machine learning algorithms are additionally able to output an `abstention` response, which is not a member of the defined answers in A.

In this chapter, I bring the two fields and the respective debates together. In doing so, the chapter starts to fill a gap in the philosophy of AI literature. Philosophy of AI is concerned with describing and evaluating AI systems with the help of philosophical terms, norms, and debates. So far, this has not been done for abstaining machine learning, although this area provides an enormous potential for philosophical investigations.

Abstaining ML is a field in ML research that is still considered only by a small group of researchers (Campagner et al., 2019; Ferri and Hernández-Orallo, 2004) and largely unknown to philosophers. This is surprising, considering that AML systems show a promising way to uncover and deal with uncertainties in decision processes. As argued by Phillips et al. (2020), the awareness of its own knowledge limits is one key principle of an explainable artificial intelligence. Abstaining Machine Learning provides a direct method for explicitly defining these knowledge limits and communicating them to users. As I outlined in the Introduction (Chapter 1), this thesis aims to contribute to the fields of trust and explainability in AI systems by underlining the significance of uncovering and effectively communicating uncertainties and the limits of knowledge. To achieve this, I intend to enhance the awareness and comprehension of abstaining machine learning among both AI researchers and philosophers.

The way in which the chapter aims to bring the two fields together is as follows: In Section 6.2, the chapter first addresses the task of explaining the idea of AML, giving an overview of the different kinds of AML

[1] The analogy between belief and disbelief can be drawn best for propositional questions that have only "Yes" and "No" in their answer set.

[2] At least this is so for a classification problem, which I will concentrate on.

systems, and clustering the different algorithms into classes based on two dimensions. One dimension describes different reasons for abstention, i.e., different situations in which an abstaining output is issued (Subsection 6.2.2). The second dimension describes different ways in which abstention is (conceptually and technically) implemented in the system (Subsection 6.2.3).

In the second part of the chapter, the philosophical analysis takes place. In a sense, this philosophical analysis is analogous to the investigations conducted for the other two frameworks. It involves demonstrating how certain types of AML systems meet the criteria for suspending judgment. First, I will draw comparisons between the reasons for abstention detailed in Part 6.3.1.a and the various reasons (or norms) for suspension that I examined in Section 2.2. Secondly, in Part 6.3.1.b, I will compare the methods of implementing abstention to the nature and the forms of suspension explored in philosophy (as presented in Section 2.3), addressing the question of which types of AML systems possibly correspond to suspension.

Additionally, this chapter seeks to explore the broader topics within the philosophy of artificial intelligence that have not been previously applied to this specific category of machine learning systems. As my focus in this chapter is on AML systems, which I have identified as a potential type of ML system capable of suspension, I will expand specific questions in the philosophy of AI to this kind of system. In particular, I will delve into matters concerning the autonomy and explainability of machine learning-generated responses. I will apply these two questions to the abstaining output of ML systems and discuss how autonomous (Subsection 6.3.2) and how explainable (Subsection 6.3.3) the abstaining output is or can be. I will argue that the different types of abstaining systems presented in Section 6.2 offer different answers for these two questions.

6.2 Abstaining Machine Learning

In this chapter, I consider *predicting* ML systems. In general, the task of those kinds of ML systems is to select a defined answer from an answer set A for a question Q. The examples considered here refer to cases where the answer set A is a finite, discrete set. A familiar example is that of

an image classifier. If an image classifier is to identify the breed of dog depicted in an image, the system is asked the question Q_2 = "Which breed of dog is displayed in the image?", and a possible set of defined answers is $A_2 = \{\texttt{Husky}, \texttt{Labrador}, \texttt{Dachshund}, \texttt{Retriver}\}$.

This type of ML is often referred to as *predicting* ML and is distinct, for example, from ML in robotics, where physically acting systems are in focus, and from generative AI, where the task of the AI is to generate text, images, or other data. Moreover, the predicting systems considered here differ from other predicting systems that have a continuous, i.e., infinite, set of possible answers available.[3] What is considered here is often referred to as a *classifier*.

Moreover, I only consider so-called *supervised* ML algorithms. This characteristic concerns the way the system is trained. In ML, one generally distinguishes between an application phase, in which the system solves the task that it is supposed to solve, e.g., answering a question, and an earlier training or learning phase, in which the system learns how to solve the task. In the training phase, the system is equipped with some kind of training data. Supervised systems learn to establish a relationship between the input and the desired output through *labeled* training data. The image classifier, for example, is supposed to establish a relationship between some kind of image-input (which can be identified with the question) with some kind of breed-output (which can be identified with the answer). This means that in the training phase, the algorithm is provided, for example, with many images of huskies and is told to answer to these images with the label `Husky`, many images of retrievers and told to answer to these images with the label `Retriever`, and so on. By this, the algorithm learns in the training phase to connect certain features of the input (image) with a certain output (breed label). In other words, the training data consists of many sample questions "Which dog breed is displayed in image $\boldsymbol{x}$?" (inputs) and corresponding correct and desired answers (outputs) like `Husky`. For the question Q_1, whether a certain tumor is malignant or benign, an input data point will not consist of a whole image but of a list of measured features of the tumor, e.g., its size, the number of concave points, its perimeter, and so on. The output will be the answer, i.e., either

[3] Most of the literature on AML deals with discrete classifiers. There are some studies on abstention in regression models (Asif et al., 2020), but I will not consider these here.

`benign` or `malignant`. In Subsection 6.2.1, I will illustrate how training data for question Q_1 could be visualized and provide an explanation of the mathematical properties of the training data points.

When the system has learned in the training phase to connect certain questions (or certain images or lists of features) with certain correct answers (or certain labels or classes[4]), it can later apply this knowledge in the application phase by answering new, previously unanswered questions, i.e., new, unseen images or new, unseen tumors.

What distinguishes abstaining classifiers from conventional classifiers is the option to choose none of the defined answers of the answer set A as an output. AML can issue an *abstaining output* as a response to the question Q allowing an alternative to the defined answers. Therefore, AML systems are often referred to as possibly *rejecting* the task or refusing to give an answer. This rejection may be issued in the form of an output saying `I do not know`, `I abstain`, `I reject a prediction`, etc. For example, the dog breed classifier can produce the output `I don't know` (Thulasidasan et al., 2019b) for a given image in addition to the defined answers: `Husky`, `Labrador`, `Dachshund`, `Retriever`.

This seems to be appropriate in many application domains. Most prominently, researchers have argued that in high-stakes scenarios like medical decision-making, ML systems with an abstaining option are clearly preferable as diagnostic tools (for example for cancer, COVID-19, or liver disease detection) (Kompa et al., 2021; Brinati et al., 2020; Hamid et al., 2017; Kempt and Nagel, 2022). But also, in other application areas like weather and climate diagnostics (Barnes and Barnes, 2021) or simple spam filters (Artelt et al., 2022), the abstaining option is often considered desirable. If ML systems are to serve as expert or advice systems, it is recommended that these systems liberally admit their own uncertainty in critical situations instead of making a decision at any cost. This also

[4]The responses generated by a machine learning system are usually called "outputs." Moreover, the terms "label" and "class" are commonly employed in literature, particularly within the context of classifiers. These terms — output, class, and label — are frequently used interchangeably. Strictly speaking, the output usually signifies the classifier's result, while the label typically refers to the ground-truth label in the training dataset. Both outputs and labels usually take up the same possible values, the values of the distinct classes. One could consider classes as abstract categories into which the data points fall. A label and an output indicate membership within one of these classes.

corresponds to our expected behavior of human experts, as Ferri and Hernández-Orallo (2004, p. 1) point out: "When we use human assistance for supporting decision making, there are some cases where the expert says 'I don't know' and asks for further assistance (to other experts) or just prefers to postpone the decision. Frequently, we say a person is an expert or a wise person when she prefers to be silent (and ask other experts) rather than to make a mistake." Moreover, as Campagner et al. (2019, p. 292) point out, when abstaining ML systems alert us to uncertainties, this often gives us the opportunity to improve the basis for decision-making: " [...] because it could be used in a human in the loop setting, to point out to the human decision-maker which instances might require the acquisition of further or more precise information."

In the following, I will illustrate the domain of AML classifiers using two dimensions. Along the first dimension, I distinguish the different reasons for abstention. Thus, I give an overview of situations in which abstaining ML is in play. For this purpose, I distinguish between *outlier abstention* and *ambiguity abstention*. The second dimension describes the composition of the algorithms. Here, I basically distinguish two ways in which the abstention option can be technically and conceptually integrated into an ML algorithm. I call these two types of AML systems *attached* and *merged abstention*. The two dimensions are fundamentally independent. One dimension concerns the reasons for abstention, and the other dimension concerns the implementation of abstention. In principle, therefore, any combination of outlier or ambiguity abstention with attached or merged abstention is possible.

In presenting the AML systems and their distinctions along the two mentioned dimensions, I will employ the two example questions Q_1 and Q_2. Specifically, I will revisit the question Q_1 concerning cancer detection and furnish an example with real-world parameters and training data points. I will explicate the training data using both a graphical representation and a mathematical framework. The mathematical terminology will be used not only to describe the training data but also to describe an optimization problem employed to identify an optimal classifier. By doing so, I can explain the process of training a standard classifying ML algorithm, using the cancer detection example of Q_1 as a guide.

6.2.1 An ML Example for Cancer Detection

The Training Data

The data points presented in the following example can serve as training data for a supervised ML algorithm aiming to classify breast tumors as either benign or malignant. The ML algorithm would thus be a *binary classifier* since only two possible class-labels are available. Consequently, the trained ML system is supposed to answer the question Q_1 = "Is this tumor malignant or benign?" with one of the defined answers from the set $A_1 = \{\texttt{malignant}, \texttt{benign}\}$.

A data set for benign and malignant points that is often used can be found in (Wolberg et al., 1992). This data set comprises multiple features, i.e., input variables, from which I have selected two (the smallest nucleus perimeter and the proportion of concave points) to visualize a two-dimensional input space. In Figure 6.1, an extract of these training data points is sketched.[5]

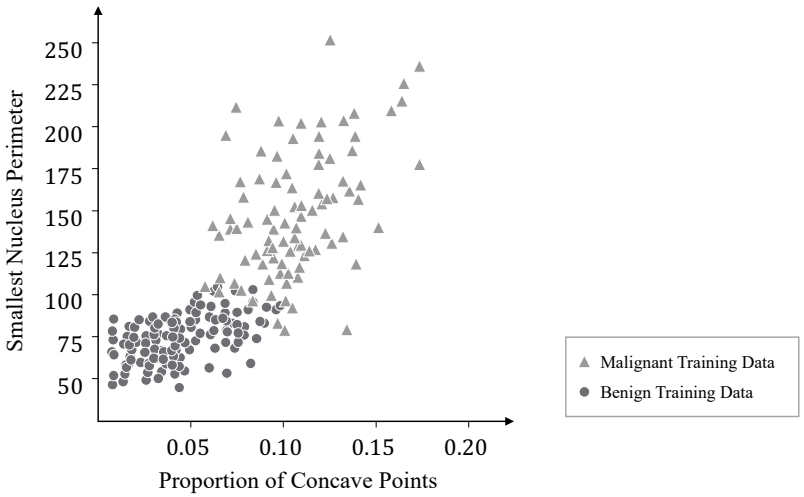


Figure 6.1: Training Data for Cancer Detection: Malignant data points are represented by triangles; benign data points by circles.

[5]The values of the visualized data points are not extracted from the data set. Rather, for this particular case study, the rough distribution of the malignant and benign data points in the data set is only sketched in order to obtain a better visualization. The range in which the data points occur is still correct.

Figure 6.1 illustrates possible training data points for training an algorithm to answer Q_1. The training data points are illustrated by the circles and the triangles in the two-dimensional coordinate system in the figure. Mathematically, each training data point can be described by a tuple $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$, $i = 1, \ldots, n$.

In this tuple, $\boldsymbol{x}^{(i)}$ is a two-dimensional vector, which represents two input parameters: the smallest nucleus perimeter and the proportion of the concave points. For example, it could be $\boldsymbol{x}^{(i)} = (0.17, 152)$ with 0.17 being the proportion of the concave points (ranging from 0 to 1) and 152 the value for the smallest nucleus perimeter (in micrometers). As each of the two entries of $\boldsymbol{x}^{(i)}$ is real-valued, $\boldsymbol{x}^{(i)}$ is an element of the two-dimensional real space, i.e., $\boldsymbol{x}^{(i)} \in \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$. In Figure 6.1, the value $\boldsymbol{x}^{(i)}$ is represented by the *position* of the circle (or triangle) in the coordinate system, i.e., by *where* the circle (or triangle) lies with respect to the horizontal and vertical axis. The space that contains all the training data points is called *input space*, which is in general denoted by X. For our example, it is $X = \mathbb{R}^2$.[6]

Since we consider supervised ML, a training data point, $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$, however, consists not only of the input values but also of the respective (ground-truth) label. In the breast cancer example, we not only know for a specific training data point its smallest nucleus perimeter and its proportion of concave points, but we also know whether that training data point *is in fact* a malignant or a benign one. This information is stored in $y^{(i)}$. In our example case, $y^{(i)}$ can have one of the values: `malignant` or `benign`. In Figure 6.1, the value of $y^{(i)}$ is represented by the *shape* drawn in the graph. If $y^{(i)} =$ `malignant`, the point is represented by a triangle, if $y^{(i)} =$ `benign`, the point is represented by a circle. The set of the potential labels is also called the *output set*, as the task of the ML system becomes to predict these labels. It is in general denoted by Y. For our example, it is $Y = \{\texttt{malignant}, \texttt{benign}\}$, which is identical to the set A_1, the set of possible answers to Q_1.

[6] In fact, it makes sense to restrict the space of X to a subset of $\mathbb{R}^2$ for this example. The proportion of concave points is measured in a value between 0 and 1, suggesting the interval $[0, 1] \subseteq \mathbb{R}$ and the smallest nucleus perimeter is measured in micrometers suggesting to at least restrict the input space to the space of all positive-valued reals $\mathbb{R}^+ \subseteq \mathbb{R}$. A medical reasonable subspace would be even smaller, as the nucleus perimeter can certainly not become arbitrarily large.

In total, one example of a training data point $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$ with $\boldsymbol{x}^{(i)} \in X$ and $y^{(i)} \in Y$ is always an element of the Cartesian product of the input and the output set, i.e., $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle \in X \times Y$. For our breast cancer example, one concrete training data point could be $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle = \langle (0.17, 152), \texttt{malignant} \rangle \in \mathbb{R}^2 \times \{\texttt{malignant}, \texttt{benign}\}$. The complete training data set is denoted by T, i.e., $T = \{\langle \boldsymbol{x}^{(1)}, y^{(1)} \rangle, \langle \boldsymbol{x}^{(2)}, y^{(2)} \rangle, \ldots, \langle \boldsymbol{x}^{(n)}, y^{(n)} \rangle\} \subseteq X \times Y$.

The Training Phase

As for every supervised ML classifier, the goal is to build a classifier that tells you for any arbitrary input (any vector $\boldsymbol{x} \in X$), representing a *new, unseen tumor*, whether that input is benign or malignant. For this, a training phase is necessary where a connection between certain input values and the different output classes can be established, based on the given training data.

For example, it might be determined that a proportion of concave points above 0.15 occurs only in malignant cases.[7] This means that the algorithm tries to find a *decision boundary*[8] between the different training data points that separates the data points that belong to the malignant class from the data points that belong to the benign class. An example of such a boundary can be visualized by a line in the input space, separating malignant and benign training data points.

Mathematically, the separation of the data points (in the input space) can be represented by a function f which maps *any* input vector $\boldsymbol{x} \in X$ to an output $y \in Y$. According to the above definitions, X is called the input space (or set) and Y is the output set of the function f. How can we find such a function? We can start by considering those functions $f : X \to Y$ that use the simplest decision boundary, i.e., a line, as we will see in Figures 6.8 and 6.9. This means, we consider a linear model.[9] Overall, the possible

[7] Commonly, these rules found by the algorithm are not that simple and are not even expressible in a way that the user or programmer would understand. Rather, they are encoded, e.g., via the enormous number of parameters of a deep neural network.

