Self-Explaining Artificial Intelligence: On the Requirements for Autonomous Explanation Generation

by

Hjörleifur Rörbeck

Thesis of 60 ECTS credits submitted to the School of Technology,
Department of Computer Science
at Reykjavík University in partial fulfillment
of the requirements for the degree of
Master of Science (M.Sc.) in Computer Science

May 2022

Examining Committee:

Dr. Kristinn R. Thórisson, Supervisor
Professor, Reykjavík University, Iceland

Dr. Pei Wang, Co-advisor
Professor, Temple University, USA

Dr. Patrick Hammer, Co-advisor
Postdoctoral Fellow, Stockholm University, Sweden

No Examiner

Self-Explaining Artificial Intelligence: On the Requirements for Autonomous Explanation Generation

Hjörleifur Rörbeck

May 2022

Abstract

Explainability is an important feature of any artificial intelligence (AI) system, for numerous reasons. Firstly, it provides interrogability, allowing any outsider to investigate why the system did what it did, or why some things happened and not others. Secondly, it can guide an autonomous learner in its quest to learn more, as the ability to find explanations for failure or success can help with learning how the world works. And thirdly, it provides transparency, since (useful) explanations can form the basis of (valid) plans, summaries, justifications, predictions, etc. To be trustworthy, especially when applied in critical domains, transparency in all activities is needed. Providing transparency to modern machine learning and AI systems, such as reinforcement learners and deep neural networks, requires considerable post-hoc effort and skill in interpreting algorithms, as they typically do not represent knowledge in a human-readable form. Prior work has focused almost exclusively on interpretation, while purposely avoiding any assessment of causal attribution, even though this is the most important aspect of explanations. In contrast to prior work, the focus here is on self-explanation: The ability of systems to generate explanations of their own behavior, and their task-environment, automatically. We evaluate the potential of two systems, AERA and NARS, for automatically generating explanations about knowledge that they have learned, and compare their mechansims to other approaches to explainable AI, as a step towards scientific evaluation of explainability. AERA and NARS are based on principles that differ radically from mainstream AI methods, relying heavily on principles of reflectivity. The focus here is on two important aspects of explanations in AI: Firstly, what constitutes a useful explanation; secondly, what are the knowledge acquisition and application requirements for generating such explanations. We examine the potential of existing systems for automating the generation of such explanations.

Keywords: Explainability, General Machine Intelligence, Artificial General Intelligence, AGI, Autonomy, causality

Sjálfskýrandi gervigreind: Um kröfur fyrir sjálfvirkar útskýringar

Hjörleifur Rörbeck

May 2022

Útdráttur

Útskýranleiki er mikilvægur þáttur í þróun gervigreindar, af margvíslegum ástæðum. Fyrst og fremst eflir útskýranleiki samskipti manns og tölvu, þannig að utanaðkomandi aðilar geti spurst fyrir um af hverju kerfið hafi tekið þær ákvarðanir sem það tók, og gerði það sem það gerði, eða hvers vegna það framkvæmdi ákveðnar aðgerðir frekar en aðrar. Í öðru lagi getur útskýranleiki leitt gervigreindina áfram í að efla lærdómshæfileikana, þar sem hæfileikinn til þess að útskýra mistök og bæði slæman og góðan árangur getur hjálpað til við að læra hvernig hlutirnir virka. Í þriðja lagi veitir útskýranleiki gegnsæi, þar sem viðeigandi útskýringar geta verið grunnurinn að viðeigandi áætlunum, útdráttum, réttlætingum, spám, o.s.frv. Til þess að vera traustsins vert, sér í lagi þegar kerfi er beitt þar sem mikið liggur undir, þarf gagnsæi að vera til staðar í öllum þáttum gervigreindra kerfa. Að veita gagnsæi í nútíma vélrænu gagnanámi og öðrum gervigreindarkerfum svo sem styrkingarnámi og djúptauganetum krefst töluverðrar vinnu eftirá af hálfu sérfræðinga, þar sem þekking slíkra kerfa er yfirleitt ekki framsett á læsilegan hátt. Rannsóknir í gervigreind hingað til hafa lagt áherslu á túlkanleika, en forðast alla greiningu á orsakasamhengi, þó svo að það sé í raun mikilvægasti þáttur útskýringa. Ólíkt fyrri rannsóknum er áherslan hér á *sjálfvirka útskýringu:* hæfileika kerfa til þess að framleiða sjálf útskýringar á eigin hegðun, umhverfi og markmiðum á sjálfvirkan hátt. Við metum möguleika tveggja kerfa til þess að framleiða útskýringar sjálfvirkt á þekkingu sem þau hafa tileinkað sér og berum saman við aðrar nálganir á útskýranlegri gervigreind, sem skref í áttina að því að geta lagt samanburðarhæft mat á útskýranleika gervigreindra kerfa. Kerfin tvö, AERA og NARS byggja á hungmyndafræði sem er gjörólík almennum aðferðum í gervigreind, þar sem mikil áhersla er lögð á sjálfsskoðun. Megináherslan hér er á tvo mikilvæga þætti útskýringa í gervigreind: Af hverju samanstendur gagnleg útskýring, og hvaða kröfur þarf kerfi að uppfylla til þess að geta framleitt slíkar útskýringar. Við metum möguleika núverandi kerfa til þess uppfylla þessar kröfur.

Efnisorð: Útskýranleiki, gervigreind, alhliða gervigreind, sjálfvirkni, orsakasamhengi

Self-Explaining Artificial Intelligence: On the Requirements for Autonomous Explanation Generation

Hjörleifur Rörbeck

Thesis of 60 ECTS credits submitted to the School of Technology,
Department of Computer Science
at Reykjavík University in partial fulfillment of
the requirements for the degree of
Master of Science (M.Sc.) in Computer Science

May 2022

Student:

....

Hjörleifur Rörbeck

Examining Committee:

....

Dr. Kristinn R. Thórisson

....

Dr. Pei Wang

....

Dr. Patrick Hammer

No Examiner

....
date

....
Hjörleifur Rörbeck
Master of Science

Acknowledgements

This work was funded in part by the Icelandic Institute for Intelligent Machines and Cisco Systems. The author would like to thank Luca Aceto and the Office of the Department of Computer Science for their help in enabling this thesis to be finished, Kristinn Thorisson for his constant support, advice, and time spent in discussions, Patrick Hammer and Pei Wang for agreeing to be on the committee for this work, and for their helpful comments and feedback, Arash Sheikhlar and Leonard Eberding for their invaluable help, Jeff Thompson, and Matteo Belenchia for all their assistance.

Contents

List of Figures

List of Abbreviations

M.Sc.	Master of Science
AI	Artificial Intelligence
AGI	Artificial General Intelligence
GMI	General Machine Intelligence
XAI	Explainable AI
AERA	Auto-Catalytic Endogenous Reflective Architecture
NARS	Non-Axiomatic Reasoning System
GDPR	General Data Protection Regulation
LRP	Layer-wise Relevance Propagation
PSCM	Problem-space Computational Model
LIDA	Learning Intelligent Distribution Agent
LHS	Left-hand side
RHS	Right-hand side

Chapter 1

Introduction

1.1 Background

There are several reasons why explanation has made an entrance onto the stage of AI research topics at this point in the history of the field. Firstly, explanations in human-computer interaction can foster much needed user trust when applied in critical domains such as the medical domain and air traffic control [1], by providing a justification and basis for human understanding of how the machines operate. In fact, explanations of machine operation are important for this reason in any domain where machine-aided or fully automated decisions may have ethical implications, by increasing operational transparency and understandability for the human users [2], [3]. Additionally, the potential for collaboration between humans and AI systems in solving problems is increased by providing explanations. Thus, it enables humans to stay in the loop and gain insights from what the AI system has learned [4], [5]. Another benefit of explainable AI is that explanations facilitate more effective training, validation, and debugging, as transparency is improved for developers [3], [4].

Another aspect comes from legislation. For example, in the European Union, every citizen is entitled to a detailed explanation for any usage of their personal data or automated decisions that affect their lives in any way. Voigt & Von dem Bussche [6] show that the European Union's General Data Protection Regulation (GDPR) states this right as follows:

> The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her [...]. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an **explanation** [emphasis added] of the decision reached after such assessment and to challenge the decision [6, pp. 291–292].

Making explanations a necessity if AI systems are supposed to work without human interference, in order to be able to fulfill the obligation to produce explanations in accordance with the new rights provided by GDPR.

Lastly, we posit that explanations can enhance an AI system's understanding of *itself*. The explanations come from within the system itself and are generated in such a way that they can also be processed by the system for the purpose of further knowledge acquisition [7] . In other words, given that the system is a *self-explaining* AI.

Recent studies on explainable AI so far have focused almost exclusively on neural networks, and the identification of relevant input that leads to certain outputs. This certainly has benefits, especially with regard to validation and verification that the output is in fact justified, and debugging when this is not the case. However, the knowledge representation used by neural networks is ill-suited for explanations, since it is purely covariational. No causal chains of note can be derived, no complex tasks analyzed - they are limited to the mapping between a certain input and a corresponding output. Additionally, interpreting these systems typically requires considerable effort.

We envision the goals of 'explainable AI' differently. First and foremost, we recognize that the primary practical application of AI is all sorts of automation, and therefore autonomous explanation generation should be a goal for explainable AI as well. In short, the human effort needed to arrive at an explanation should be minimized as far as possible, delegating the explanation generation to the machine. Furthermore, in the context of learning machines, the scope of explanations needs to be expanded so as to provide insight into what knowledge the AI has acquired. In particular, they should enable humans to learn from an AI how they accomplish a complex task. In theory, it should even allow one AI to learn from another.

We therefore argue that systems without such abilities cannot be said to be truly explainable, as they are incapable of inspecting their acquired knowledge for the purpose of extrapolating explanations. Since any such extrapolation requires post hoc processing, we would instead refer to these types of systems as interpretable. Interpretation is in fact different from explanation, not being qualified by the same requirements. In order to show this, we must first define what properties are necessary for an explanation.

"Explainable" and "interpretable" are often used interchangeably in AI, but what is the difference, if any? Palacio et al. [8], in their attempt to create a unified framework for explainable AI, define explanation as "the process of describing one or more facts, such that it facilitates understanding of aspects related to said facts", and interpretation as "the assignment of meaning (to an explanation)," citing the philosophical origins of the term. In other words, explanations describe the facts relevant to the explanandum, and interpretation relates those facts to the task or goal being attempted. While we do not disagree with this definition, we yet find it lacking in substance, owing to the fact that they are simply dealing with a completely different type of learners - the goals of both the learning and the explanation are different. Machine learning algorithms like artificial neural networks (ANNs) aren't currently designed to acquire causal knowledge, and they do not learn how to perform complex tasks with many steps, where causality is needed. Therefore, the kinds of explanation that are needed are different. What we derive as the foremost necessity of an explanation is to determine the *most relevant causes.* Interpretation (in machine learning) concerns itself with the discovery of patterns in the input data leading to the corresponding output, that is the "meaning" when explaining classification tasks, determining the relevant patterns in the input that lead to a belief or action. This can be very useful, for instance in working with artificial neural nets to discover correlations and weed out those that aren't actually related to the class label, such as watermarks in images that cause them to be classified as pictures of horses [9]. This qualifies interpretation as an explanation of the mechanics of the classifier, but not anything else, such as the subject matter (in this example being horses). We therefore suggest the following definitions of, on the one hand, *explainable AI* as “AI that is capable of valid explanation”, and on the other, *interpretable AI* as “AI that can be interpreted”. The phrase ‘explainable AI’ has to date been used mostly in the latter sense—since no ANN is yet able to produce valid explanations[1] of its own operation from scratch. Therefore, any and all ANNs that are labeled as “explainable” can essentially only be, at best, interpretable AI.

The concept of ‘grounding’ an explanation is of importance for any explainable AI. A system which is able to explain must be able to create an “argument” for its explanation, in light of other knowledge which has itself been validated, thus *grounding* it in a web of other knowledge that supports its claims. To do this, knowledge of cause and effect is essential. This concept therefore separates statistics-based AI systems from those which are able to generate valid, causality-based explanations. A detailed description of what an explanation is and how it is defined can be found in Chapter 4.

Physical events and actions of agents are influenced by other physical events and processes. So, an explainable AI system must capture the chain of events by detecting the transitions, changes of the values and direction of influences. In other words, the system must know how and in what direction variables influence each other. This feature can only exist in AI systems that create an directed acyclic graph (DAG) of variables, by which one knows which variables control which variables. A DAG also allows tracing backwards from the goals to the potential commands, a.k.a. abduction [10].

The above feature does not exist in artificial neural nets (ANN), since they learn associations between variables, which could be useful for prediction but not for goal achievement and explanation. In fact, ANNs are even not able to give a reason for why a prediction is made, due to their “black box"-like nature.

An improvement over interpretable AI is Auto-explainability, which is the ability to generate explanations on the fly, without the need for human intervention or any additional training. A step further is a “self-explaining” AI system. It is capable of automatically generating (valid) explanations for its own beliefs and actions. This is the kind of ability this thesis aims to examine the potential for in AI systems.

1.2 Research aims and contribution

The conceptual framework in this thesis is autonomous systems that automatically learn about complex tasks and environments in a self-supervised manner, and can automatically and naturally generate *valid explanations* about these. More specifically, the focus being on the principles that would enable us to build machines with such capabilities, which we refer to here as *‘self-explanation’*.

The explainability of different AI systems has been investigated many times from many angles. However, the feature of self-explanation in AGI architectures has not been fully addressed yet.

[1] The use of the term ‘valid’ refers to a process by which the explanation is validated through the generation of arguments that references (a network of) causal relations. Therefore, the system must be able to a) model causality, and b) abduce arguments which “argue for” the explanation.

To define self-explanation, we must first take a step back and define ‘explanation’, particularly in the context of AI. It is important to look at other sciences here, as explanations have been extensively researched in various fields, such as psychology and social studies. An explanation has a practical purpose, so its definition must encompass the fulfillment of that purpose.