[8] If we have a multi-class problem, one boundary will not be enough.

[9] Considering only linear models is one possible *model choice*. Instead, one could also make a different model choice, like a quadratic or logarithmic model, returning curved decision boundaries. In principle, though, the set of possible functions is always restricted by a particular choice of a model, e.g., to avoid overfitting or too much computational complexity, see Murphy (2022).

candidate functions of a particular model choice can be collected in a set $\mathcal{F}$. The goal is then to choose one function, to be denoted $\hat{f}$, in $\mathcal{F}$ that has the property of performing the mapping of the input parameters of *the training data* in the best possible way. This means that the task in our binary classification problem is to find a $\hat{f}$ for which $\hat{f}(\boldsymbol{x}^{(i)}) = y^{(i)}$ for as many $i = 1, \ldots, n$ as possible.

But how can we determine $\hat{f}$ and derive a boundary that separates the training data labeled `malignant` from the training data labeled `benign` best? One option would be *to try different* functions in $\mathcal{F}$ and choose the one that makes the fewest mistakes (trial and error).

The different functions in $\mathcal{F}$ then have to be evaluated in order to find the "best one," i.e., the one that maps the most $\boldsymbol{x}^{(i)}$ $(i = 1, \ldots, n)$ to their associated $y^{(i)}$.[10] We do this be determining for each f in $\mathcal{F}$ how "bad" it is, i.e., *how much loss* it produces for the different training data points. For this, we introduce a *loss function* l which determines how much loss a particular function f generates for each training data point. This loss occurs when a data point is assigned a different label, according to the decision boundary set by f, compared to its ground-truth label from the training data. For example, the training data point is labeled `benign`, and the label assigned by the algorithm (according to that boundary) is `malignant` (or vice versa).

In general, the loss function is the heart of a learning algorithm. It determines the loss a candidate function $f \in \mathcal{F}$ generates. The total loss (also often referred to as "cost") is usually determined by summing up the single losses that occur when evaluating a training data point by the candidate function f.

[10] In reality, for most applications, the optimal function has not only the objective to map as closely to $y^{(i)}$ as possible, but also to be "simple enough" to avoid the problem of overfitting. Therefore, the objective usually consists of one part that is to reduce the prediction error and a second part that *regularizes* f, i.e., avoids that f perfectly fits the training data by being overly complex. With this second part, one wants to ensure that the function not only maps the specific training data points well, but can also reasonably well *generalize* beyond the training data. For more information about this regularization see, for example Murphy (2022). For reasons of simplicity, I will only consider the first objective of mapping the training data as good as possible here. Moreover, as I limited the model choice to linear models, regularization is not relevant after all, as the model's complexity is restricted to linear functions.

A simple loss function could in general look like this: $l : Y \times Y \to \{0, 1\}$,

$$l(y^{(i)}, f(\boldsymbol{x}^{(i)})) = \begin{cases} 1 & \text{if} \quad y^{(i)} \neq f(\boldsymbol{x}^{(i)}), \\ 0 & \text{if} \quad y^{(i)} = f(\boldsymbol{x}^{(i)}). \end{cases} \tag{6.1}$$

Given a particular candidate function f, the loss function l for one training data point is 0 if the ground-truth label *is equal* to the label determined by f and is 1 if the ground-truth label *is unequal* to the label determined by f.[11]

The optimal function $\hat{f}$ is then $f \in \mathcal{F}$ for which *the sum* of the values of the loss function over *all* training data points $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$ $(i = 1, \ldots, n)$ is *as small as possible*.[12] Mathematically, we find this $\hat{f}$ by solving the following optimization problem:

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y^{(i)}, f(\boldsymbol{x}^{(i)})).$$

In the following, I will call $\hat{f}$ sometimes also the *regular predictor*, to allow for a distinction from other predictors that are obtained in an abstaining setting.

The Application Phase

Once we have found $\hat{f}$ in this way, we thereby found a model and a separation boundary, and we can *apply* the ML model. The application phase can be represented in the following way: We take a new input vector $\boldsymbol{x}$ from X, which the system has not seen before, and put it through the ML system, i.e., the regular predictor $\hat{f}$. The output $\hat{f}(\boldsymbol{x})$ then indicates the assigned label for the input $\boldsymbol{x}$. This application phase is visualized in Figure 6.3.

[11] In accordance with Footnote 4, this suggests a terminology in which we distinguish the "ground-truth label" from the "output label," the latter being the label determined by f.

[12] The solution of such an optimization problem is often not guaranteed to be unique.

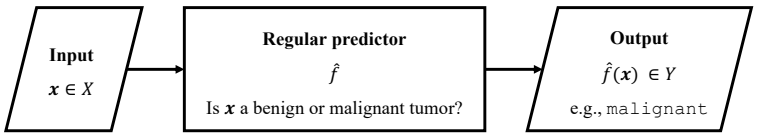


Figure 6.2: Flowchart of the application phase of a regular (non-abstaining) ML classifier: The input $\boldsymbol{x}$ is processed through the regular (non-abstaining) predicting function $\hat{f}$ and an output $\hat{f}(x)$ from the output set Y is generated.

The real-world example for question Q_1 will be revisited and applied in the next two subsections when introducing the domain of AML classifiers, highlighting the two differentiations between ambiguity vs. outlier abstention (Subsection 6.2.2) and attached vs. merged abstention (Subsection 6.2.3).

6.2.2 Reasons for Abstention: Ambiguity versus Outlier Abstention

The first distinction in abstaining machine learning revolves around the *reasons* prompting a system to abstain. This distinction describes the handling of a *new* data point during the *application phase* of an AML algorithm. Therefore, the following elaborations have to be considered at a stage where the system is already trained and is applied to new data points.

In general, if it is too uncertain whether the system will produce the correct output for the new data point, an AML system will abstain. This uncertainty can arise in many ways. While some uncertainties concern the general structure of the model (e.g., an inappropriate model choice for the kind of training data), other uncertainties are due to some characteristic of a specific input.

The different uncertainties can be categorized by means of a common distinction in abstaining machine learning: the distinction between ambiguity and outlier abstention. Roughly speaking, when an input is too far away from or too dissimilar to the training data, we are dealing with an outlier; when the input is such that more than one output is likely for the input, we are dealing with ambiguity. This distinction can be found in early works (Dubuisson and Masson, 1993; Denoeux, 1995) and is sometimes referred to with different names, such as novelty rejection versus ambiguity rejection (Hendrickx et al., 2021), distance rejection versus

ambiguity rejection (Dubuisson and Masson, 1993) or distance rejection versus confusion rejection (Mouchère and Anquetil, 2006a).

Outlier Abstention

In outlier abstention (Lotte et al., 2008; Mouchère and Anquetil, 2006b,a), the system abstains on data points that are very dissimilar to the training data. This is useful for (at least) two scenarios. First, if an input is very far away from *all* training data points, it is likely that the input might belong to none of the classes that are in the scope of the classifier. If a classifier is trained to classify different breeds of dogs and the new input is an image of a cat, the cat image will likely be very dissimilar to *all* of the different dog images that were used for training the classifier. The classifier here really should abstain, as it is only capable of classifying dogs and will not be able to solve the task of classifying a cat. The correct answer for this input of a cat image (and for the question about what is displayed in the image) is not included in the set of defined answers $A_2 = \{\texttt{Husky}, \texttt{Labrador}, \texttt{Dachshund}, \texttt{Retriver}\}$ that the system operates on. Hence, it is reasonable that the algorithm chooses none and abstains.

Secondly, even in cases where the correct label of an input might be one of the considered labels of the classifier, i.e., the correct answer to the question is one of the defined ones, outliers appear. If an input dog image is very dissimilar to the training images, this suggests that any prediction the system could make will be prone to error. The data point can be dissimilar to the training data for various reasons: There could be measurement inaccuracies, there could be adversarial examples (that are meant to trick the system), or the training data have been just not diverse enough (Hendrickx et al., 2021). In this sense, outlier detection is often used to actually improve the prediction system. If a certain dog image is characterized as an outlier (although the system should recognize the type of dog in the image), this might suggest that the system was trained on too uniform and not sufficiently diverse data, which could be improved based on the detected outliers. Maybe the system was trained on images of dogs that were taken during summertime and the detected outlier is a dog image in the snow. Detecting this outlier can suggest retraining the system with more diverse data; in this case: images taken in different seasons.

Figure 6.3 illustrates a typical case of outlier abstention. Similar to Figure 6.1, the triangles represent the training data with the label `malignant` and the circles represent the training data with the label `benign`. Besides the training data, an additional data point is represented by a star. The star represents a to-be-classified new data point that is taken to be an outlier.

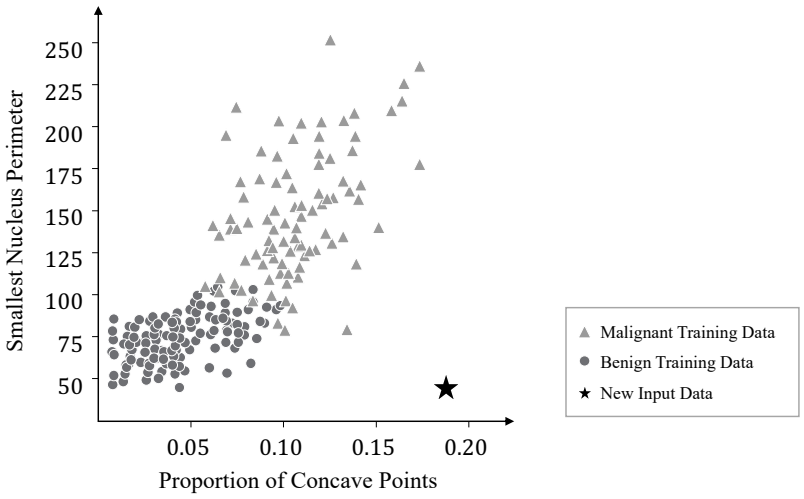


Figure 6.3: Outlier Abstention: A to-be-classified data point (star) is too dissimilar to training data (circles and triangles).

For the example training data presented in Figure 6.3, we can see that all training data points have a proportion of concave points between 0 and 0.18 and a smallest nucleus perimeter between 45 and 250, i.e., $x_1^{(i)} \in (0, 0.18)$ and $x_2^{(i)} \in (48, 250)$ for all $i \in \{1, \ldots, n\}$.

A new data point $\boldsymbol{x} \in X$ could now, for example, be considered an outlier, if $x_1 \notin (0, 0.18)$ *or* if $x_2 \notin (48, 250)$. Mathematically, there are plenty of other methods and metrics for determining the amount of dissimilarity (mathematically, the distance)[13] of a new data point to the training data. For example, the metric of Euclidean distance either between the new point and the closest training data point, between the new point and the training data sphere, or between the new point and the center of the training data could be computed. One could set a threshold t for a maximum distance

[13]Distance measures like the classical Euclidean distance are used mathematically to determine dissimilarity between data points.

between a new data point and the training data points and if the distance is above this threshold, the new data point is considered an outlier.

Of course, it depends on the context how dissimilar (how "distant") a new data point has to be from the training data in order to be considered an outlier, i.e., how large the threshold t for the maximum distance should be. In general, one plausible requirement for choosing t could be that it should be high enough such that the distance among the different training data points is on average smaller than t.

Ambiguity Abstention

In contrast to outlier abstention, the problem in ambiguity abstention is not that none of the answers seem likely, but rather that too many of the answers seem likely for the input (Barnes and Barnes, 2021; Campagner et al., 2019; Sarker et al., 2020; Thulasidasan et al., 2019b). Ambiguity is at play when an input appears to belong to more than one class. This can be the case when the input is on a boundary, but also can be due to the structure of the training data itself.[14] Often training data is not perfectly separable. When this is the case, the training data is called *noisy*. This means that there are certain regions in the training data that overlap (see Figure 6.4). If an input sample lies in such an overlapping (or noisy) region, ambiguity is present and a prediction for one class or the other would be error-prone. This type of uncertainty can also arise for a variety of reasons. Maybe the input data point simply has certain characteristics of one class as well as characteristics of another class. For example, the size of the dog in an input image might be indicative of a retriever, while the coat color is clearly indicative of a Labrador.

A case for ambiguity abstention for the Example from Subsection 6.2.1 can be visualized like this:

[14] In the latter case, the uncertainty is not purely due to some characteristic of the input sample but also due to the composition of the training data being not perfectly separable or the model choice being inappropriate to perfectly separate the data.

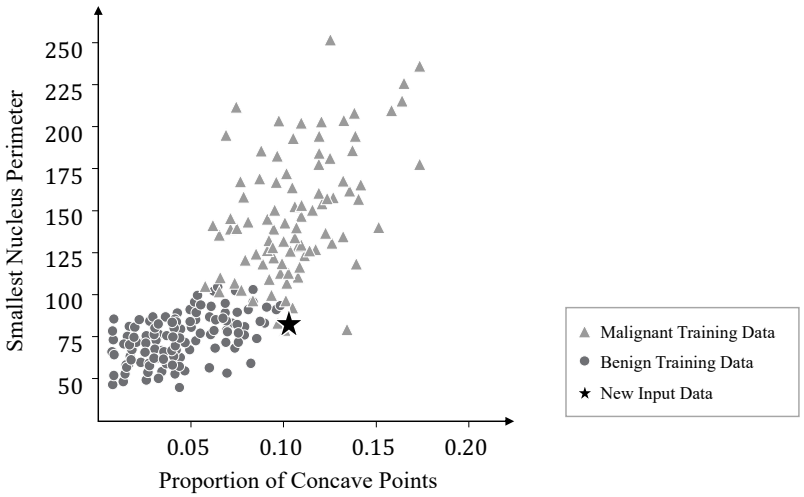


Figure 6.4: Ambiguity Abstention: A to-be-classified data point (star) lies in an overlapping, ambiguous area of the training data.

One important distinction between outlier and ambiguity abstention lies in how a data point can be identified as an outlier or an ambiguous point. Detecting an outlier typically does not require any information about the *labels* of the training data points. As illustrated in Figure 6.3, the outlier could be identified without distinguishing between the triangle-shaped and circle-shaped training data points. The only essential information is the input values of the training data points (i.e., *where* they are located in the two-dimensional space) and the input value of the new data point. The labels $y^{(i)}$ of the training data are not needed.

In contrast, to identify a new data point as an ambiguous case, information about the labels of the training data is essential (i.e., the information $y^{(i)}$ is necessary). Furthermore, determining whether a new data point $\boldsymbol{x} \in X$ is an ambiguous case often depends on the specific trained model and cannot be directly inferred from the training data and $\boldsymbol{x}$ alone. While the potentially ambiguous region is visually discernible in Figure 6.4, this is not always the case, especially not for higher-dimensional data and more complex models. This consideration is picked up again in the distinction between two forms of attached abstention, as discussed in Section 6.2.3.

6.2.3 Implementation of Abstention: Attached versus Merged Abstention

In this section, I introduce the second dimension for classifying AML systems. Here, I distinguish different *types* of AML systems with respect to the technical implementation of the abstention option. Although there are many ways to incorporate the abstention option into a classifier, I will present two main categories under which many systems can be subsumed and that I consider to be fundamentally different approaches. In contrast to many other reasonable approaches to categorizing different abstaining models (see especially Hendrickx et al. (2021)), my distinction between attached and merged abstention models is chosen for being most relevant and useful for the philosophical questions considered in Section 6.3. In Section 6.3, we will see that the different types of abstaining models behave differently regarding the questions about their similarity to suspension, their autonomy, and their explainability.