We must also understand this purpose in the discussion of why self-explanation is needed, to determine what self-explanation adds to the equation - what further insights will be made possible by the user and by the AI itself.

In this research, we

- define self-explanation,
- discuss why it is needed,
- what its requirements are, and
- which current cognitive architectures have the greatest potential for self-explanation, and to what extent they fulfill the aforementioned requirements

Given that self-explanation is a worthwhile goal, which we hypothesize it is, the first step towards developing such capabilities is to determine what the requirements are. That is the main contribution of this work, along with an analysis of current cognitive architectures in two phases; first a shallow look at some of the best known ones to assess their potential to fulfill our requirements. Then a more detailed analysis of what we find to be the most promising candidates, with some examples.

1.3 Organization of Thesis

This thesis is organized into eight chapters. Excluding the introduction, the second chapter goes over current related work, both in machine learning and general intelligence. Chapter three outlines our methodology, and in four we examine explanations from a multidisciplinary standpoint. Chapter five details our requirements for self-explanation, chapter six contains a more shallow analysis of known cognitive architectures with regard to the requirements put forward in the previous chapter, whereas chapter seven holds a more detailed analysis of the most promising candidates, AERA and NARS. Finally, chapter eight consists of our conclusions and future work.

Chapter 2

Related Work

While the focus on explanations in AI research is a relatively recent development, and can mostly be traced to the recent increase in popularity of black-box algorithms – and the inevitable need to make them more understandable – extensive research has been done on the phenomenon in other disciplines as well, the results of which should not be ignored in the definition and design of explainable AI. Miller [11] provides an extensive survey of work in the social sciences on this topic. We must also consider what has been done in cognitive architectures and general intelligence - these fields are of particular interest, as they are not by necessity so encumbered by the lack of transparency as are most machine learning algorithms. The following chapter first gives an overview of explanations in AI and other research fields before moving to the field of machine learning and the notion of "explainable AI" (XAI). Lastly, current work on systems aiming for general intelligence and their explanation capabilities are introduced.

2.1 Explanations

Most sources agree that *causal attribution*, or identifying underlying causes of a class of (or particular) events or state of affairs, is a vital part of explanation [11]–[15] [1]. In fact, this is often how explanation is defined. Josephson equates finding possible explanations with finding possible causes [18], and Pearl claims that explaining a set of events necessitates the acknowledgment of the cause of those events [15]. Miller expands on this, arguing that explanation is a twofold process which begins with the cognitive process of identifying causes, followed by a social process of conveying the knowledge acquired by the cognitive process to the intended recipient. As he also points out, causal attribution is a twofold process of inferring the key causes and then selecting a subset of those causes as the most relevant for an explanation [11].

Palacio et al went with a broader definition of explanation, arguing that causation is not necessary for all explanation: "An explanation is the process of describing one or more facts, such that it facilitates the understanding of aspects related to said facts (by a human consumer)." [8, p. 5]. They further argue that understanding is

[1] Other types of explanation than causal do exist and should be mentioned here. Teleological explanations are explanations focused on utility, to explain by defining the purpose or intent of the thing to be explained [16]. But when explaining complex tasks this does not always apply – not all things in need of explaining have intent or utility behind them. Reductionist explanations also exist [17], but it is not apparent how such explanations could be automated and therefore they fall outside the scope of this work.

unique to humans, and therefore explanation from machine to machine is merely verification. While it may be true that not all explanations require causation,[2] certainly explanations involving complex tasks with multiple steps and subgoals will invariably do so. We will therefore treat causation and causal knowledge as a necessary element in our research. (It is worth mentioning that we disagree with their assertion that 'understanding' is something limited to humans only; we consider it among the goals of AGI research [19]).

The processes of causal attribution have been systematically analyzed in the work of Pearl [20]. He defines causal models as having sets of endogenous and exogenous, or internal and external variables, respectively. Defined in this way, an agent using his approach to generate an explanation of an event can avoid selecting irrelevant causes, as they are treated as external to the model [15]. For example, if the explanandum (thing to be explained) is "why did the cat get wet," and it is known (by explainer and explainee both) that it is raining outside, this would be an exogenous variable, as it does not represent relevant information. A more suitable cause for an explanation would be "the cat went outside". If it were not known that it was raining outside, but only that the cat went outside, the situation would be reversed. In summary, the most relevant causal structures in an explanation must be those not observable by the explainee, as conveying these maximizes the knowledge gained. We find this to be the most important aspect that makes an explanation *good*.

2.2 Machine Learning - XAI

There exist different approaches to explainable AI in current machine-learning research. Doran et al. [21] therefore divide existing approaches into three categories: 1) opaque systems (unexplainable or black-box), 2) interpretable, and 3) comprehensible. Additionally [21] introduces a fourth category, which are truly explainable systems.

A lot of research in XAI focuses on post hoc interpretation of an AI system's rather than on actual explanations [22] [23] [24] [25]. This comes mostly from the fact that AI systems, in particular neural networks, often belong to the first category of opaque, or black-box systems [26]. Interpretability is added afterwards by analyzing the neural network. Let us consider what this interpretation of ML models is. The need to develop methods for "visualizing, explaining and interpreting deep learning models", was expressed in Samek et al [27], and the reasons for doing so identified as (i) verification and improvement of the system, (ii) learning from the system, and (iii) compliance with legislature.

Verification is closely related to the user trust issue mentioned above; if the reason why a decision was made based on the output of a machine learning model can be interpreted so as to rationalize or justify that decision that would certainly increase the trust the user has in it.

Humans learning from a system is potentially a very useful aspect of explainable AI, and this is both the case when interpreting black-box models and for other types of explainable AI, because computers have different data processing capabilities than humans. It therefore stands to reason that there is something to be gained by insight into the patterns learned by these systems.

[2] This is of course partly a question of definition.

Improvement comes here from the ability to determine what factors adversely affect the outcome of a model. This will inevitably involve some human intervention, where whoever trains the model studies the interpretation of it and acts on it somehow, if possible. Such improvement can therefore never amount to the kind of autonomous improvement mentioned earlier; in order for a system to be capable of using interpretations of its own models to improve itself, a significant level of understanding would be required. Black-box machine learning models can only be attributed a very limited level of understanding, if we judge them by the tests suggested in [28].

Many methods for visual explanations of machine learning models have emerged in an attempt to accomplish the above. TCAV, or "Testing with Concept Activation Vectors" is one such approach that has gained significant popularity, due to the ease with which a user can define a concept and test the model for sensitivity to it [23]. Several methods exist to visualize learned concepts by generating prototypes using activation maximization to cause neurons associated with the given concept to fire strongly [29].

A number of techniques use heatmaps to identify the most relevant parts of the input in making a decision [30], [25]. Of particular note is Layer-wise Relevance Propagation (LRP) [24], which provides a general framework for pixel-wise decomposition of most kernel-based classifiers and thus visualizing their contributions with heatmaps. SpRAy, introduced in [9], further improves upon this by applying clustering on LRP decisions in order to identify and explain different decision behaviors. This is particularly useful in the detection of so-called "Clever Hans" decision making, or when models have learned to make correct predictions for the wrong reasons (e.g. identifying horse images by watermarks, or wolves by snow).