Attached Abstention

The first class I will consider is the class of what I will call attached abstaining machine learning systems. In these systems, the part that is relevant for the abstaining activity is in some sense *attached* to the core machine learning algorithm, i.e., to the predicting algorithm (Sarker et al., 2020; Mouchère and Anquetil, 2006a). Hence, the predicting and the abstaining activities are separated from each other and one can speak about "the predictor" (which I refer to as $\hat{f}$) and "the rejector," r (i.e., the part of the system that is relevant for abstaining). There are two ways in which the rejector can be attached to the predictor. The rejector can be attached *prior* or *posterior* to the predictor.

(a) *Pre-algorithmic attachment*
In pre-algorithm abstention models, the abstaining part is executed prior to the predicting classifier[15] (Wu et al., 2007; Mouchère and Anquetil, 2006b; Homenda et al., 2014; Coenen et al., 2020). This means that that for a given input, the rejector decides whether or not to abstain for the input even *before* the prediction algorithm starts.

[15]What Hendrickx et al. (2021) call a "separated rejector" can best be compared to pre-algorithm abstention models.

If the input is not rejected, the predictor starts running; if the input is rejected, the predictor will not even be started in the first place. This has the clear advantage of low computational effort since the additional model, the predictor, is started only for those inputs for which the prediction is considered certain enough.

Pre-algorithmic abstention is especially relevant for outlier abstention (Coenen et al., 2020; Lotte et al., 2008). For a given input, the decision of whether the prediction will be too uncertain is made before the prediction is computed. Therefore, it must be a property that is inherent to the input data that determines whether the input will be rejected. This does not work well for ambiguity rejection because ambiguity arises not only due to the input but due to the relationship of the input and the trained model. In order to detect ambiguity, the model generally has to be run. In contrast, for outlier rejection, it is possible to determine whether an input data point is an outlier (i.e., very dissimilar to the training data) before calculating a prediction for that data point. The concept of pre-algorithmic attachment is visualized in Figure 6.5.

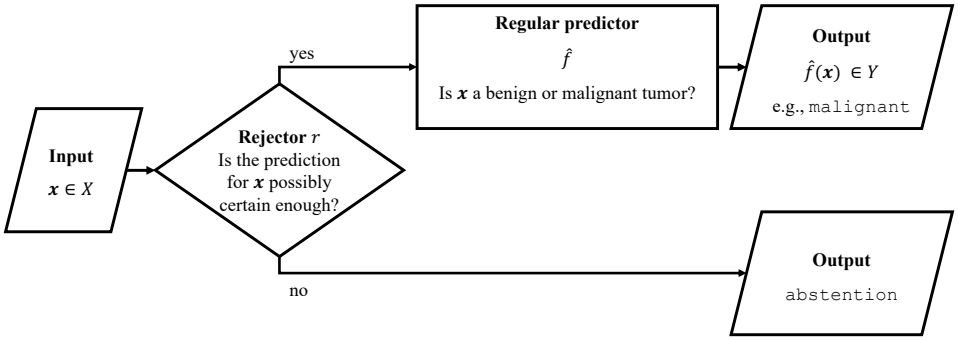


Figure 6.5: Pre-algorithmic attachment of abstention.

The question of the rejector, "Is the prediction for $\boldsymbol{x}$ possibly certain enough?", could, for example, be decided by determining the *distance* d of $\boldsymbol{x}$ to the training data. For example, the rejector could decide that, if $\min_{i \in \{1,\ldots,n\}} d(\boldsymbol{x}, \boldsymbol{x}^{(i)}) > t$, the output `abstention` is generated and

if $\min_{i \in \{1,\ldots,n\}} d(\boldsymbol{x}, \boldsymbol{x}^{(i)}) \leq t$, the data point is forwarded to the predictor $\hat{f}$ with $d(\boldsymbol{x}, \boldsymbol{x}^{(i)})$ being some distance metric between $\boldsymbol{x}$ and a training data point $\boldsymbol{x}^{(i)}$.

(b) *Post-algorithmic attachment*

For post-algorithmic abstention, the rejector is downstream of the predictor (Campagner et al., 2019; Brinati et al., 2020; Artelt et al., 2022). For every input data point, an ordinary prediction is calculated. This is done independently of any abstention activity. The prediction is computed in the exact same way the prediction would be computed in a non-abstaining system. This means that the question that is under discussion, Q, is answered by choosing one of the defined answers from A. In the second step, the certainty of the prediction, i.e., the likelihood of the selected defined answer being the correct answer is measured. This certainty can be provided by the predictor itself (e.g., as some kind of probability value in a neural network, distance in a support vector machine, or some "soft probabilistic classifier" (Campagner et al., 2019; Brinati et al., 2020)) or it can be calculated additionally by some uncertainty or reliability measure (Linusson et al., 2018; Mouchère and Anquetil, 2006b; Lotte et al., 2008). This certainty value is then used in the posterior attached rejector. In the simplest version, the rejector only consists of a certainty threshold and two *if*-clauses. If the certainty of the calculated answer being correct is above the threshold, the prediction is passed through and revealed; if the certainty is below the threshold, the predicted answer is rejected, and the system abstains.[16] The concept of post-algorithmic attachment is visualized in Figure 6.6.

[16] Although the abstaining part of this type of model is attached, it corresponds best to what is called a "dependent rejector" in Hendrickx et al. (2021). The term "dependent rejection" used by Hendrickx et al. (2021) implies that the rejection of a particular input *depends* on the previously calculated output of the predictor. This stands in contrast to what I refer to as the (attached) *pre-algorithmic* abstention models, wherein the rejection of an input occurs prior to the predictor's calculation and is, in that sense, *independent* of the predictor.

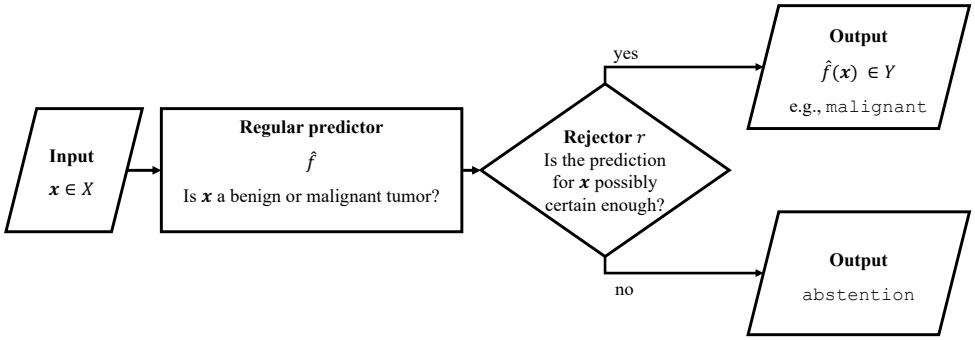


Figure 6.6: Post-algorithmic attachment of abstention.

Both pre-algorithmic and post-algorithmic attachment are attached forms of abstention since the abstaining part is in both forms an additional, separated algorithm that is attached (either prior or posterior) to the predictor. Attached abstention could also be called *threshold abstention* as the decision whether a sample is rejected or not is usually based on comparing some certainty (in the case of post-algorithmic abstention) or similarity (in the case of pre-algorithmic outlier abstention) to a defined threshold, see Hendrickx et al. (2021).[17]

Merged Abstention

The crucial difference between merged and attached AML systems is that for the merged systems the abstaining and predicting activity are to some extent inseparable. The abstaining activity is neither upstream nor downstream of the prediction but is included in the predicting activity. Therefore, it is not practical anymore to refer to "the predictor" and "the rejector." Instead, the predictor is modified to have the capability to reject as well. For merged AML systems, we can aptly name the modified predictor an "abstention predictor."

In a classifier, an extra, abstaining output is introduced. In addition to the outputs represented by the defined answers, there is also the abstaining

[17]Note that there are varieties of attached AML systems that do not include a pre-set certainty threshold. For example, it is possible to reject a fixed fraction of the samples. In this approach, it is not a matter of rejecting all samples below a specific *certainty threshold*; instead, a *fixed fraction* of the most uncertain samples, for instance, the bottom 10%, is rejected.

output. For a given input (e.g., a dog image), the system can either output one of the defined answers (e.g., `Husky`, `Labrador`, etc.) or output the `abstention` output.

The property of being "merged" can be observed both in the application phase and in the learning or training phase of the algorithm. In the application phase, the fact that the AML system is "merged" is illustrated by the fact that decisions about whether to abstain on an input are made neither before nor after the decision about which output to assign (if any). The decision about abstention is made simultaneously with, and as part of the decision about the appropriate output. In the application phase, `abstention` is simply one additional output among others and in this sense one additional answer. For this, we do not use the regular predictor $\hat{f}$, but a special abstention predictor $\bar{f}$, which also allows for abstention. The application phase of a merged AML system can be visualized in the following flowchart:

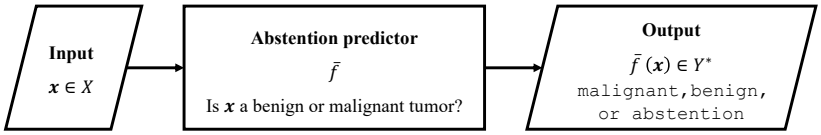


Figure 6.7: Merged Abstention: The decision about abstaining or not is made simultaneously to the decision about the class by an adapted abstention predictor $\bar{f}$.

In order to obtain such an abstention predictor $\bar{f}$, the training phase of a merged AML system has to be adapted. Those adaptions in the training phase, i.e., the way in which the abstaining option is learned, illustrate the second dimension in which merged AML systems differ from attached AML systems. For merged AML, the tasks of rejecting and predicting are blended into one task that is *learned simultaneously* in the training phase.[18] While it is possible for an attached AML to have the same learning phase as a non-abstaining classifier, the learning phase of a merged AML is necessarily different from a non-abstaining classifier.

[18]This is also why Hendrickx et al. (2021) call this type of learning *simultaneous learning* as contrasted with sequential learning.

With Labeled Abstention (a) and Unlabeled Abstention (b), I will distinguish again between two ways of how the learning phase of a merged AML system can allow for abstention-learning. This distinction concerns only the training phase and the way the abstaining class is learned.

I will explain this by means of the cancer detection example from Subsection 6.2.1. There, I introduced how a *regular, non-abstaining* classifier $\hat{f}$ can be trained on the training data visualized in Figure 6.1. This training or learning phase can now, in principle, be adapted in two ways in order to allow for abstention.

(a) *Labeled Abstention*

A simple solution for training a system when to abstain is to extend the general method of supervised learning from the normal outputs to the abstaining output. In the training phase, a classifier is usually given examples of inputs (e.g., images of dogs) along with the correct (ground-truth) label or output we want for that particular image. For the dog classifier, in the training phase, the system would be presented with multiple images of huskies all labeled `Husky`, multiple images of retrievers all labeled `Retriever`, etc. The system is shown what a conventional input of a dog image looks like, for which we want to have `Retriever` as the output. Analogously, we can now proceed for the abstention class. One can label inputs for which one would consider abstention appropriate with the label `abstention` and put them into the training phase just like the examples of all other classes (Lotte et al., 2008; Mouchère and Anquetil, 2006b; Singh and Markou, 2004).[19] For example, one could label images of Shepherds, Bulldogs, or images of cats by hand with `abstention` since these images should be considered outliers. Moreover, blurry images or images where the dog is only partially visible can also be labeled `abstention` by hand. Thereby the set of defined answers is in a sense extended from $\{$`Husky`, `Labrador`, `Dachshund`, `Retriver`$\}$ to $\{$`Husky`, `Labrador`, `Dachshund`, `Retriever`, `abstention`$\}$.

Considering the example in Figure 6.1, in the original,

[19] This need not to be the end result of training the classifier. In Singh and Markou (2004), the authors use the rejected training data to retrain the classifier with potentially new classes earlier detected as outliers.

non-abstaining case, a training data point was a tuple $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$ with $\boldsymbol{x}^{(i)} \in X = \mathbb{R}^2$ and $y^{(i)} \in Y = \{\texttt{malignant}, \texttt{benign}\}$. In the case of labeled abstention, some of the training data points have the label `abstention`, i.e., $y^{(i)} = \texttt{abstention}$. Hence, for a training data point $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$, it is $y^{(i)} \in Y^*$ with $Y^* = \{\texttt{malignant}, \texttt{benign}, \texttt{abstention}\}$. The training data for this would be the set $T^* = \{\langle \boldsymbol{x}^{(1)}, y^{(1)} \rangle, \langle \boldsymbol{x}^{(2)}, y^{(2)} \rangle, \ldots, \langle \boldsymbol{x}^{(n)}, y^{(n)} \rangle\} \subseteq X \times Y^*$. In this approach, there is no categorical change required for the loss function. The loss function only needs to be extended to accommodate the extra class. The loss function for the non-abstaining, binary classification from Equation (6.1) is a function from $Y \times Y$ to the loss $\{0, 1\}$. A loss function for the labeled abstaining case can be the same as l, only mapping from the extended sets, i.e., from $Y^* \times Y^*$. The training data for labeled abstention is visualized in Figure 6.8.

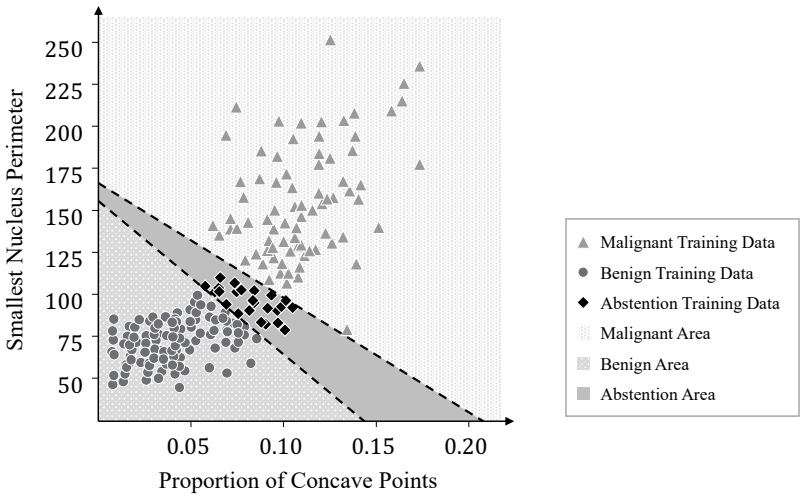


Figure 6.8: Labeled Abstention: Training data is either labeled `malignant` (triangle), `benign` (circle), or `abstention` (diamond). The model learns three different areas for the three different classes: a malignant area (top), a benign area (bottom), and an abstaining area (middle).

In this classification problem, simply three classes instead of two are considered. This means that the model needs to learn two boundary lines instead of just one as evident in Figure 6.8. Using the dog

image example, if there are initially *three* separating boundaries, the model would have to learn *four* to additionally separate the `abstention` class from the others. So, the difference between the labeled abstention training and the normal, non-abstaining training is the same as the difference between the training of a classifier with 5 dog breed classes and the training of a classifier with 4 dog breed classes.

This approach has two major drawbacks, though. First, labeling training data points with `abstention` by hand can be very time-consuming. Second, often it is not useful to label training data as `abstention`. While in some application domains, we know exactly what a prototypical abstention case might look like (e.g., a blurred image for an image classifier), often we do not, or at least not in advance. In particular, when the uncertainties are due to factors that cannot be readily detected by humans looking at the training data, we cannot tell which samples will be error-prone. Often, the samples that are difficult for the algorithm to process are easy for a human expert and vice versa. This suggests that the human expert will not be able to identify the difficulties for the machine, so that it is unclear how the abstention labels are determined in the training data.

(b) *Unlabeled Abstention*

Besides the straightforward way of inserting abstention as an extra output in the learning process as described in case (a), there is a more indirect, but also more sophisticated way. Here, the training data is not explicitly labeled `abstention`. In systems like those of Thulasidasan et al. (2019b); Geifman and El-Yaniv (2019); Mozannar and Sontag (2020); Wegkamp and Yuan (2011); Barnes and Barnes (2021); Yuan et al. (2020), the training data looks exactly the same as in a training situation of a *non-abstaining* classifier. There are images of the different dog breeds, and each image is labeled with one of the normal (defined) labels, i.e., `Husky`, `Retriever`, `Dachshund`, or `Labrador`. No training image has the label `abstention`. Hence, for our main working example from Subsection 6.2.1, the set of training data for the unlabeled abstention case

would be $T = \{\langle \boldsymbol{x}^{(1)}, y^{(1)} \rangle, \langle \boldsymbol{x}^{(2)}, y^{(2)} \rangle, \ldots, \langle \boldsymbol{x}^{(n)}, y^{(n)} \rangle\} \subseteq X \times Y$ with $\boldsymbol{x}^{(i)} \in X = \mathbb{R}^2$ and $y^{(i)} \in Y = \{\texttt{malignant}, \texttt{benign}\}$.

Therefore, the usual supervised way in which an ML system learns to associate an input with a desired output is not applicable to the abstention cases. In order for the system to learn a connection between certain images and the abstention output, the underlying learning process, i.e., the loss function itself must be adjusted.[20]
This can be implemented when for a given training data point, it is possible not only to produce a full loss (if the point is misclassified) or no loss (if the point is classified correctly), but also a small loss if the point is not classified at all. For the breast cancer classifier, the normal (non-abstaining) loss function of Equation (6.1) was introduced as a function that takes the value 1 for each misclassified data point and the value 0 for each correctly classified point. The abstaining loss function could then include an additional loss of, say, 0.2 if the system does not classify `benign` or `malignant` but instead chooses the `abstention` output for a given input (regardless of what the point's actual ground-truth label is).[21]

In the case of unlabeled abstention, we look for $\bar{f}$ in the set of the candidate functions $\mathcal{F}^*$, which consists of functions of a particular model choice that maps from $X = \mathbb{R}^2$ to

[20] In Hendrickx et al. (2021), the authors present another approach to learning to abstain and predict in what they call a "simultaneous learning" way. This does not require labeling the input data or directly adjusting the loss function. This workaround is usually based on combining different algorithms, each of which executes only one predicting task. For example, if there are four ordinary classes, i.e., four defined answers, one could train four different classifiers in a "one vs. all" training. This can, for example, be implemented via several support vector machines (SVM), as it is done in Wu et al. (2007). The combination of the four trained SVMs then possibly yields areas of overlap or areas that none of the classifiers considers to belong to its trained class. These areas can then be seen as abstaining areas. In my framework, I do not consider these types of algorithms to be merged systems, though. Although they do not perfectly fit the prototype of attached systems either, abstaining and predicting still happen in different parts of the algorithm. Plus, the systems do not really learn what abstaining cases look like. This will become relevant for my considerations in Subsection 6.3.3.

[21] Depending on the context, it might actually make sense to assign different penalties for abstaining for different ground-truth labels. In my example, it might make sense to rate "false negatives" worse than "false positives." Consequently, abstention for benign cases could be penalized more than abstention for malignant cases (Zheng et al., 2011).

$Y^* = \{\texttt{malignant}, \texttt{benign}, \texttt{abstention}\}$. While the set of the candidate functions was also $\mathcal{F}^*$ for the case of labeled abstention, in unlabeled abstention training, the loss function l^* needs to be adjusted, too. For each single training data point $\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$, $l^*(y^{(i)}, f(\boldsymbol{x}^{(i)}))$ can add either a loss of 1 for misclassification, a loss of 0 for correct classification, or a loss of some α if the system abstains on this point. Hence, $l^* : Y \times Y^* \to \{0, 1, \alpha\}$,

$$l^*(y^{(i)}, f(\boldsymbol{x}^{(i)})) = \begin{cases} 1 & \text{if} \quad y^{(i)} \neq f(\boldsymbol{x}^{(i)}) \text{ and } f(\boldsymbol{x}^{(i)}) \neq \texttt{abstention}, \\ \alpha & \text{if} \quad f(\boldsymbol{x}^{(i)}) = \texttt{abstention}, \\ 0 & \text{if} \quad y^{(i)} = f(\boldsymbol{x}^{(i)}). \end{cases} \tag{6.2}$$

Note that $\alpha \in (0, 1)$ since for $\alpha \leq 0$ the system would always abstain and for $\alpha \geq 1$ never abstain. If the same α is chosen for all classes, it has been noted in Ramaswamy et al. (2018) that $\alpha \leq \frac{m-1}{m}$ for m being the cardinality of Y, the number of possible ground-truth labels.[22] In our example, $m = 2$. This means that choosing to abstain has to be always less costly than making a random guess for a particular point. The closer α is to 0, the less it costs for the system to abstain, i.e., the more the system will abstain. If α is close to $\frac{m-1}{m}$, the system will learn to abstain only rarely, since abstention is almost as costly as making a random guess.

The distinction between l and l^* shows the principle of how a loss function can be adapted to allow the system to learn abstaining. It should be noted that this is a simplified loss function used for illustrative purposes. The loss functions in the literature are more complicated and designed to be handled numerically well (Thulasidasan et al., 2019b,a; Geifman and El-Yaniv, 2019; Yuan et al., 2020; Barnes and Barnes, 2021).

[22] This can be seen following Chow's rule for an optimal abstention rate (Chow, 1970). According to this rule, Equation (6.2) states that the system should abstain iff the probability of the likeliest output is smaller than $1-\alpha$. Note that this is only one *necessary* upper bound for α. If the prior probabilities for the different classes are highly unequally distributed, α should be bounded even more. In fact, in this case, considering different α values for the different classes is reasonable as noted in Footnote 21.

In Equation (6.2), we see that the option to abstain is *merged* into the loss function l^* and thereby merged into the training of the classifier. Predicting and abstaining are trained at the same time. A trained unlabeled classifier is illustrated in Figure 6.9. In contrast to this, attached AML systems can only learn in a sequential way. First, for example, it is learned how to classify and only then it is learned how to abstain. Moreover, the prototypical systems of attached AML systems that I presented here do not even *learn* to abstain but are rather *told* by the programmer when they should abstain.

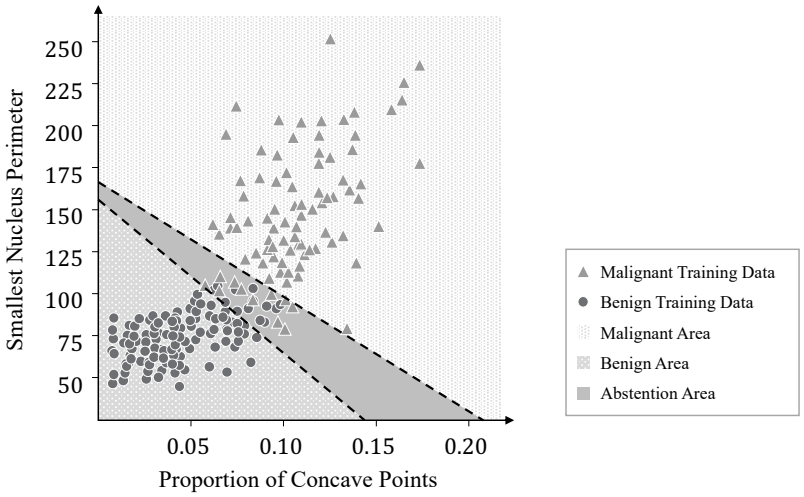


Figure 6.9: Unlabeled Abstention: the training data consists only of malignant (triangle) and benign (circle) points. Due to the adaption of the loss function, the model learns to separate three areas: a malignant area (top), a benign area (bottom), and an abstaining area (middle).

6.3 Philosophical Analysis

6.3.1 Comparison of Suspension and Abstention

In the following section, I want to investigate how far the phenomenon of abstention, described in the previous section, matches its epistemological counterpart: suspension of judgment as it was presented in Chapter

2. Here, I draw parallels between abstention and suspension, but also stress the points where the analogy ends. First, I compare the reasons to abstain that were presented in Subsection 6.2.2 with reasons to suspend (Part 6.3.1.a), and second, I compare the different ways abstention is implemented, which was investigated in Subsection 6.2.3, with different forms of suspension, which I presented in earlier Section 2.2.

As elaborated before, in doxastic terminology, belief and disbelief are characterized by taking one of the defined (or complete) answers to a question to be true. Suspension is characterized by not choosing or not committing to the truth of any of the defined or complete answers, i.e., being neutral towards the defined answers. As I have done already in Chapter 2, I will sometimes talk about suspension towards a proposition p rather than towards a question Q in the following.

6.3.1.a Reasons for Suspension and Abstention

In Chapter 2, we saw that epistemologists describe at least two different types of justifications for suspension. One type of justification involves factors that *positively support* suspension, while the other type of justification arises when there is neither a reason to believe nor a reason to disbelieve. In other words, there is no compelling reason to choose any of the defined answers to a question. In this case we have *privative* justification for suspension.

In AML systems we find a correspondence with both these types of justifications. When we look at the justifications for abstention, we can observe that the different justifications considered in epistemology are at play in abstention for ambiguity cases and in abstention for outlier cases. When a system abstains due to ambiguity, a particular input data point is considered ambiguous, meaning that the point is in a region where two (or more) classes overlap. In ambiguity abstention, we have positive evidence for class A *and* positive evidence for class B. For example, some features of the input image speak for the class `Husky`, while others speak for the class `Retriever`. Therefore, we abstain in a *privative way*. There is no positive evidence for abstention, but conflicting evidence for different classes.

This is different for outlier abstention. Here, the system abstains from classifying an input sample *because* the sample is an outlier. The sample

being an outlier is *positive evidence* for abstention on that sample. This can be seen in cases of pre-attached abstention, where it is decided that abstention is the correct output even before the system is queried about the question and the possible defined answers. This particular input or question is not to be decided by the algorithm. Hence, we can conclude that both, cases of being neutral due to privative reasons and cases of being neutral due to positive reasons are present in AML systems. In conclusion, the results for the reasons for suspension and abstention are summarized in Table 6.1.

Reasons for Abstention in Machine Learning	Justification for Suspension in Philosophy
Ambiguity Abstention	Privative Justification
Outlier Abstention	Positive Justification

Table 6.1: Different reasons for abstention in AML and different corresponding justifications for suspension.

6.3.1.b Nature of Suspension and Abstention

The more complex question pertains to the relationship between the nature of suspension and the implementation of abstention in AML systems. In the broader context of assessing the actual "intelligence" of various AI systems and their ability to mimic to human reasoning processes, it is crucial to explore whether different AML systems can mimic what we call "suspension of judgment" when abstaining on a specific question. In Chapter 2, I outlined how philosophers characterize this phenomenon and how they identify various forms of it. Subsequently, I will illustrate whether abstention within the different AML system implementations aligns with the nature of suspension as illustrated in Section 2.3. I will show how we can draw parallels between the different methods of abstaining in AML and the different forms of suspension.

One way to start investigating these topics is to precisely describe which question is addressed by suspending or abstaining. As described in Chapter 2, suspension can be characterized as a way of behaving doxastically to a

question that is under discussion (or to an answer to the question). This means that suspension is one way of responding to a question Q by *not choosing* one of the defined answers. Suspension is characterized as one possible stance towards the question under discussion, e.g., "What kind of dog is on this image?", different from both belief and disbelief. Basically, abstention in ML algorithms describes a similar phenomenon, namely the generation of an output with respect to a question that does not match any of the defined answers.

In the case of attached systems, the analogy between suspension and abstention can be drawn only to a limited extent, though. In attached systems, two different questions play a role in generating the abstention output. One question is the actual question under discussion, i.e., "Which kind of dog is in the image?" that is to be answered by the predicting algorithm. The second question is of the type: "Is the (possible) answer to the first question certain enough?"

In the case of post-algorithmic attachment, the question under discussion is answered first. This is done in a conventional sense, i.e., in exactly the same way as in a non-abstaining system. A defined answer (e.g. `Husky`) is generated.[23] Only afterward the second question ("Is this answer certain enough?") is asked. This is the question that is answered by the abstaining part of the algorithm. Hence, in this picture, abstention is not a response or attitude towards the question under discussion, but a response to the second question asked about the certainty of the first answer. In the case of pre-algorithmic attachment, we find a similar situation, but the order of the questions is reversed.

Therefore, for attached systems, the analogy between suspension and abstention fails in so far as suspension is supposed to address the same questions as the other possible doxastic attitudes. Suspension is a response towards the question under discussion. Abstention in attached systems is an answer to a different question than the question under discussion.

This is different for merged systems. Here, abstention is considered

[23] One has to acknowledge that the answer is more informative than just choosing one class, i.e., when the question is answered, there is more information present, e.g., about the probability for this answer being the correct one, and about the probability for other answers.

an extra class among the other options for classification. Thus, abstention is one response to the question under discussion. The system is asked: "What kind of dog is on this image?" and responds either by providing a defined answer (e.g. `Husky`) as the output class or responds by choosing the abstaining output class. As described earlier, abstention and prediction are parts of the same process and occur simultaneously. Thus, abstention addresses the question under discussion directly.

In addition, the different implementations of merged systems (labeled vs. unlabeled) can also be compared with different forms of suspension found in the philosophical literature. In Section 2.4.1, I introduced a distinction between epistemic suspension and indeterminacy suspension that goes back to Ferrari and Incurvati (2022).[24] This distinction consists of different attitudes towards whether the question is in general answerable or not. One stereotypical case for indeterminacy suspension is a case of mathematical indeterminacy for which a subject can conclude that the proposition is in fact neither true nor false but ontologically indeterminate.[25] In cases of epistemic suspension, the subject will take the question to be in principle decidable, but not according to their current epistemic stance.[26]

This difference in attitude regarding the question is also found to some degree in the labeled and unlabeled implementations of the merged systems. On the one hand, we have merged systems that learn abstention in a labeled way. We externally tell the system in the training phase which input data (e.g., images) should trigger the response `abstention`. Here, `abstention` is considered one ground-truth label of the image. In a certain sense, we ascribe an indeterminate state to these images, which is supposed to be accompanied by abstention. We basically say, no matter how the parameters of the classifier are selected, this image is not to be classified

[24] Note again that Ferrari and Incurvati (2022) take the term "agnosticism" to refer to the broad concept that subsumes different versions. I take "suspension" to be this broad term. Hence, I will use the term "suspension" in the following when Ferrari and Incurvati (2022) would talk about "agnosticism."

[25] As noted in Chapter 2, the most prominent case is the continuum hypothesis (Gödel, 1947).

[26] As noted in Section 2.4.1, this distinction does not align precisely with the one described by Ferrari and Incurvati (2022). In their work, both epistemic and indeterminacy suspension are considered "pessimistic," indicating that the subject does not believe that further inquiry will ultimately resolve the question in a positive or negative manner. In my slightly modified interpretation of this concept, such pessimism is not a requirement. Here, especially in the case of epistemic suspension, it can coexist with the hope of improving one's evidential situation.

(by a defined answer or label).

Moreover, abstention in such an implementation no longer exactly fulfills the role we ascribed to it in the description of the overarching phenomenon. I described both suspension and abstention as ways of responding to a question *without* selecting one of the defined answers. I diverge from this picture when abstention is learned in a labeled way. Then, abstention no longer represents the non-selection of a defined answer but represents a defined answer itself. In the training phase, abstention is treated exactly analogously to the other classes, and the abstention output is learned in exactly the same way as the other outputs. In the loss function, the abstention output is not handled separately. The loss calculated for misclassifying a point with the label `abstention` is conceptually equal to that of misclassifying a point with any other label. By labeling certain training data as `abstention`, we treat abstention as a regular class among the others and, thus, as one of the defined answers.

Ferrari and Incurvati (2022) draw a similar picture regarding indeterminacy suspension. They argue that this kind of suspension could be argued to not count as suspension at all if the question is opened to the extent that indeterminacy is one of the conventional, defined answers. The answer set is just expanded, such that it can account for indeterminacy cases. However, choosing this answer is no different from choosing any other answer.