Another approach that has been followed in recent years is the identification of counterfactuals through input adjustment. To generate counterfactuals, the system simulates inputs which are only slightly different, but result in a different output. While this approach has been used in the past, it is difficult to define what "good" counterfactuals are (what are their properties, their differences to the original input, etc.) [31], [32].

Stepin et al [33] identify the ethical challenges in explainable AI using counterfactuals as the following:

- Safety: This refers to the risk associated with revealing too much information about the model and thereby the training data as well, opening up the possibility for model stealing, or data leakage [34]. However, this is only a problem if the user of the system or intended audience of an explanation is not supposed to have access to the model and the data.

- Fairness: When decisions made by an AI affect people's lives, it's important to avoid bias and discrimination inadvertently being included in the process. There are many known examples of this happening, and explainability can actually help address this problem by applying counterfactual scenarios to identify any bias [35].

- Feasibility: When explanations are used to suggest alternative actions, there are two important considerations. Firstly, the suggestion needs to be the most feasible option available, based on the knowledge of the system. Secondly, it needs to provide a confidence level, or at least have a confidence threshold.

- Accountability: Here, the concern is mostly on how an AI system is used rather than how it is designed or implemented. However, for the system to be explainable by design and to have traceability of decisions are the most important steps to ensure that the user is in fact able to use the AI in a responsible manner and thus achieve accountability.

Like much of present work on XAI, this list clearly betrays its primary focus on artificial neural nets (ANNs) – while the four key areas may apply to any system's (and possibly person's) ability for producing explanations, the details of Stepin et al.'s breakdown is by and large limited to narrow AI approaches. Explanations in such systems are mostly ad-hoc, because narrow AI systems are hand-crafted, their ability to explain is also hand-crafted. While an ad hoc interpretation is often desirable, it might be unnecessary in tasks which are well defined and do not provide much diversity,

The ability to explain can be contradictory to the original purpose of a machine learning system. One should therefore ask the question whether it is necessary for a system to explain, given its task [36]. Especially concerning narrow AI systems this is of importance. It could be argued that explanations are even more important in the field of general machine intelligence (GMI) than in narrow AI, due to the difference in the environments in which the systems are deployed. While narrow AI systems usually are only deployed in simple, well structured environments, AGI systems live in more open worlds, thus needing a more general repertoire. In such open worlds a system can use the ability to generate explanations ad hoc to explain interrelations to *itself*.

2.3 General Machine Intelligence

In the following, four GMI-aspiring architectures – SOAR, LIDA, AERA, and NARS – are introduced, and their abilities to create explanations are reviewed.

The SOAR cognitive architecture is based on the problem-space framework, where an agent is in a state and must choose its next action from a set of operators [37]. *Debrief* [38] is an explanation system that is integrated with SOAR agents in the simulations. The explanation system stores states of the agents over the simulation. Then, when the simulation is finished, the user can ask questions about what has occurred for which the system can generate a textual explanation. Also, the system can describe the causes of its actions by rerunning the simulation, in which some states have changed. This, however, is restrictive, since a reflective system must be able to generate explanations on the fly while it performs its task. In addition, Debrief is specific to the AI systems controlling the simulated entities, and is not directly applicable to other AI systems.

The LIDA (Learning Intelligent Distribution Agent) cognitive architecture has different cognitive cycles. In its perceptual memory, LIDA assigns interpretations to its perceived sensory data, based on which it decides which incoming stimuli are more important than others [39]. However, LIDA uses the interpretations to only answer the question "what do I do next", which is used in its action selection mechanism. In other words, potential explanations can be created by LIDA only to help its architecture to focus on specific features of environments, which means that the system may not capture the cause and effects and thus answer **why** an event occurs in its surrounding.

The AERA system is based on acquisition of causal knowledge [40], which allows an explanation system to trace back to potential causes of events. AERA is equipped with a visualizer that clearly demonstrates prediction and simulation steps that occur in each time frame, such that the user can trace how the system commits to certain sets of actions that realize a high-level goal. This explanation generation can be done while the system is learning, and the system does not have to try different actions in reality to infer what would have happened in counterfactual scenarios. Instead, the system performs simulations at each time frame, which makes it capable of both explanation and self-explanation through a set of causal networks, all leading to the same high-level goal.

The NARS system [41] has a few explanation tools, one of which is the capability of directly interfacing with the system using the Narsese language and posing questions, which lead to the generation of a goal to find an answer to the posed question. In other words, one of the two main tasks of NARS is question-answering. The system receives a set of evidences in the form of Narsese statements and generates different Narsese statements based on a set of inference rules (deduction, abduction, induction and analogy) and the given evidence. Then, if the user asks questions, the system can provide the inferred statements along with their truth values.

2.4 Causal explanations

Pearl et al [42] define causal explanations using structural equations, for the purpose of determining and conveying an actual cause of an explanandum. To accomplish this, they envision a causal model, and work under the assumption that all relevant facts are known to said model. Much of their work is useful and applicable in our context, such as their definition of causal explanations and the importance of the epistemic state of the agent. However, what is lacking is a focus on tasks and goals rather than simply explaining, and particularly complex situations where complete knowledge of relevant facts is nigh impossible to have. We must therefore adapt their definitions for a more practical, problem-solving based approach.

2.5 Summary

Current research in explanation shows that causality and counterfactuals have long been introduced in the concept of an explanation. However, as current deep-learning research is solely based on correlational knowledge it fails to integrate causal explanations. Current XAI research tries to tackle this issue by generating correlational counterfactuals through input variation. However, using probabilistic approaches, instead of causality, leads to the generation of counterfactuals which have close to no meaning to humans (see e.g. [31]). While work has been done making counterfactual explanation of deep learning systems more understandable to humans [31] it still fails to meet the definitions of Pearl [15] and Miller [11].

Chapter 3

Methodology

The goal of this thesis is to evaluate the ability of learning AI systems to generate explanations. This focus on learning stems from the fact that since the purpose of explanations is knowledge transfer and the subsequent elevation of understanding, it stands to reason that the most valuable knowledge to be gained from such a transfer is knowledge that was learned first-hand, i.e. from *experience*.

Another important consideration is that the explanations must be good, that is, they must satisfy some requirements regarding the knowledge they effectively provide. We have therefore chosen to identify what constitutes a good explanation and analyze what is required, both in terms of information and cognitive abilities, to produce such explanations on demand, where the precise topic or purpose of the explanation is not known beforehand.

The top-level questions that we want to answer are:

1. Explanation is a transfer of knowledge and understanding from one agent to another. However, what are the features that make an (autonomously generated) explanation a "good" explanation?

2. Self-explanation brings about self-reflection. However, why is it important to have a self-explanation capability for AI systems?

3. An AI system that is capable of generating self-explanations must fulfill a set of requirements. What are the requirements for such a self-explainable AI system?

4. Explanation methods for AI systems have been investigated for different architectures and algorithms. However, the topic of self-explanation has not been addressed yet. So, we want to know to what extent current cognitive architectures fulfill the requirements for self-explanation?