In merged systems, in which abstention is learned in an unlabeled way, the situation is different. Here, abstention is also a possible output class, but it has a special role compared to the other classes. The abstaining response addresses the question in a different way than the other outputs (the defined answers). Abstention is not learned by explicit abstention prototypes, but by giving the system the option not to select any of the other classes in cases of unclear data. In this case, abstention is a way of opting out of choosing one of the defined answers. It reflects epistemic uncertainty. There is uncertainty about the correct defined answer, but it is not assumed that the correct defined answer could not be found in a better evidential situation, or that the correct answer to this question *is* `abstention`. This is similar to the case of epistemic suspension.

The special role of abstention here goes hand in hand with the different

accounts of (sophisticated) suspension that I discussed in Chapter 2. This alignment is particularly evident when comparing it to other neutral stances like mere non-belief or deep ignorance and contrasting it with its doxastic counterparts of belief and disbelief. As I outlined in Subsection 2.3.3, authors arguing for a meta-cognitive account of suspension clearly express their viewpoint that suspension carries a role of more "cognitive sophistication," setting it apart from belief and disbelief (Crawford, 2004; Raleigh, 2021). According to this perspective, the special role of suspension (compared to belief and disbelief) lies in its nature as a higher-order attitude. According to this picture, suspension presupposes indecision, which is then qualified to suspension as the subject acknowledges their own indecision by forming a belief about this lack of certainty (Crawford, 2004; Raleigh, 2021). Similarly, in the competing model of Wagner (2022), suspension occupies a distinct position and is viewed as a more sophisticated attitude than belief and disbelief. In the account of Wagner (2022), suspension also presupposes indecision, which is then qualified as suspension through the process of endorsement. In a parallel manner, abstention in unlabeled merged systems plays a special role compared to all other standard output choices. This is characterized by a certain overview when recognizing that choosing one of the defined answers would be problematic. The parallel is especially evident during the learning phase of these systems. Although the system is assigned the task of determining a predefined regular answer for all data points, in certain cases, it evaluates that abstaining is a more favorable option (in terms of cost) than providing a specific answer.

It might be argued that the meta-cognitive form of suspension, which consists of a belief about the own evidential situation, can be found in attached systems, too. (Post-) attached systems can be said to evaluate their evidential situation in terms of probabilities or certainty for specific outputs. While this process might have a meta-cognitivist appearance, it is distinct from what philosophers have in mind when talking about suspension being meta-cognitive. For suspension as a meta-cognitive attitude, there *first* must be indecision as such, which is *then* evaluated by

a kind of introspection on a second level.[27] For post-attached systems, we find two disanalogies with this picture. First, in post-attached systems, there is no indecision at all, since an answer has de facto already been selected. As I have argued, the question under discussion is here answered in a non-abstaining way by selecting one of the defined answers; abstention addresses a different question than the question that is under discussion. Second, it seems arguable whether there really is an evaluation of one's *own* evidential situation. On the contrary, it could be argued that the predicting and abstaining parts are two systems. In this respect, it is difficult to speak of the abstaining part evaluating *its own* evidential situation. The results for how the different implementations of abstention correspond to suspension are summarized in Table 6.2.

Implementation of Abstention	Qualification for Suspension?	Form of Suspension
Attached	*no*	–
Merged	*yes*	Indeterminacy for *Labeled Abstention* Epistemic for *Unlabeled Abstention*

Table 6.2: Correspondence of the different implementations of abstention in AML with the nature of suspension as well as with different forms of suspension.

6.3.2 Autonomy of Abstaining

In this section, I aim to explore the autonomy of abstention in various AML systems. The level of autonomy in the outputs of ML systems is an important topic when philosophically assessing the appropriateness of ascribing intelligence to artificial systems (Russell and Norvig, 2021).

[27] The connection between indecision and the second-order belief is different in the account presented by Raleigh (2021). In his model, the second-order belief is constitutive for indecision and, in this context, takes precedence. Nevertheless, the crucial point is that, in practice, all of these approaches involve a state of indecision concerning the proposition p.

Consequently, it becomes imperative to examine the autonomy of AML systems, especially concerning their abstaining output. The term "autonomy" is discussed controversially in the philosophy of AI and is not easy to define. Nevertheless, there are two (connected) desiderata that are emphasized repeatedly and that emerge as commonly accepted criteria in debates around autonomous AI. First, the way from the input to the output is *not* supposed to be completely *hard-coded* by the programmer, and second, some kind of *flexible learning* has to be involved.

Johnson and Verdicchio (2017, p. 576), for example, define autonomous AI as "computational artefacts that are able to achieve a goal without having their course of action fully specified by a human programmer" and claim that "learning can play a significant role in seeming to expand the autonomy of computational artefacts" (Johnson and Verdicchio, 2017, p. 583). Anderson and Anderson (2011) also stress that autonomy can only be present if the behavior of the system is not micro-managed by humans. Russell and Norvig (2021, p. 42) claim that "to the extent that an agent relies on the prior knowledge of its designer rather than on its own precepts and learning processes, we say that the agent lacks autonomy."

The two criteria are also emphasized in the discussion on artificial agency which is a concept that is closely related to autonomy (Russell and Norvig, 2021). As noted in Eva et al. (2022), a model of an artificial agent has to make sure that the agent is set up in a way such that it can make its own decisions and is not pre-programmed for all actions and all circumstances. Also, Müller and Briegel (2018) emphasize that "free agents have to be learning agents" and that the learning history of an agent becomes part of the agent's identity and explains the agent's behavior. These learned but flexible behavior patterns make it possible to attribute actions to the agent itself (see also Briegel and Müller (2015)).

Apart from these two necessary criteria for artificial agency and autonomy, Bradshaw et al. (2013) emphasize that it makes sense to speak of autonomous *capabilities* rather than of autonomous *systems* as such since there will always be some activities or capabilities of one system that are autonomous while others may not. I agree with this shift of perspective. In this section, I specifically ask about the autonomy of the *abstaining* capability rather than about the autonomy of the predicting activity or the autonomy of the system itself.

To determine the autonomy of the *abstaining* capability of a systems the two minimal demands for autonomy should be assessed for the abstaining activity *in the same way* as for the predicting activity. This means that I demand that (a) the way in which a system arrives at the abstaining output should not be completely hard-coded and (b) the connection from the input to the output `abstention` should be in some way learned by the system.

The system should be able to independently establish a correlation between certain aspects of the inputs and an `abstention` output. Not all ML systems belonging to the class of abstaining ML meet this requirement. Attached systems typically consist of an ML system that is trained on the data and that is responsible for predicting, *and* an additional rejection part that is responsible for the abstention task. Thus, in the attached AML systems, the act of abstention is performed by an algorithm that is separate from the algorithm that performs (in a fairly autonomous ML fashion) the task of prediction. Often, the abstention part of the algorithm is itself a simple, hard-coded piece of the program that is not connected to the machine learning part.[28] Therefore, the kind of autonomy that is present for the predicting capability in ML systems is not present for the abstaining capability in attached AML systems. We can say that attached AML systems do not abstain *as autonomously* as they predict.

This is different for merged AML systems. Merged abstention systems autonomously abstain to the same extent that (regular) ML systems make decisions autonomously. In merged systems, the option of abstention is offered in the training phase, and the system establishes a connection between the features of the input data and an abstention output. Though in different ways, this connection is made both in labeled and unlabeled merged systems. A merged system can be described as learning to identify situations where a prediction is too risky and thus can be viewed as evaluating its own evidential situation independently of the programmer. In this sense, a merged abstention system can be described as "knowing when

[28] However, it is possible that the attached abstention part involves some kind of learning. For example, the optimal rejection threshold may also be learned (De Stefano et al., 2000). Still, this type of learning does not involve (autonomously) establishing a link between the input data and an abstention output.

it doesn't know" (Thulasidasan et al., 2019b). Note that I do not claim that merged AML systems abstain autonomously, but rather that in contrast to attached systems, they meet the minimal criterion of autonomous abstaining. The abstaining activity is not hard-coded but learned in some way. Merged AML systems are *as autonomous* in abstention as they are in prediction.

6.3.3 Explainable Abstaining

Beyond the issue of autonomy, explainability is a widely debated topic in the field of (the philosophy of) artificial intelligence, often interconnected with concepts such as interpretability and understanding. This subsection is intended to give a first idea of how investigations about the explainability of AI systems can be extended to abstaining ML systems.

One of the four key principles of explainable AI that are established in Phillips et al. (2020, p. 2) is the *Explanation Principle*, which states that "Systems deliver accompanying evidence or reason(s) for all outputs." This can be issued, for example, in a procedural way (How did the system reach this output?), in a contrastive way (Why did the system output *this* instead of that answer?), in a recourse way (What do I need to change in the input in order to get another output?). Here, I will focus on local (or instance) explanations, i.e., explaining why a *particular* input sample produces a *particular* output (Burkart and Huber, 2021).

The explanation principle of Phillips et al. (2020) requires *all* outputs to be accompanied by a reason or explanation. Hence, when considering AML systems, we must apply this demand not only to the defined answers but also to the abstaining output.[29] In particular, if we want to *learn* something from the abstaining response by improving the input data, examining certain characteristics more closely, or making the training data more diverse, it is useful to know *why* the system reports that it cannot make a decision. Some first approaches to provide explanations for abstaining responses can be found in Artelt et al. (2022); Artelt and

[29]There are certainly cases where we would intuitively demand an explanation for the defined answers but are fine without an explanation for the abstaining output. Abstaining represents precisely the cautious reaction that does not directly provide us with a decision-making aid in any direction. Therefore, it is sometimes not necessary to ask for an explanation for this option, as long as it is seen as a fallback option that can be used when all other options fail. Still, we would become skeptical if it was used too much.

Hammer (2022); Thulasidasan et al. (2019b).[30]

When we ask for a (local) explanation about the system's abstention on a particular input, we ask about *why* the system abstained on that input or about the *reason* for abstaining on this input. Therefore, the explanation should refer back to the input in some way and point out which parts of the input were responsible for the response (abstaining in this case). For outlier abstention, this is rather trivial. Abstaining on an outlier can always be explained by referring to the relationship between the training data and the input data point that makes the point an outlier. An explanation is always available and not very informative. The more interesting cases are cases of ambiguity abstention. Thus, I will focus on these in the following.

The distinction between merged and attached systems, which I made in Subsection 6.2.3 again becomes relevant for this question about explainability because merged and attached systems allow different options for explanations.

In merged systems, it is (in principle) possible to refer back to the characteristics of the input that are responsible for the abstaining output. If we ask for a reason why the system abstains on a particular input, a merged system can provide such an explanation by pointing to particular features of the input sample just as it can point to the input features that are responsible for, e.g., the output `Husky` or the output `Dachshund`. This possibility arises from the fact that merged systems learn to associate certain input characteristics with an abstention. The system thus establishes correlations between characteristics of the input data and an abstention label and can provide the reasons (i.e., some characteristics of the input sample) for abstention. This can serve as a local explanation.

While this seems rather obvious for labeled merged systems, it is interesting to see that this possibility is also available for unlabeled systems. For example, Thulasidasan et al. (2019b) use visualization techniques like the one of Selvaraju et al. (2017) to visualize the areas in input images that were relevant for abstaining. Thulasidasan et al. (2019b) tested their

[30] On a different note, it is also interesting to evaluate how well the AML classifiers do. An explicit approach to provide metrics for evaluating the results of abstaining classifiers can be found in Ferri and Hernández-Orallo (2004).

(merged) deep abstaining image classifier ("DAC") for different abstaining situations. They never labeled the training data with `abstention`. In a first case, they took 10% of the training data images and randomized the ground-truth labels. Hence, the ground-truth labels of these images were not correct. There was no regularity in the image-label connection. For tracking, they included a "smudge" on these images with randomized labels. In a second experiment, they took all the training images of one class (all monkey images) and randomized the labels while not providing any smudge. In comparison to the first experiment, the noise they created here was "structured." In both experiments, they applied a heat map to the test data, which was supposed to visually highlight the areas of the image that are especially relevant for a certain output. In the first experiment, they found that the system established a correspondence between the smudge and the abstention output. In the heat map, the smudge was highlighted as the part of the image that was decisive for the abstention output. In the second experiment, the typical monkey features were highlighted. This means that the system established a correspondence between either the smudge or typical monkey features and an abstention output, even without being provided with labeled prototypical abstaining cases in the training phase.[31]

This shows how even a merged system that learned abstention not through explicitly `abstention` labeled training data can still find a connection between certain features of the input space and an abstention output. Thus, one can exploit the full range of local explanations that is available for conventional non-abstaining classifiers. Not only heat maps but any explainable method that is available for regular outputs can be applied to these systems.

[31] A comparable experiment setup can be found in Barnes and Barnes (2021). The authors also experiment with corrupting the labels of exactly one (or two) classes. In another experiment, the authors simply corrupt a certain percentage of labels from the training data of all classes. Barnes and Barnes (2021, p. 3) notice that "in this case, there is no systematic relationship between the input maps and whether the sample is corrupted or not. For [these] mixedLabels, we would like the CAN [controlled abstention network] to learn to abstain on the corrupted training samples by identifying them as those that do not behave like the majority of the training samples." It is interesting to see that in this setup there is no intended or pre-specified correlation between input features and the abstaining output. Still, when they test the abstaining system and compare it to the results of a non-abstaining, all-knowing oracle, which serves as an upper bound for accuracy, the results in terms of accuracy (i.e., how many test data points are classified correctly) are nearly ideal.

For attached systems[32] this is not possible. The system does not find any connection between the characteristics of the input and the abstaining output. It merely learns to connect the characteristics of the input with the conventional outputs. The abstention option, however, is imposed on the system afterward. The attached system abstains when issuing a conventional response is associated with too much uncertainty. So, if we ask for the reason why the system abstains for the specific input $\boldsymbol{x}$, the answer (and thus explanation) can only be: "because the certainty for providing a correct answer is below the threshold." Of course, the system can give us information beyond that, such as how far the certainty is from the threshold or the exact probabilities for each answer. If the predicting system itself is explainable, we can possibly even get an answer about which characteristics of the input speak for class A and which for class B and thus concoct an explanation for the abstention ourselves (in the sense of "the system thinks the head region of the dog looks like a Husky, but the tail looks like a Retriever, hence it abstains"). This could then be seen as an indirect explanation (via the reasons or explanations of the different classes). However, the system itself cannot provide a straightforward, informative reason for the abstention. Hence, also in terms of explaining the abstaining output, merged systems surpass attached systems, offering more advanced possibilities for providing explanations.

6.4 Conclusion

This chapter explored the realm of suspension within data-based, machine learning AI systems, with a particular focus on abstaining machine learning (AML) systems. AML systems stand out as the closest approximation to what might be termed "suspending AI" in this domain. AML systems introduce a novel approach for responding to questions (or tasks like classifying) by refraining from selecting one of the defined answers, essentially opting out. This unique feature enables them to communicate uncertain situations effectively and allows to bring a human in the loop when stakes are too high to allow for decisions that are prone to error.

[32] As presented here in the post-algorithmic attachment form for ambiguity abstention. Pre-algorithmic attachment can be neglected as this is mostly possible for outlier abstention.

The objectives of this chapter were manifold. Firstly, it aims to shed light on this type of ML systems that has thus far received limited attention, both within the computer science community and the philosophical community. Secondly, it strives to offer an accessible and informative characterization of these systems. Thirdly, it aligns with the overarching theme of this thesis: exploring the various forms and norms of suspension within different AI systems, here in particular AML systems. Lastly, the chapter extends beyond the identification of norms and forms of suspension in machine learning, by pioneering the first philosophical analysis of abstaining machine learning. The inquiry delves into essential questions in the philosophy of AI, especially concerning autonomy and explainability. AML systems have not yet been considered in these discussions. Thereby, this chapter provided the first philosophical analysis of abstaining machine learning.