This calls for a methodology that ideally allows us to do a conceptual analysis of the candidate systems' capabilities as well as produce experimental data from attempts to leverage their respective frameworks to generate explanations. Both are challenging and could each be the subject of one or more theses. Here, we position ourselves between the two: We analyze to some extent the potential of candidate systems, selecting only the ones we find most promising to autonomously run experiments on and produce data. From those results we can then clarify some aspects of their abilities and limitations in producing explanations.

To this end, the methodology employed in this work consists of the following:

1. **Outlining and justifying the need for explanations in AI**. The ability of an agent to explain a phenomenon can show the level of the agent's understanding [19]. When the agent explains a phenomenon, it transfers some knowledge about aspects of the phenomenon that are hidden to another agent that receives the explanation. In the case of self-explanation, it reflects on different aspects of the phenomenon and infers some other knowledge that allows it to better understand why the phenomenon exists. Thus, self-explanation leads to having higher levels of understanding in AI systems.

2. **Investigating current definitions of explanation and explanation generation, classification of methods, and choosing those that are most appropriate**. As seen in chapter 2, the current explanation methods in the literature of machine learning are usually designed to increase the trust level in human users in relation to decisions and predictions AI algorithms make. In other words, the methods provide some information that is understandable to humans but do not necessarily help the system learn better, if fed back to the system. In most AGI-aspiring architectures, the same issue exists, due to the fact that the knowledge they create is not causal and therefore the reflective reasoning through self-explanation becomes impossible. In fact, few systems exist that create models of their world that can support reasoning about different hypothetical scenarios that might occur.

3. **Analyzing the requirements for being able to generate explanations**. An AI system needs to be able to generate explanations on-the-fly, such that they improve the ability to learn the assigned tasks over time. In addition, the knowledge representation that an AI system utilizes is a vital factor for its capability to create explanations of itself, or the knowledge it has acquired. Causal knowledge acquisition allows AI systems to test and modify their knowledge and apply reasoning via a causal graph based representation of knowledge.

4. **Evaluating current AI systems and cognitive architectures with regard to their ability to fulfill the above requirements**.

   Although in most AGI architectures, comparable cognitive functionalities are implemented, their knowledge representation can be radically different, which leads to significant differences in the architectures' capacity for self-explanation. In addition, fundamental reasoning rules, by which a system can infer and thus produce more knowledge, are not compatible between architectures. This makes the learning process in those systems totally different. But, for self-explanation, only systems based on causal knowledge and inference can create knowledge graphs that allow reasoning about potential causes of events and phenomena.

5. **Selecting the most promising candidates from the evaluations to run experiments on, for further in-depth analysis**. Only a few AI architectures are capable of self-explanation and thus self-reflection. We have chosen the AERA and NARS systems since we find they have higher potential compared to others in generating causal explanations. NARS creates event-based statements that can show causal chains between two or more events. AERA can create simulation branches in which the system inspects different hypothetical scenarios that are possible after learning causal models, and explain why a particular

action was taken by the system. AERA can in fact be said to be capable of producing a “good" explanation [1].

[1] For how we define a “good" explanation, see Good Explanations

Chapter 4

Explanations

In this chapter we will attempt to get closer to a crisp definition of what is an 'explanation.' Before we can consider how to do that for the purposes of this thesis, and particularly what constitutes a *good* explanation, it is a good idea to investigate briefly why explanations even exists, and when and how they are used.

4.1 Why Do We Explain?

From the standpoint of psychology, Lombrozo defines explanations as "employed as a basis for constraining inference and guiding generalization" as well as promoting understanding [13]. In other words, explanation is a means to facilitate learning, and not just in the explainee but also the explainer. Of particular interest is her finding that the cognitive process of causal attribution benefits the understanding of the explainer in similar ways to how it benefits the explainee, resulting in judgments and generalizations based on more than covariation [13].

Because AI systems are built to perform tasks (whether novel or known), and intelligence is only needed if there is some variation in a task that could not be handled by pre-programmed routines, having AI systems be able to explain things – their understanding of a task and situation, intent, plans, and reasons for action (whether erroneous or successful) – can be extremely valuable for a number of reasons.

Any AI system is designed to provide some utility. The utility of a given system then drives it in the form of goals, whether it is simple classification or the completion of complex tasks. Explanation, then, is a type of goal, driven by the utility of exposing hitherto unknown causal relations to the explainee. For humans, the cognitive process of constructing the explanation with causal attribution deepens the understanding of the explainer, and is followed by the social process of transfering that knowledge to the explainee, thus deepening his understanding. In most cases the top-level goal of explanation is to increase the understanding of the explainee. However, the concept of understanding has a shallow foundation in AI research. Thórisson et al. [19] proposed four critera to evaluate understanding of a phenomenon Φ in an AI system:

1. To predict Φ
2. To achieve goals with respect to Φ
3. To explain Φ
4. To (re)create Φ

We find this definition of *pragmatic* understanding proposed by Thórisson et al. (2016) most suitable for our purposes here, as we are concerned with pragmatic explanations relating to complex tasks. Furthermore, the most important criterion for evaluating the capacity for explanation of an AI system is to what extent it can facilitate this very understanding to an explainee, and the ease with which this can be accomplished.

We see at least three reasons, or ***goals***, related to the production of explanations:

1. To identify relevant causes and contexts for goal achievement (unexpected failure and unexpected achievement). This may proceed, for instance, through hypothesis generation based on counterfactual argumentation[1] (see Section 4.2).
2. To highlight variables, patterns, or other aspects relevant to an observed event that an explainee missed (failed to consider).
3. To produce evidence (reasoning) for a plan intended to achieve a particular set of goals.

In our view, these form three fundamental categories of goals related to the production of explanation.[2] The relevance of these to AI, and especially general machine intelligence, should be rather self-evident; we will come back to these in Section 4.4.

4.2 How Do We Explain?

One would assume that, when devising a means to automate an explanation of complex processes, the first thing to do would be to examine how humans have been explaining things to each other. Yet, this has been mostly ignored in explainable AI research to date.

Hilton [43], [44] did extensive research on explanations from the psychological perspective. They point out the inherent fallacy in using covariational criteria for causal attribution, as there are numerous examples of events occurring at the same time without one being the cause of the other. Furthermore, they proposed an alternative model of explanation, based on findings in ordinary language philosophy that humans make use of contrastives and counterfactuals as criteria for causal attribution.

This is in fact one of the major findings of Miller in his survey [11], that explanations most commonly come in response to contrastive questions, for instance "Why did you do A and not B?" rather than simply "Why did you do A?" Halpern and Pearl [15], [42] also built on this idea, modeling counterfactuals to define actual causes.

But what is the benefit of contrastive explanations? For one, they point to what knowledge is missing for the explainee, enabling the explainer to identify the most

[1] In our view, such hypothesis generation and counterfactual argumentation does not need to be explicit in the system—they can be implicit and thus not available for introspection by the system's internal processes. Any learning-capable systems that are incapable of producing explicit explanations are likely constructed this way.