I have presented and categorized the different AML systems along two dimensions. I distinguished different reasons to abstain and different ways to abstain. I used these distinctions to evaluate the systems based on philosophical demands. It was shown that the different reasons to abstain in ambiguity and outlier abstention find correspondence in different philosophical norms regarding suspension. I have also examined the technical implementation of AML systems, distinguishing between attached and merged systems. I showed that merged systems generally meet the requirements for suspension that I explained in Chapter 2 on suspension and that different versions of suspension correspond to different implementations of learned abstention (labeled and unlabeled). I have shown that in artificial systems there is both a possibility to implement a type of abstention that is structurally similar to the other responses and a possibility to implement abstention with a conceptually more sophisticated special role. This is of particular interest from a philosophical perspective since the explanations I provided for qualified suspension (Subsection 2.3.3), especially the meta-cognitive accounts, characterize suspension by its sophisticated, distinctive role and its deviation from belief and disbelief.

I have also shown that merged systems exhibit a higher level of autonomy and that these systems have more room for different opportunities to explain the abstention responses. As a result, this philosophical analysis provides

compelling reasons for computer scientists to favor the development of such systems.

However, the findings presented here mark just the initial stage of the philosophical analysis of abstaining machine learning. The two aspects of autonomy and explainability should be further explored, and additional topics, e.g., on consciousness and cognition or understanding, warrant investigation. Even in the context of autonomy and explainability, it would be interesting to study the relationship with AML from a different perspective. While this study primarily examined how explainable and autonomous abstaining outputs are, one could also investigate the extent to which the mere capacity to abstain already yields a more autonomous or explainable machine. I am confident that the trust in artificial intelligence is strengthened when these systems acknowledge their own uncertainty and effectively communicate it.

6.4.1 Answers to the Research Questions

1. Does the considered framework allow for a way to deal with conflicting or uncertain information?

   Abstaining machine learning systems have a distinct approach to handling uncertain information. Unlike most machine learning systems, which eliminate uncertainties in the form of transferring soft probabilities into hard probabilities, AML systems can abstain from providing a (defined) answer when uncertain. This abstention communicates that making a decision for a particular input would entail a high degree of uncertainty.

2. Is there something in the light of suspension of judgment present in the framework?

   While there are other attempts within machine learning to communicate uncertainties, such as providing probabilities for different potential outputs, AML systems distinguish themselves by their ability to outright abstain from giving a defined answer. This makes them the most promising candidates within the realm

of machine learning for incorporating something akin to suspension of judgment. Furthermore, I have demonstrated that the abstaining behavior in the identified subclass of AML systems, which I referred to as merged systems, aligns better with the requirements of suspension of judgment compared to the subclass that I called attached systems. In merged systems, abstention explicitly addresses the question under discussion, while in attached systems, abstaining seems to address a separate question.

3. Can we find and distinguish different forms and epistemological norms of doxastic neutrality in the framework?

   The doxastic neutrality represented within merged systems closely aligns with the criteria for qualified suspension. Within merged systems, the differentiation between labeled and unlabeled abstention systems corresponds to the difference between indeterminacy and epistemic suspension. Additionally, the common reasons for abstention, namely outliers and ambiguity, match the differentiation between positive and privative justification for suspension.

Chapter 7

Conclusion

Research Aim

This dissertation investigated the integration of suspension of judgment into artificial intelligence, a topic positioned at the intersection of epistemology, philosophy of mind, and AI. The aim of this work was to explore the different possibilities of incorporating suspension in AI as a tool to uncover uncertainties. The primary objective was to demonstrate how various AI systems already possess or could potentially be adjusted to incorporate the capability to navigate uncertain situations by remaining neutral and delivering responses such as "I don't know."

I examined the potential of suspension in both logic-based and data-based AI systems. Consequently, this thesis made significant contributions to three distinct research domains: (1) philosophical investigations on suspension of judgment, (2) research on logic-based AI systems, and (3) research on data-based AI systems.

For each of these research areas, I will now summarize how the thesis addressed the area, outline the key findings and results, and illustrate the contributions of my work to each respective domain.

For the research areas of logic-based AI and data-based AI, I will present the key results and findings by delineating how the subsequent research questions of this thesis were answered for the various AI frameworks:

1. Does the considered framework allow for a way to deal with conflicting or uncertain information?
2. Is there something in the light of suspension of judgment present in the framework?
3. Can we find and distinguish different forms and epistemological norms of doxastic neutrality in the framework?

Summary, Key Findings, and Contributions

Philosophical Investigations on Suspension of Judgment

Chapter 2 established the philosophical background of this thesis, analyzing the concept of suspension of judgment. This chapter was structured into three main sections: first, epistemological considerations regarding

the normative aspects of suspension (Section 2.2), second, descriptive considerations on the nature of suspension (Section 2.3, and third, overlapping considerations bridging both domains (Section 2.4).

For the epistemological standpoint, several key insights emerged. Firstly, a notable distinction was observed between positive and privative justification for suspension. This distinction highlights the complex nature of suspension, in contrast to belief and disbelief, which only allow for positive justification. Secondly, an analysis of the logic of suspension revealed that the logic of suspension is more intricate than the logic of its counterparts, belief, and disbelief. Logical rules, such as closure under conjunction, and frameworks aiming to represent graded beliefs, like Bayesianism, fall short in capturing suspension's complexity.

The key findings from Section 2.3 on the nature of suspension were the following. Firstly, it became evident that there exists a notable lack of consensus regarding the categorization and terminology surrounding phenomena related to doxastic neutrality within this domain. Secondly, this thesis emphasized a categorical distinction between the phenomena encapsulated by the term "suspension" and other forms of doxastic neutrality, such as mere non-belief and (deep) ignorance. In addition, a gradual scale for embedding suspension within a spectrum of doxastic neutrality was provided, situating it at the far end, opposite to (deep) ignorance. Thirdly, particular emphasis was placed on the meta-cognitive account for describing qualified suspension, which serves as a pivotal reference point for subsequent investigations in artificial intelligence, especially for those pertaining to default logic.

In Section 2.4, the discussion shifted to overlapping considerations, where I emphasized the significance of differentiating between epistemic and indeterminacy suspension. These two forms of qualified suspension are primarily distinguished by their normative profile. Additionally, I explored suspension's connection with inquiry. Given the focus of my investigations on utilizing suspension as a *response to queries* within AI systems, it is appropriate to interpret suspension as a means of providing a response to or concluding an inquiry, rather than initiating one.

This research project makes important contributions for the research

area of suspension. Firstly, the thesis provided a comprehensive philosophical investigation into the multifaceted debates surrounding suspension of judgment, drawing insights from epistemology, philosophy of mind, and their intersections. While this inquiry is broad in scope, it is tailored towards practical and formal applicability, particularly for AI systems. This approach yielded new perspectives on suspension, for instance, the potential illustration of suspension as both categorically distinct from other forms of doxastic neutrality and as part of a scale of neutrality.

Secondly, my investigations shed light on which perspectives on suspension are technically feasible to implement. For instance, my research demonstrates that the differentiation between ignorance and suspension is often representable in AI systems. Additionally, the perspective that suspension is fundamentally a meta-cognitive attitude can be integrated into some formal systems. Moreover, my findings indicate that in the field of AI, suspension can function as an authentic tool for responding to queries, countering the increasingly popular notion that suspension should initiate rather than conclude an inquiry. Lastly, novel insights into the logic of suspension have emerged, particularly through its examination within default logic.

Logic-based AI

Chapters 3, 4, and 5 all dealt with suspension in logic-based systems. I investigated the phenomenon of floating conclusions (Chapter 3), the non-monotonic framework of default logic (Chapter 4), and the non-monotonic framework of argumentation theory (Chapter 5). Chapter 4 and 5 delved into two specific frameworks of non-monotonic reasoning, which serves as a genuine basis of logic-based AI. While in default logic, the main objects of investigation are propositions, argumentation theory allowed to explore suspension regarding an argument, too. The case study on floating conclusions in Chapter 3 is independent of a specific framework and aimed to provide an initial understanding of the role of suspension in non-monotonic reasoning systems.

My investigations on floating conclusions revealed that floating conclusions represent a prototypical case of continuing reasoning with suspended

propositions, as they follow from two conflicting propositions. I presented a collection of different examples of floating conclusions and hypotheses to explain why some floating conclusions seem acceptable while others are deemed to be suspended. I showed that no single hypothesis manages to describe the different intuitions. I presented also constructive results, providing my own framework for evaluating the acceptability of floating conclusions, which is based on the idea of testing different reasons for suspending about the floating conclusions. A key result is that no uniform treatment of all floating conclusions can manage to describe our intuitions appropriately. The acceptance of floating conclusions depends not purely on the logical circumstances, but on the context.

The answers to the research questions about the possibilities of suspension in the different logic-based frameworks can be summarized like this: Upon examining the current capabilities of these logic-based systems to accommodate suspension as a strategy for managing conflicting information, I found that argumentation theory offers substantial possibilities for incorporating various forms of suspension. Argumentation theory already allows for diverse forms and norms of doxastic neutrality through the labeling of statements as undecided, through merging these labels, and through employing non-standard approaches. Default logic, in contrast, does not involve any possibility for neutrality at all. This becomes even more apparent when translating default logic into argumentation theory. Default logic does not leverage the possibilities for indecision from argumentation theory, being purely binary in nature and lacking any space for neutrality. I argued that this is particularly worrisome in conflicting situations, where existing conflict resolution solutions are often unsatisfactory.

Motivated by this lack, additional constructive results for default logic were generated. I adapted the default logic system to accommodate both suspension and ignorance. In this adapted framework, four second-order attitudes are included to distinguish the different epistemic status of the considered propositions. I provided a plurality of results concerning the logic of this framework, yielding insights into the logic of suspension. For instance, I showed that inferences from suspended propositions need not be suspended themselves; they are, at the very last, believable. I also showed how the adapted framework can be interpreted in a deontic way, too.

The contributions of this thesis to the field of logic-based AI are manifold. In general, this thesis is the first to focus specifically on suspension within logic-based AI and to emphasize the intrinsic value of neutrality. In the chapter on floating conclusions, my work offers a structural investigation of the topic, considering all examples and hypotheses from the literature. Additionally, it provides a constructive proposal for a framework that allows to systematically evaluate different examples of floating conclusions. Another significant contribution lies in establishing the close connection between floating conclusions and suspension.

In the realm of argumentation theory, this research is pioneering in demonstrating important parallels between the various uses of indecision in argumentation theory and philosophical insights on suspension. This offers a fresh perspective on indecision within argumentation theory and allows for philosophically informed considerations about refining existing approaches even further.

Within the framework of default logic, the primary contribution is of a constructive nature. My detailed creation of an adapted version of default logic significantly enhances the conflict resolution capabilities of this framework. In my solution, relevant information about conflicts and evidence is stored. The framework stands out through the enrichment of the consequences of the default theory. Through the introduction of second-order propositions, my framework allows for higher-order reasoning, a powerful category of reasoning that can subsume many reasoning processes. Notably, suspension is representable in this framework in a way that aligns well with the philosophical meta-cognitive accounts for suspension. Furthermore, a clear distinction between ignored and suspended propositions can be made. The thesis also provides nuanced observations on the logic of suspension and other second-order attitudes within this framework.

In the overarching view, this work contributes also to the unification of various non-monotonic reasoning frameworks, demonstrating options to transfer the findings about suspension between default logic and argumentation theory.

Data-based AI

In Chapter 6, the focus shifted to data-based AI, specifically abstaining machine learning (AML) systems. In the first part, covered in Section 6.2, abstaining machine learning systems were introduced by explaining their divergence from conventional machine learning classifiers. The various AML systems were categorized along two dimensions: the reasons behind abstention and the implementation of abstention. In Section 6.3, I investigated how suspension manifests within AML systems and considered questions about the explainability and autonomy of the abstaining output.

I emphasized the importance of investigation AML systems, as they introduce a unique approach to uncovering uncertainties. Unlike the various methods available in ML for communicating uncertainties through gradual, numerical values, AML systems offer the option to reject providing an answer, and hence, deliver a categorical abstaining output. Consequently, these systems emerge as the most suitable candidate for incorporating suspension into ML systems. The main result from the categorization of different AML systems is the distinction between attached and merged AML systems. Moreover, the distinction between outlier and ambiguity abstention characterized the different reasons for abstention well.

My research questions were addressed in the philosophical analysis of AML systems. Here, I demonstrated how my differentiation between attached and merged systems can be applied to the philosophical evaluations of AML systems. I showed that merged systems are equipped to approach the qualified concept of suspension. Moreover, I showed that the outlier, ambiguity distinction allows for the distinction between different norms of suspension. The chapter also emphasized the potential of merged systems in delivering more autonomous and explainable outputs.

In the research field of philosophy of data-based machine learning, to my knowledge, my research was the first to focus on suspension and abstaining machine learning systems. By providing a comprehensive explanation of their technical implementation, tailored for a broader philosophical audience, I established the groundwork for future philosophical explorations in this domain. I organized the diverse AML systems into coherent categories and demonstrated that my categorization proves valuable for the

philosophical evaluation of AML systems.

Furthermore, in this work, I pioneered the first *philosophical analysis* of abstaining machine learning. Through this analysis, I showed how the different AML systems address philosophical demands for suspension differently well, offering insights that can inform their further development. By examining suspension within AML systems, my research contributes to our understanding of how ML systems can incorporate epistemic practices per se. Additionally, I expanded the discourse on autonomy and explainability in the philosophy of AI to encompass abstaining outputs, which contributes to the ongoing discourse surrounding these crucial topics.

Besides the significant contributions made in the various research fields, the scope of this thesis necessitated some limitations, which present several opportunities for further research, as I will discuss in the following.

Limitations and Outlook

Concerning the notion of suspension employed, one limitation of this work is the one-sided interpretation of suspension, which is exclusively viewed as a means of responding to queries, and thus concluding the inquiry. An alternative perspective, advocated notably by Friedman (2017), posits suspension as a mechanism for maintaining an open mind at the start of each inquiry. Exploring this interpretation of suspension in AI systems could offer valuable insights, particularly in understanding how biases in model choices and prior probabilities might hinder genuine open-mindedness during the initial stages of inquiry. Furthermore, investigating inquiry and zetetic reasoning themselves within AI systems seems promising for further exploration.

This observation is closely related to another constraint observed in my adaptation of default logic. As a static, logical system, my framework treats all floating conclusions uniformly, without distinguishing between different examples. As the incorporation of my second-order propositions is based on the synthesis of various extensions rather than scenarios, floating conclusions are accepted with certainty in my framework. This outcome conflicts with the demanded treatment of floating conclusions advocated in Chapter 3. Future investigations aimed at facilitating dynamic, zetetic

reasoning within these systems could mitigate such uniform treatment and lead to more appropriate outcomes for such cases.

Regarding the philosophical analysis of abstaining machine learning, my work just represents the initial phase of the possible philosophical inquiry into this area. Further investigation into autonomy, explainability, and related topics is crucial to comprehensively understand the implications of abstaining machine learning systems. Moreover, my work on suspension in data-based AI is primarily limited towards classifiers, which are designed to address specific tasks or questions. As the importance of more general artificial intelligences continues to grow, exploring the feasibility of incorporating suspension into such systems emerges as a critical next research step.

From my perspective, there are two specific groups of AI systems, which seem especially worth for further investigation into suspension within AI.

Hybrid systems: These AI systems integrate both data-based and logic-based approaches, aiming to leverage the advantages of each paradigm. By exploring suspension mechanisms within hybrid systems, we can delve into how suspension interacts within the context of logical reasoning structures and statistical patterns derived from data simultaneously. Although I have demonstrated possibilities for incorporating suspension within the primary subfields of hybrid systems — logic-based and data-based systems — it can be beneficial to examine hybrid systems as such. In hybrid systems, the coexistence of logical reasoning structures and statistical patterns allows for both logical conflicts and uncertainties in the data, representing different sources of error.

Large Language Models: Given the immense potential of recently introduced generative models, it almost seems like a matter of social responsibility to prioritize these models for future philosophical inquiries on AI. Especially, large language models (LLMs) like GPT[1] of OpenAI, and related LLMs such as Gemini[2] of Google DeepMind and Claude[3] of Anthropic, but also rather recent European LLMs like the French

[1] https://chat.openai.com,
[2] https://gemini.google.com
[3] https://claude.ai/

Mistral[4] of MistralAI and the German Luminous[5] of Aleph Alpha, demonstrate remarkable proficiency in various tasks. LLMs offer an ideal framework for studying suspension, as they have the potential to explicitly express indecision to the user through linguistic expressions. However, when evaluating their ability to suspend, we must broaden our focus beyond a specific epistemological perspective. While LLMs can respond to straightforward questions like "Who is the president of France?", their core function is generative in nature — they generate text. Consequently, the output generated by LLMs often transcends the specific question-answer format, encompassing a wide range of content creation tasks from poetry to code debugging. When considering the integration of abstention capabilities into LLMs, it is essential to expand the concept of knowledge limits to encompass capabilities limits. The possibilities for investigating how to train LLMs to offer abstention are diverse, ranging from developing abstaining systems from scratch, to fine-tuning and unlearning existing LLMs to enhance suspension in sensitive application domains.

Concluding Remarks

In conclusion, this research contributes to the important objective of empowering AI systems to effectively handle and articulate uncertainty. It not only provides important insights into the academic discourse on suspension of judgment in AI but also paves the way for future explorations that bridge theoretical insights with practical applications. Encouraging AI systems to suspend judgment in unclear or critical situations serves as an effective approach to appropriately uncover and communicate uncertainties that occur in AI contexts. By integrating suspension of judgment into AI systems, their decision-making capabilities are enriched, providing a more nuanced alternative. These investigations have far-reaching implications for numerous socially significant aspects of AI systems, such as:

> **Human-AI Interaction**: In many scenarios, such as collaborative settings or decision support systems, AI systems are expected to interact with human users or experts. In such contexts, the ability to suspend judgment can enhance communication and collaboration.

[4] `https://docs.mistral.ai`
[5] `https://app.aleph-alpha.com`

When AI systems effectively communicate uncertainties, human users or experts can intervene and make more informed decisions.

Learning: Incorporating mechanisms for suspension of judgment can also facilitate learning in AI systems. It serves as a starting point for enabling AI systems to adapt and actively learn from their experiences in uncertain environments. By recognizing when they lack sufficient information to make confident decisions, retraining and fine-tuning in the identified areas of uncertainty can be prioritized.

Explainability, Transparency, and Trust: Instead of offering potentially misleading or overly confident predictions, when employing suspension, the system acknowledges its limitations and refrains from making unfounded assertions. When an AI system suspends judgment, it encourages users to engage in a dialogue about the decision context and the uncertainties involved, fostering a deeper and more transparent understanding of the decision-making process for the user. This transparency helps users grasp the level of uncertainty associated with the AI system's decisions. As noted by Phillips et al. (2020), meaningful explanations for an AI system's output require the system to recognize its own knowledge limits. Suspending judgment provides the most straightforward way to uncover these limits. This is also demonstrated by a discussion on different LLMs, where Aleph Alpha was considered more transparent, revealing not only the references of its information, but also potential conflicts of information (Bomke and Holzki, 2024). AI systems capable of suspending judgment and communicating their uncertainty offer hence more transparent explanations for their decisions. This enhances trust and understanding of AI systems, particularly in critical applications where decision justification is necessary.

The implications on these crucial aspects of AI research once again underscore the importance of this work and the need for further research in this field. As we stand on the precipice of a new era in AI development, the nuanced understanding of suspension of judgment and its implications for AI systems will undoubtedly play a critical role in shaping the trajectory of responsible AI research and its application across diverse domains.

Notation

The following list provides an overview of the major symbols used in the different chapters of this thesis. Symbols that are mathematical conventions are not necessarily included here:

Default Logic (and Floating Conclusions)

Δ	default theory
$\mathcal{W}$	set of world descriptions or knowledge base of a default theory
$\mathcal{D}$	set of defaults of a default theory
δ	a default in D
$\mathcal{S}$	scenario of a default theory
$a, p, s_1, s_2 \ldots$	propositions (represented by small letters)
$prem(\delta)$	function that picks out the premise of a default
$con(\delta)$	function that picks out the conclusion of a default
$out(d_i)$	statement that a default δ_i is undercut
$\mathcal{E}^{\mathrm{D}}$	extension of a default theory
Γ	the consequences of a default theory
$\mathcal{C}(p)$	second-order attitude that represents certainty in p
$\mathcal{B}(p)$	second-order attitude that represents believability in p
$\mathcal{S}(p)$	second-order attitude that represents suspension about p
$\mathcal{I}(p)$	second-order attitude that represents ignorance about p

Argumentation Theory

Ar	set of arguments of an argumentation framework
att	set of attack relation of an argumentation framework
$A, B, C, \ldots$	arguments in set Ar of an argumentation framework (represented by capital letters)
$\mathcal{E}^{\mathrm{A}}$	extension of a argumentation framework
`in`, `out`, `undec`	three labels of a standard argumentation framework
Lab	labeling of an argumentation framework

$Lab(\texttt{in})$	set of arguments that are labeled `in` according to Lab
$Lab(\texttt{out})$	set of arguments that are labeled `out` according to Lab
$Lab(\texttt{undec})$	set of arguments that are labeled `undec` according to Lab

Machine Learning

$\mathbb{R}^2$	two-dimensional real space $\mathbb{R} \times \mathbb{R}$
X	input space, in our main example equal to $\mathbb{R}^2$
$\boldsymbol{x}$	vector variable in X
Y	output set, in our main example equal to $\{\texttt{malignant}, \texttt{benign}\}$
y	value in Y
T	set of training data points, subset of $X \times Y$
n	number of training data points in T
m	number of possible labels/outputs in Y
$\langle \boldsymbol{x}^{(i)}, y^{(i)} \rangle$	i-th (of n) training data point in T
$\boldsymbol{x}^{(i)}$	(two-dimensional) input vector of i-th training data point
$x_1^{(i)}$	first entry of (two-dimensional) vector $\boldsymbol{x}^{(i)}$
$x_2^{(i)}$	second entry of (two-dimensional) vector $\boldsymbol{x}^{(i)}$
$y^{(i)}$	output value and (ground-truth) label of i-th training data point
f	candidate functions for predictors
$f(\boldsymbol{x}^{(i)})$	predicted output or label for i-th training data point by f
$f(\boldsymbol{x})$	predicted output or label for variable $\boldsymbol{x}$ by f
$\mathcal{F}$	set of candidate functions f from X to Y
$\hat{f}$	optimal regular predictor from X to Y
l	regular loss function from $Y \times Y$ to $\{0, 1\}$
r	rejector in an attached abstaining ML system
t	threshold for maximum uncertainty or distance measure
Y^*	output set of the abstention-labeled training data, in our main example equal to $\{\texttt{malignant}, \texttt{benign}, \texttt{abstention}\}$

T^*	set of abstention-labeled training data points, subset of $X \times Y^*$
$\mathcal{F}^*$	set of candidate functions f from X to Y^*
$\bar{f}$	optimal abstention predictor from X to Y^*
l^*	abstention-allowing loss function from $Y \times Y^*$ to $\{0, 1, \alpha\}$
α	loss for abstaining in abstention-allowing loss function

Bibliography

Anderson, M. and Anderson, S. L. (2011). *Machine Ethics*. Cambridge University Press.

Archer, A. (2018). Wondering about what you know. *Analysis*, 78(4):596–604.

Arieli, O. (2016). On the acceptance of loops in argumentation frameworks. *Journal of Logic and Computation*, 26(4):1203–1234.

Armendt, B. (2010). Stakes and Beliefs. *Philosophical Studies*, 147:71–87.

Artelt, A. and Hammer, B. (2022). "Even if..."–Diverse Semifactual Explanations of Reject. *arXiv preprint arXiv:2207.01898*. `https://doi.org/10.48550/arXiv.2207.01898`.

Artelt, A., Visser, R., and Hammer, B. (2022). Model Agnostic Local Explanations of Reject. *arXiv preprint arXiv:2205.07623*. `https://doi.org/10.48550/arXiv.2205.07623`.

Asif, A. et al. (2020). Generalized Neural Framework for Learning with Rejection. In Roy, A., editor, *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Barnes, E. A. and Barnes, R. J. (2021). Controlled Abstention Neural Networks for Identifying Skillful Predictions for Regression Problems. *Journal of Advances in Modeling Earth Systems*, 13(12):e2021MS002575.

Baroni, P., Caminada, M., and Giacomin, M. (2011). An introduction to argumentation semantics. *Knowledge Engineering Review*, 26(4):365.

Baroni, P., Giacomin, M., and Guida, G. (2004). Towards a Formalization of Skepticism in Extension-based Argumentation Semantics. In Grasso, F., Reed, C., and Carenini, G., editors, *Proceedings of the 4th Workshop on Computational Models of Natural Argument*, pages 47–52. ECAI.

Baroni, P., Giacomin, M., and Guida, G. (2005). SCC-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*,

168(1-2):162–210.

Baroni, P., Giacomin, M., and Liao, B. (2015). I don't care, I don't know... I know too much! On Incompleteness and Undecidedness in Abstract Argumentation. In *Advances in Knowledge Representation, Logic Programming, and Abstract Argumentation*, pages 265–280. Springer.

Baroni, P., Governatori, G., Lam, H.-P., and Riveret, R. (2016). On the Justification of Statements in Argumentation-based Reasoning. In Baral, C., Delgrande, J., and Wolter, F., editors, *Proceedings, Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 521–524. Citeseer.

Baroni, P. and Riveret, R. (2019). Enhancing Statement Evaluation in Argumentation via Multi-labelling Systems. *Journal of Artificial Intelligence Research*, 66:793–860.

BBC-News (2017). Google's 'superhuman' DeepMind AI claims chess crown. `https://www.bbc.com/news/technology-42251535`, 2023-12-11.

Beirlaen, M., Heyninck, J., Pardo, P., and Straßer, C. (2018). Argument strengh in formal argumentation. *IfCoLog Journal of Logics and their Applications*, 5(3):629–675.

Belnap, N. (1977). How a computer should think. In Ryle, G., editor, *Contemporary Aspects of Philosophy*, pages 30–55. Oriel Press.

Bengio, Y. and Marcus, G. (2021). AI DEBATE! Yoshua Bengio vs Gary Marcus. Youtube, `https://www.youtube.com/watch?v=8NNDCuSgls0`, 2023-12-22.

Bergmann, M. (2005). Defeaters and higher-level requirements. *The Philosophical Quarterly*, 55(220):419–436.

Besold, T. R., D'Avila Garcez, A. S., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kphnberger, K.-U., , Lamb, L. C., Lowd, D., Vieira Lima, P. M., de Penning, L., Pinkas, G., Poon, H., and Zaverucha, G. (2022). Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342(1):327.

Bomke, L. and Holzki, L. (2024). Mistral gegen Aleph Alpha. Wer gewinnt das KI-Rennen in Europa? *Handelsblatt*, `https://www.handelsblatt.com/technik/ki/mistral-gegen-aleph-alpha-wer-gewinnt-das-ki-rennen-in-europa/100021767.html`, 2024-03-25.

Bradshaw, J. M., Hoffman, R. R., Woods, D. D., and Johnson, M. (2013). The Seven Deadly Myths of "Autonomous Systems"'. *IEEE Intelligent*

Systems, 28(3):54–61.

Briegel, H. J. and Müller, T. (2015). A Chance for Attributable Agency. *Minds and Machines*, 25:261–279.

Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., and Cabitza, F. (2020). Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *Journal of Medical Systems*, 44:1–12.

Bringsjord, S. and Govindarajulu, N. S. (2022). Artificial Intelligence. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition. `https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/`, 2023-12-22.

Broersen, J. (2017). Rethinking the BOID with an eye on making it more responsible. *Unpublished Manuscript*.

Bundesministerium für Wirtschaft und Klimaschutz (2019). Wie Künstliche Intelligenz die Energiewende voranbringt. *Energiewende direkt*, `https://www.bmwi-energiewende.de/EWD/Redaktion/Newsletter/2019/10/newsletter_2019-10.html?__act=renderPdf&__iDocId=1483930`, 2023-01-26.

Burkart, N. and Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Caie, M. (2012). Belief and Indeterminacy. *Philosophical Review*, 121(1):1–54.

Campagner, A., Cabitza, F., and Ciucci, D. (2019). Three–Way Classification: Ambiguity and Abstention in Machine Learning. In Mihálydeák, T., Min, F., Wang, G., Banerjee, M., Düntsch, I., Suraj, Z., and Ciucci, D., editors, *Rough Sets: International Joint Conference, IJCRS 2019, Debrecen, Hungary, June 17–21, 2019, Proceedings*, pages 280–294. Springer.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46.

Christensen, D. (2010). Higher-Order Evidence. *Philosophy and Phenomenological Research*, 81(1):185–215.

Coenen, L., Abdullah, A. K., and Guns, T. (2020). Probability of default estimation, with a reject option. In Webb, G., Zhang, Z., Tseng,

V. S., Williams, Graham Vlachos, M., and Cao, L., editors, *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 439–448. IEEE.

Cohen, P. J. (1966). *Set theory and the continuum hypothesis*. Courier Corporation, 2008 Edition.

Conee, E. and Feldman, R. (2004). *Evidentialism: Essays in Epistemology*. Clarendon Press.

Crawford, L. (2022). Suspending Judgment is Something You Do. *Episteme*, 19(4):561–577.

Crawford, S. (2004). A solution for Russellians to a puzzle about belief. *Analysis*, 64(3):223–229.

De Stefano, C., Sansone, C., and Vento, M. (2000). To reject or not to reject: That is the question - an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(1):84–94.

Delgrande, J. P. and Schaub, T. (2000). The Role of Default Logic in Knowledge Representation. In Minker, J., editor, *Logic-based artificial intelligence*, pages 107–126. Springer.

Denoeux, T. (1995). A k-Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813.

Descartes, R. (1641). *Meditationes de Prima Philosophia*, volume VII of *Oeuvres de Descartes*. Paris. Edition from 1996.

Dubuisson, B. and Masson, M. (1993). A statistical decision rule with incomplete knowledge about classes. *Pattern recognition*, 26(1):155–165.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

D'Avila Garcez, A. and Lamb, L. C. (2023). Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406.

D'Avila Garcez, A., Lamb, L. C., and Gabbay, D. M. (2009). *Neural-Symbolic Cognitive Reasoning*. Springer.

European Parliament (2023). Artificial intelligence act. `https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf`, 2024-02-21.

Eva, B., Ried, K., Müller, T., and Briegel, H. J. (2022). How a Minimal

Learning Agent can Infer the Existence of Unobserved Variables in a Complex Environment. *Minds and Machines*, 33(1):185–219.

Fantl, J. and McGrath, M. (2002). Evidence, Pragmatics, and Justification. *The Philosophical Review*, 111(1):67–94.

Feldman, R. and Conee, E. (1985). Evidentialism. *Philosophical Studies*, 48(1):15–34.

Feldman, R. and Conee, E. (2018). Between Belief and Disbelief. In McCain, K., editor, *Believing in Accordance with the Evidence: New Essays on Evidentialism*. Springer.

Ferrari, F. and Incurvati, L. (2022). The Varieties of Agnosticism. *The Philosophical Quarterly*, 72(2):365–380.

Ferri, C. and Hernández-Orallo, J. (2004). Cautious Classifiers. *ROCAI*, 4:27–36.