[2] Note that the above goals can be active in the context (and for the higher-level purposes) of communicating to another cognitive agent as well as producing information in a new form for other internal processes in your own mind. To us, these situations are therefore comparable, because the contexts do not change the above explanation-related goals. (Therefore, we do not make an effort to explicate whether a "explainee" is an agent, a cognitive process, or some other receiver of the explanation.)

relevant causal chains for an explanation. Lipton [45] also argues that it is easier to explain a difference between A and B, rather than providing a complete explanation for A. Since for any situation or phenomenon that calls for explanation there may be multiple competing hypotheses, some of which are more likely to achieve a desired outcome in the current context than others, contrastive (meta-) explanations may in fact serve an important purpose in planning and learning.

This is a valuable lesson to be learned from how humans explain. Constraining the explanation by indicating what knowledge is missing using contrastives (or possibly by other means) not only makes for a better explanation as it is more relevant, but also makes it easier to find, or compute.

Counterfactuals are conditional statements such as "if p, then q," where p is false [46]. That is to say, they represent an alternative state of the task-environment, given that certain conditions had been different. For example, "The door would have opened, had it not been locked" is a counterfactual statement where the door being in state 'open' is a hypothesized alternative state to the actual one (being 'closed'), that would have been observed, given the same action (attempt to open) but with different starting conditions (being unlocked).

The work of Halpern and Pearl in particular [15], [42] has made counterfactuals the center of focus for (human) explanations. One problem with Halpern & Pearl's approach is that the "counterfactual inference" they studied is mostly "hypothetical inference" because they only discussed the inference from "if p then q" without considering the effect of the "not p" part, which requires selectively deny the logical implications of "not p". Counterfactuals are certainly important for human-level explanations, but because human explanations rest on assumptions that cannot be upheld at the present in AI, including the assumption of a lot of knowledge, "common sense," and human-motivated (psycho-social) purposes. Here we are focused primarily on machines, however, and since the current state of AI is not anywhere close to human-level, our focus is less ambitious, focusing more on the more basic low-level aspects and functions of explanations.

4.3 How We Define Explanation in This Thesis

Of primary concern in this thesis are explanations of complex tasks. We define a complex task as a task that is composed of multiple elementary steps, many of which may be performed in a variety of ways, but which form groups that contain constraints between them, such that e.g. one group must be finished before another can begin, all fall under a maximum time (taking longer in one part will thus diminish time available for others), etc. In addition, all physical tasks are layered, where some subsets of a task involve finer or coarser granularity (e.g. the transportation mode for one leg of a trip may be more detailed than that of another). As proposed by Belenchia et al. [47], level of detail is considered part of the task – that is, if the levels of details of a task are changed, the task itself is changed. Therefore, all known levels of detail relevant to a task must be defined in relation to a task performer, because it is the elementary operations that the performer is capable of performing that determine how the task could be performed, and how difficult it is.

A task like driving to the countryside may involve several subtasks that are highly time-sensitive (in that missing for instance an exit may be a matter of 10 seconds, which in a 5-hour trip is a very small part of the achievement of that task). While such temporal constraints can be added to any instance of explanation generation, generating explanations does not seem subject to the same constraints, that is, an explanation may be produced slowly or quickly, somewhat independently of the passage of time according to the clock-on-the-wall. Therefore, we do not consider time to be an integral part of whether an explanation is valid, or whether it is good.

According to our view, explanation generation *is itself a task*. The task of producing explanations requires certain information, and follows certain requirements to count as an explanation. More specifically, explanation is the generation of a *compact description* that references (one or more) causal relations that – if not present, or structured differently – would result in a different outcome [48]. The necessary ingredients, therefore, are

1. knowledge of causal (and other) relations,
2. named entities (and appropriate grammar) for producing this description, and
3. a fulfillment of a (possibly hypothesized) goal that the explanation is intended to meet.

The task of an explainer (or explanation-generating process) is to meet the goal of the explanation, which may be, as mentioned above, to provide missing information (for the explainee – whether it is the system itself or an outside agent) or to guide the attention (of the explainee) to particular hidden information (that the explainee knew about, but did not consider).

4.4 Good Explanations

A *good* explanation, in the context of this thesis, is an explanation that (i) meets a goal that the explanation is supposed to fill, and (ii) can be validated through experimentation. Validation of an explanation – both validation of the general ability to explain, as well as the establishing the validity of a particular explanation – is a fundamental method for establishing the value of an explanation. Such validation can come in the form of experiments, whereby a counterfactual, for instance, is put to the test (e.g. a door that didn't open, with the explanation that this was because it was locked, is unlocked to see if that counts as an actual cause).

With reference to the three goals that an explanation may fulfill (see Section 4.1), there are at least as many methods to validate whether an explanation is good or not: Its goodness is simply a measure of how well it meets its goal(s).

Chapter 5

Requirements for Autonomous Explanation Generation

Now that we have established a definition of explanations and, more importantly, what is meant by 'good explanations,' we need to establish the requirements for an AI to be capable of autonomous explanation. What kind of knowledge representation and learning capabilities are needed for it to be able to automatically generate good explanations for its beliefs and justify its actions?

We propose the following:

1. The ability to process causal knowledge.
2. The ability to inspect, evaluate and modify its own knowledge.
3. The ability to estimate the confidence level of all acquired knowledge.
4. The ability to apply reasoning on current knowledge and input.

5.1 Causal Knowledge

This first requirement is immediately apparent from the previous chapter, where it was established that causality plays a central role in explanations. Causality in explanations is the single most important factor in determining what changes the course of events, and provides the foundation on which arguments for such explanations' validity can be based, via counterfactuals.[1] The bare minimum requirement for explaining complex tasks is then knowledge of its underlying causal structures. More strictly speaking, since we also aim for the explanations to be autonomous, the explaining agent should not only be able to process the causal knowledge but also acquire it. We cannot assume the environment will never change, so the causal models must be kept up to date, or the explanations may become invalid. This is where the next requirement comes in.

[1] The only exceptions to this might exist in explanations about simple things, like e.g. in the case of simple Boolean choices ("Pressing button A, not B, locks the door"), but these do not apply when explaining complex tasks involving multiple steps, which is our main concern here.

5.2 Self-Inspection

This requirement comes in three parts, the first two being more obvious. For generating explanations of a known phenomenon, it is necessary to access the knowledge one has of said phenomenon, and evaluate it in order to arrive at a proper explanation. Modifying it may also become necessary, when new knowledge is acquired and combined with current knowledge, possibly contradicting, correcting or enhancing what was known before. This is always necessary in dynamic, fast-changing environments where one cannot simply rely on axioms, as well as environments too complex to learn before doing, and then shut off the learning, and these are exactly the kinds of environments where explanations are the most useful.

5.3 Confidence Estimation

As discussed above, acquired knowledge may be contradictory or conflicting in some way. For example, say we observe a bouncing ball, and establish a model of its behavior. Then a different factor comes into play, suddenly there's wind and it behaves differently and the model must be updated. The same thing happens when the wind changes directions. Most of the time, things are not so clear-cut as to be predictable with absolute certainty, and in complex environments there are always improvements to be made. This is why it's necessary to establish the confidence in each piece of knowledge obtained. A model of a coin-toss is not very useful if it just predicts that heads are the most likely outcome each time because it has observed heads more often. Much more useful would be to predict heads with a given percentage of confidence, say 52 percent.

This not only drastically improves the output of the model, but also any explanations that use it; it is far more valuable for an explanation to say that there is a 51 or 99 percent likelihood of something occurring than to simply say it is likely. As the complexity of the state space is increased, so is the advantage of these estimates, because the estimates stack up.

5.4 Reasoning

Reasoning is needed to produce new knowledge from what is currently known. Learning is an obvious requirement for AI, but why is it required for explanations? Generating explanations is a task like any other, certainly a complex one, but it is a goal to be accomplished by an AI, and therefore learning is required to automate the process. If learning were not required, then the explanation could be documented beforehand and there would be no need for automation. But we have seen that this is not the case, and there is a need to make the decisions made by AI more accessible to human users. Therefore learning is necessary, and thereby so is reasoning.

Chapter 6

Candidate Architectures

We will examine four systems and evaluate to what extent they fulfill the aforementioned requirements. The four systems we have chosen are SOAR, LIDA, NARS and AERA.

6.1 SOAR

The SOAR cognitive architecture is based on the problem-space framework, where an agent is in a state and must choose its next action from a set of operators. Selecting an operator results in its application, and a new state emerging. A problem space is thus the set of states the agent can possibly achieve by applying the operators in a given task environment, and a problem consists of an initial state and a set of goal states.

SOAR uses a problem-space computational model (PSCM) to organize its knowledge and behavior. It has several types of memory, a short-term memory where perceptions are stored, a working memory acting as the global workspace, and three independent long-term symbolic memories, each with separate learning mechanisms.

These three symbolic memories are:

- Procedural memory, which stores procedural knowledge, represented as production rules. Operator selection is controlled by preferences, which are generated from these production rules by matching conditions against the working memory. The procedural memory uses two learning mechanisms. Chunking is the process by which new production rules are learned, and reinforcement learning is used to tune the actions for these rules, thus creating preferences for operator selection.

- Semantic memory, which stores general facts.

- Episodic memory, which stores snapshots of the working memory.

When the agent reaches an impasse, complex reasoning occurs, creating a substate in an attempt to resolve it. In this way, a SOAR agent is capable of subgoaling and even reflection [37].

Given these capabilities, generating explanations linking actions to production rules would be easy. However, tracing the creation of the production rules and their connection to preferences for certain actions would be quite difficult, as reinforcement learning leaves little of use for explanations. This imposes a hard limit on how deep an explanation can go, and thus fails to some extent the requirements outlined above for producing explanations from scratch.

6.2 LIDA

The LIDA (Learning Intelligent Distribution Agent) cognitive architecture has strong ties to cognitive science, evident in its theoretical similarity to how we see the human cognitive process. Like SOAR, it has many different types of memory, and different kinds of learning. The LIDA cognitive cycle is divided into several steps:

- The **Understanding Phase** is where the agent makes sense of input or sensory stimuli, passing low-level input to high-level representations and connecting the dots in the so-called perceptual associative memory that keeps track of relations between objects. The agent's model of the current situation is updated.
- The **Attending Phase** is where prioritisation takes place. Portions of the model of the current situation compete for attention, the most urgent perceptual structures are then broadcast and effectively brought to the attention of the agent for the duration of the cycle.
- Finally, the **Action and Learning Phase** is where the structures now available in the procedural memory are evaluated. Each template, or scheme of possible actions has an activation value and expected result. Once the action with the highest confidence is selected, it is executed and concurrently all the memories (perceptual associative, transient, episodic, attentional, procedural) are updated.

Learning in LIDA occurs on two levels, and there are three distinct types of learning; perceptual learning is about recognition, mapping input to internal representations, episodic learning is associative, and procedural learning is about learning actions. The two levels of learning are called instructionalist, or learning by creating new representations, and selectionist, or learning by reevaluating existing representations based on new information [49].

The focus in the design of LIDA on concepts from cognitive science leads to excessive complexity for the scope of this work, and a lack of focus on reflectivity.

6.3 NARS

NARS (Non-Axiomatic Reasoning System) and its implementation known as OpenNARS, are built on the foundation of the "assumption of insufficient knowledge and resources", or AIKR. This means that special consideration is given to resource management, time constraints, memory capacity, and prioritisation, and the necessity of providing useful feedback with respect to the confidence of any generated or acquired knowledge. NARS uses a formal language known as Narsese to represent its knowledge in the form of statements. Each statement includes at least two terms, and a copula representing the relationship between the two terms. A statement also has a truth value, consisting of its observed frequency, which is adjusted whenever the statement is observed to be either false or true, and the confidence value, which increases up to a maximum value each time the event is observed, whatever the outcome (for example, for the statement “dogs are black", the frequency is increased each time a black dog is observed, and decreased each time a non-black dog is observed, but the confidence increases each time a dog is observed whether black or not). One way to represent a causal connection in Narsese is with the temporal implication copula, which indicates that the occurrence of a subject term implies the occurrence of the predicate term after the fact. NARS receives input in the form of Narsese statements and learns, or creates new knowledge by applying a set of inference rules on currently held beliefs. The further from observed evidence a belief is, the lower its confidence will be. This is also affected by the kind of inference that led to the belief, for instance, deduction has a low negative effect on the generated belief, whereas abduction has a larger effect.

Using the properties from Langley et al [50] to characterize NARS, it becomes apparent that it is quite flexible. Narsese allows for a mixture of declarative and procedural representations, in turn enabling acquisition of conceptual as well as skill knowledge - even if the primary focus is on the former. Its memory is not limited to either episodic or semantic, but both are equally possible. As for the organization of knowledge, the granularity is determined by the input, and the same goes for whether it is flat or structured. Not being limited to any single approach, NARS is highly flexible, as befits Pei Wang’s definition of intelligence as adaptation under the assumption of insufficient knowledge and resources [51]. This, in addition to NARS fulfilling the proposed requirements for automatic generation of explanations above, is what makes it a great candidate for this research.

6.4 AERA

AERA stands for Auto-Catalytic Endogenous Reflective Architecture [40]. As the name suggests, one of the main areas of focus in its design is the ability to reflect on its knowledge, or in essence, to program itself. For that purpose, the Replicode programming language [52] was designed to encode the programs and models learned by AERA such that they would be not only human readable but also compiled and decompiled at runtime by AERA’s executive (the central processes of a runtime AERA). The architecture relies on a hierarchy of a vast amount of tiny “peewee-sized” models to make predictions and take actions towards goal attainment [53]. Each model has two sets of patterns, called left-hand-side and right-hand-side. When a pattern on the LHS is detected, the model predicts the pattern on the RHS, using a set of transformation functions called forward guards. Conversely, if a pattern represented on the RHS is detected, a hypothesis is put forth that it was produced by a pattern on the LHS, using another set of transformation functions called backward guards. Thus, models in AERA represent hypothesized causal relations. Since AERA has built-in methods to generate new models from scratch, AERA can use this to learn, by inferring the cause of observed patterns. Once learned, it can use the knowledge both to create plans, as well as predict what the results and side-effects of those plans might be.

Demonstrations of AERA have shown this approach to be capable of learning very complex tasks, such as a simulated real-time TV interview with two participants, where an AERA agent learns multi-layered plan-making from scratch by observing two people engaged in dialogue [54].