Fischer, T. (2023). Strafgesetzbuch mit Nebengesetzen, Auflage 70. *C.H.Beck, München*.

Foley, R. (1992). The Epistemology of Belief and the Epistemology of Degrees of Belief. *American Philosophical Quarterly*, 29(2):111–124.

Friedman, J. (2013a). Question-directed attitudes. *Philosophical Perspectives*, 27(1):145–174.

Friedman, J. (2013b). Rational Agnosticism and Degrees of Belief. In Gendler, T. S. and Hawthorne, J., editors, *Oxford Studies in Epistmology*, volume 4, pages 45–81. Oxford University Press.

Friedman, J. (2013c). Suspended judgment. *Philosophical Studies*, 162(2):165–181.

Friedman, J. (2017). Why Suspend Judging? *Noûs*, 51(2):302–326.

Friedman, J. (2019a). Checking again. *Philosophical Issues*, 29(1):84–96.

Friedman, J. (2019b). Inquiry and belief. *Noûs*, 53(2):296–315.

Friedman, J. (2020). The Epistemic and the Zetetic. *Philosophical Review*, 129(4):501–536.

Geifman, Y. and El-Yaniv, R. (2019). SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In Chaudhuri, K. and Salakhutdinov, R., editors, *International conference on machine learning*, pages 2151–2159. PMLR.

Ginsberg, M. L. (1993). *Essentials of Artificial Intelligence*. Morgan Kaufmann Publishers.

Gödel, K. (1947). What is Cantor's Continuum Problem? *The American*

Mathematical Monthly, 54(9):515–525.

Goldman, A. I. and Olsson, E. J. (2009). Reliabilism and the Value of Knowledge. In Haddock, A., Millar, A., and Pritchard, D., editors, *Epistemic Value*, pages 19–41. Oxford University Press.

Hájek, A. (1998). Agnosticism meets Bayesianism. *Analysis*, 58(3):199–206.

Hamid, K., Asif, A., Abbasi, W., Sabih, D., et al. (2017). Machine Learning with Abstention for Automated Liver Disease Diagnosis. In Bajwa, U. I., editor, *2017 International Conference on Frontiers of Information Technology (FIT)*, pages 356–361. IEEE.

Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., and Davis, J. (2021). Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*. `https://doi.org/10.48550/arXiv.2107.11277`.

Hern, A. (2022). AI bot ChatGPT stuns academics with essay-writing skills and usability. *The Guardian*. `https://www.theguardian.com/technology/2022/dec/04/ai-bot-chatgpt-stunsacademics-with-essay-writing-skills-and-usability`, 2022-12-04.

Homenda, W., Luckner, M., and Pedrycz, W. (2014). Classification with rejection based on various SVM techniques. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 3480–3487. IEEE.

Horty, J. F. (1994). Moral dilemmas and nonmonotonic logic. *Journal of philosophical logic*, 23:35–65.

Horty, J. F. (2001). Nonmonotonic Logic. In Goble, L., editor, *The Blackwell Guide to Philosophical Logic*, pages 336–361. Blackwell Publishers.

Horty, J. F. (2002). Skepticism and floating conclusions. *Artificial Intelligence*, 135(1-2):55–72.

Horty, J. F. (2011). *Reasons as Defaults*. Oxford University Press.

Horty, J. F. (2023). Logics for Defeasible Reasoning. *Lecture Notes. University of Maryland*.

Jakobovits, H. and Vermeir, D. (1999). Robust semantics for argumentation frameworks. *Journal of Logic and Computation*, 9(2):215–261.

Johnson, D. G. and Verdicchio, M. (2017). Reframing AI Discourse. *Minds and Machines*, 27(4):575–590.

Kempt, H. and Nagel, S. K. (2022). Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. *Journal of Medical Ethics*, 48(4):222–229.

Kim, B. (2017). Pragmatic encroachment in epistemology. *Philosophy*

Compass, 12(5):e12415.

Knoks, A. (2021). Moral Principles: Hedged, Contributory, Mixed. In Liu, F., Marra, A., Portner, P., and Van De Putte, F., editors, *Deontic Logic and Normative Systems, 15th International Conference, DEON 2020/2021*, pages 272–290. College Publications.

Kobl, K. (2020). Hier zieht Künstliche Intelligenz im Hintergrund die Fäden. *Frauenhofer IKS*, `https://safe-intelligence.fraunhofer.de/artikel/hier-zieht-ki-die-faeden`, 2023-01-26.

Kompa, B., Snoek, J., and Beam, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):1–6.

Le Morvan, P. (2011). On Ignorance: A Reply to Peels. *Philosophia*, 39(2):335–344.

Lewis, D. (1980). A Subjectivist's Guide to Objective Chance. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *IFS: Conditionals, Belief, Decision, Chance and Time*, pages 267–297. Springer Dordrecht.

Linusson, H., Johansson, U., Boström, H., and Löfström, T. (2018). Classification with Reject Option Using Conformal Prediction. In Phung, D., Tseng, V. S., Webb, G. I., Ho, B., Ganji, M., and Rashidi, L., editors, *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part I 22*, pages 94–105. Springer.

Lord, E. (2020). Suspension of Judgment, Rationality's Competition, and the Reach of the Epistemic. In Schmidt, S. and Ernst, G., editors, *The Ethics of Belief and Beyond: Understanding Mental Normativity*. Routledge.

Lord, E. and Sylvan, K. (2021). Suspension, Higher-Order Evidence, and Defeat. In Simion, M. and Brown, J., editors, *Reasons, Justification, and Defeat*. Oxford University Press.

Lotte, F., Mouchere, H., and Lécuyer, A. (2008). Pattern rejection strategies for the design of self-paced EEG-based Brain-Computer Interfaces. In Borgefors, G. and Flynn, P., editors, *2008 19th International Conference on Pattern Recognition*, pages 1–5. IEEE.

Luther, T. (2018). Wenn Künstliche Intelligenz über eine Finanzierung entscheidet. *Handelsblatt*, `https://www.handelsblatt.com/unternehmen/leasing/rating-durch-digitale-kreditpruefer-`

wenn-kuenstliche-intelligenz-ueber-eine-finanzierung-entscheidet/22593710.html, 2023-01-26.

Makinson, D. and Schlechta, K. (1991). Floating conclusions and zombie paths: Two deep difficulties in the "directly skeptical" approach to defeasible inheritance nets. *Artificial intelligence*, 48(2):199–209.

Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *arXiv preprint arXiv:2002.06177*. https://doi.org/10.48550/arXiv.2002.06177.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., and Clahsen, H. (1992). Overregularization in Language Acquisition. *Monographs of the society for research in child development*, 57(4):1–178.

McCarthy, J. (1959). Programs with Common Sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pages 75–91, London. Her Majesty's Stationary Office.

McCarthy, J. (1980). Circumscription - A Form of Nonmonotonic Reasoning. *Artificial intelligence*, 13(1-2):27–39.

McCarthy, J. (1993). History of circumscription. *Artificial Intelligence*, 59(1-2):23–26.

McGrath, M. (2021). Being neutral: Agnosticism, inquiry and the suspension of judgment. *Noûs*, 55(2):463–484.

Minker, J. (2000). Introduction to Logic-Based Artificial Intelligence. In Minker, J., editor, *Logic-based artificial intelligence*, pages 3–27. Springer.

Minsky, M. (1975). A Framework for Representing Knowledge. In Winston, P., editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill.

Modgil, S. and Prakken, H. (2014). The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62.

Mouchère, H. and Anquetil, E. (2006a). A Unified Strategy to Deal with Different Natures of Reject. In Werner, B., editor, *18th International Conference on Pattern Recognition (ICPR 06), Volume 2*, pages 792–795. IEEE.

Mouchère, H. and Anquetil, E. (2006b). Generalization Capacity of Handwritten Outlier Symbols Rejection with Neural Network. In Lorette, G., Bunke, H., and Schomaker, L., editors, *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft.

Mozannar, H. and Sontag, D. (2020). Consistent Estimators for Learning to

Defer to an Expert. In Daumé, H. and Singh, A., editors, *International Conference on Machine Learning*, pages 7076–7087. PMLR.

Muggleton, S. and Marginean, F. (2000). Logic-Based Machine Learning. In Minker, J., editor, *Logic-Based Artificial Intelligence*, pages 315–330. Springer.

Müller, T. and Briegel, H. J. (2018). A Stochastic Process Model for Free Agency under Indeterminism. *dialectica*, 72(2):219–252.

Mullins, R. (2021). Formalizing Reasons, Oughts, and Requirements. *Ergo an Open Access Journal of Philosophy*, 7:568–599.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT press.

OPEN AI (2024). ChatGPT 3.5. `https://chat.openai.com`, 2024-04-10.

Owens, D. (2002). *Reason Without Freedom: The Problem of Epistemic Normativity*. Routledge.

Pardo, P. and Straßer, C. (2022). Modular orders on defaults in formal argumentation. *Journal of Logic and Computation*, 34:exac084.

Peels, R. (2010). What is ignorance? *Philosophia*, 38(1):57–67.

Peels, R. (2012). The New View on Ignorance Undefeated. *Philosophia*, 40(4):741–750.

Phillips, N. D., Neth, H., Woike, J. K., and Gaissmaier, W. (2017). FFTrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. *Judgment and Decision Making*, 12(4):344–368.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., and Przybocki, M. A. (2020). Four principles of explainable artificial. Technical report, NIST interagency report; NIST internal report; 8312. Commerce Department, National Institute of Standards and Technology.

Pollock, J. L. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press.

Pollock, J. L. (2001). Defeasible reasoning with variable degrees of justification. *Artificial intelligence*, 133(1-2):233–282.

Prakken, H. (2002). Intuitions and the Modelling of Defeasible Reasoning: some Case Studies. In Benferhat, S. and Giunchiglia, E., editors, *Proceedings of the 9th International Workshop on Non-Monotonic Reasoning. Toulouse, France*, pages 91–102.

Putnam, H. (1973). Meaning and Reference. *The Journal of Philosophy*, 70(19):699–711.

Raleigh, T. (2021). Suspending is believing. *Synthese*, 198(3):2449–2474.

Ramaswamy, H. G., Tewari, A., and Agarwal, S. (2018). Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530 – 554.

Reiter, R. (1980). A Logic for Default Reasoning. *Artificial intelligence*, 13(1-2):81–132.

Roberts, C. (1996). Information structure in discourse: Toward a unified theory of formal pragmatics. *Ohio State University Working Papers in Linguistics*, 49:91–136.

Rosa, L. (2019). Logical Principles of Agnosticism. *Erkenntnis*, 84(6):1263–1283.

Rosa, L. (2021). Rational requirements for suspended judgment. *Philosophical Studies*, 178(2):385–406.

Rosenkranz, S. (2007). Agnosticism as a Third Stance. *Mind*, 116(461):55–104.

Rosenkranz, S. (2018). The Structure of Justification. *Mind*, 127(506):309–338.

Russell, S. and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, 4th edition.

Sarker, K., Yang, X., Li, Y., Belkasim, S., and Ji, S. (2020). A Unified Plug-and-Play Framework for Effective Data Denoising and Robust Abstention. *arXiv preprint arXiv:2009.12027*. `https://doi.org/10.48550/arXiv.2009.12027`.

Schroeder, M. (2012). Stakes, withholding, and pragmatic encroachment on knowledge. *Philosophical Studies*, 160(2):265–285.

Schuster, D. (2021). Forms and Norms of Indecision in Argumentation Theory. In Liu, F., Marra, A., Portner, P., and Van De Putte, F., editors, *Deontic Logic and Normative Systems, 15th International Conference, DEON 2020/2021*, pages 394–413. College Publications.

Schuster, D. (2023). The fixed points of belief and knowledge. *Logic Journal of the IGPL*, page jzad016.

Schuster, D., Broersen, J., and Prakken, H. (2023). On Floating Conclusions. In Maranhão, J., Peterson, C., Straßer, C., and van der Torre, L., editors, *Deontic Logic and Normative Systems, 16th International Conference, DEON 2023*, pages 199–215. College Publications.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and

Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Mortensen, E., editor, *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144.

Singh, S. and Markou, M. (2004). An Approach to Novelty Detection Applied to the Classification of Image Regions. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):396–407.

Straßer, C. and Antonelli, G. A. (2019). Non-monotonic Logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2019 edition. `https://plato.stanford.edu/entries/logic-nonmonotonic/`, 2023-12-22.

Sturgeon, S. (2010). Confidence and Coarse-Grained Attitudes. *Oxford Studies in Epistemology*, 3:126–149.

Sun, R. (1996). Hybrid Connectionist-Symbolic Modules: A Report from the IJCAI-95 Workshop on Connectionist-Symbolic Integration. *AI Magazine*, 17(2):99–99.

Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., and Mohd-Yusof, J. (2019a). Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*. `https://doi.org/10.48550/arXiv.1905.10964`.

Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., and Mohd-Yusof, J. (2019b). Knows When it Doesn't Know: Deep Abstaining Classifiers. *preprint*. `https://openreview.net/forum?id=rJxF73R9tX`.

Toni, F. (2014). A tutorial on assumption-based argumentation. *Argument & Computation*, 5(1):89–117.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, LIX(236):433–460.

van Fraassen, B. C. (1998). The Agnostic Subtly Probabilified. *Analysis*, 58(3):212–220.

Veltman, F. (1985). *Logics for Conditionals*. PhD thesis, University of Amsterdam.

Verena, W. (2019). Doxastischer Voluntarismus und epistemisches Handeln. In Grajner, M. and Melchior, G., editors, *Handbuch Erkenntnistheorie*,

pages 218–224. Metzler.

Wagner, V. (2021). Epistemic dilemma and epistemic conflict. In Stapleford, S. and McCain, K., editors, *Epistemic Dilemmas: New Arguments, New Angles*, pages 58–76. Routledge.

Wagner, V. (2022). Agnosticism as settled indecision. *Philosophical Studies*, 179(2):671–697.

Wagner, V. (2023a). Bracketing and inquiry-opening suspension. *Manuscript*.

Wagner, V. (2023b). Intentionally Off Topic: A Theory of Fake Answers. *Manuscript*.

Wedgwood, R. (2002). The Aim of Belief. *Philosophical Perspectives*, 16:267–297.

Wegkamp, M. and Yuan, M. (2011). Support vector machines with a reject option. *Bernoulli*, 17(4):1368–1385.

Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1992). Breast cancer Wisconsin (diagnostic) data set. *UCI Machine Learning Repository*. `http://archive.ics.uci.edu/ml/`.

Wu, Q., Jia, C., and Chen, W. (2007). A Novel Classification-Rejection Sphere SVMs for Multi-class Classification Problems. In Lei, J., Yao, J., and Zhang, Q., editors, *Third International Conference on Natural Computation (ICNC 2007). Volume 1*, pages 34–38. IEEE.

Wu, Y., Caminada, M., and Podlaszewski, M. (2010). A Labelling-Based Justification Status of Arguments. *Studies in Logic*, 3(4):12–29.

Yuan, B., Yue, X., Lv, Y., and Denoeux, T. (2020). Evidential Deep Neural Networks for Uncertain Data Classification. In Li, G., Shen, H. T., Yuan, Y., Wang, X., Liu, H., and Zhao, X., editors, *Knowledge Science, Engineering and Management: 13th International Conference, Hangzhou, China, Proceedings, Part II 13*, pages 427–437. Springer.

Zheng, E.-h., Zou, C., Sun, J., and Chen, L. (2011). Cost-sensitive SVM with Error Cost and Class-dependent Reject Cost. *International Journal of Computer Theory and Engineering*, 3(1):130.

Zinke, A. (2021a). Logic of suspension. *Unpublished lecture slides III-IV. Goethe University Frankfurt*.

Zinke, A. (2021b). Rational Suspension. *Theoria*, 87(5):1050–1066.

Zolfagharian, A. (2020). *Formal Representations of Suspended Judgment*. PhD thesis, University of Konstanz.