Chapter 7

Analysis of AERA and NARS

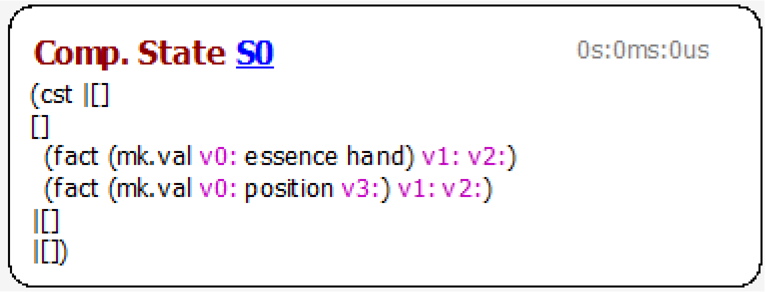
How the knowledge acquired by an AI is stored and retrieved is a major determining factor for the capacity to generate explanations from it. How human-readable it is determines to what extent a user may gleam useful insights, and how machine-readable it is the extent to which automated explanations will be possible. Let us therefore examine how AERA and NARS represent knowledge, and how their respective approaches to knowledge representation can be leveraged to create explanations.

7.1 Knowledge Representation in AERA

AERA represents knowledge using composite states and models. Composite states represent patterns that can be caught by models to produce predictions or actions that lead to the fulfillment of goals.

As an example to show AERA's capacity for explaining, we have a simulated task-environment where AERA controls a robotic arm. Initially, the arm is holding a cube, and it has three possible actions: move, grab, and release; the goal is to move a sphere to a certain location. In order to do this it must release the cube, move the arm to the location of the sphere, grab it, move to the intended destination of the sphere and release.

Using a custom-built visualizer [55] to decompile and show the entire process through which AERA decided what action to take and eventually completed the task, we can demonstrate how accessible explanations are for AERA and its users.



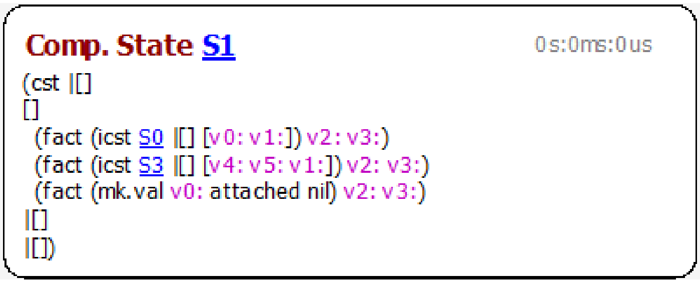
Comp. State S0

0s:0ms:0us

```
(cst |[]
[]
 (fact (mk.val v0: essence hand) v1: v2:)
 (fact (mk.val v0: position v3:) v1: v2:)
|[]
|[])
```

Figure 7.1: An example of a composite state in AERA

The patterns used by AERA's models are usually represented by the aforementioned composite states. A composite state consists of subpatterns for one or more variables that hold true for a specified time interval. The first example we have is shown on figure 7.1, a composite state with two subpatterns. The first one specifies a value for an object called v0 (within the scope of this composite state only), indicating that v0 is a hand. The second one indicates that v0 has a position v3. Both statements should hold true between timestamps v1 and v2.

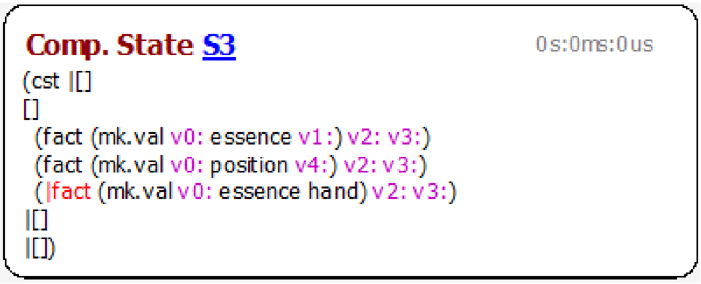


```
Comp. State S1                                  0s:0ms:0us
(cst |[]
[]
 (fact (icst S0 |[] [v0: v1:]) v2: v3:)
 (fact (icst S3 |[] [v4: v5: v1:]) v2: v3:)
 (fact (mk.val v0: attached nil) v2: v3:)
|[]
|[])
```

Figure 7.2: A second composite state example

Another example, shown on figure 7.3 has three subpatterns, again indicating something about an object v0, although this time it is a different v0, as we will see. The first one indicates that v0 has some type, or "essence" v1. The second one tells us it has some position v4. And finally, we have an antifact, which tells us that v0 is not a hand. All of these should hold true between timestamps v2 and v3.

And a third example ties the first two together on figure 7.2. Composite state S1 holds that S0 and S3 have been instantiated, that is, injected as a belief between timestamps v2 and v3. The variables v4, v5 and v1 in S1 represent the variables v0, v1, and v4 within S3, respectively. Similarly, the v0 and v1 in S1 are the v0 and v3 in S0. Note that since within S0, v3 is the position of the hand and v4 the position of the object within S3, and these two variables are both bound to v1 in S1, they must have the same position. Finally, v0, or the hand object, is "attached nil", which means it is not holding anything.

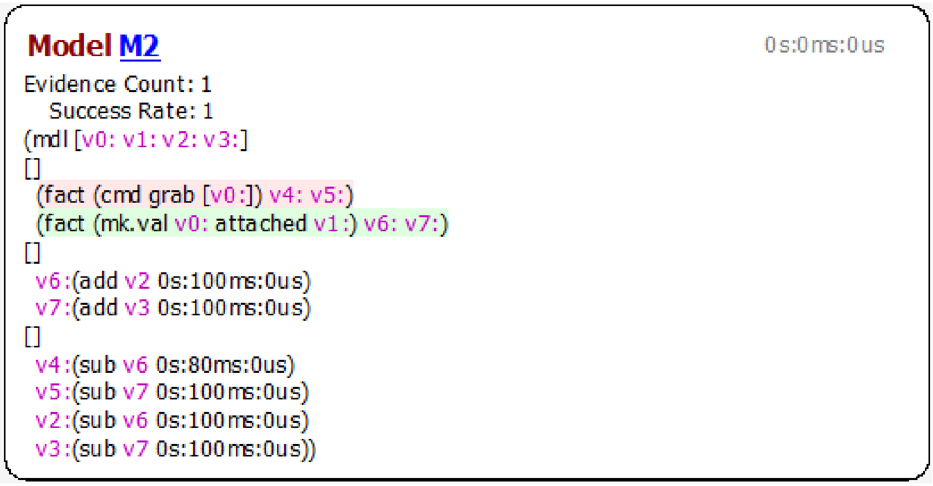


```
Comp. State S3                                  0s:0ms:0us
(cst |[]
[]
 (fact (mk.val v0: essence v1:) v2: v3:)
 (fact (mk.val v0: position v4:) v2: v3:)
 (|fact (mk.val v0: essence hand) v2: v3:)
|[]
|[])
```

Figure 7.3: A third composite state example

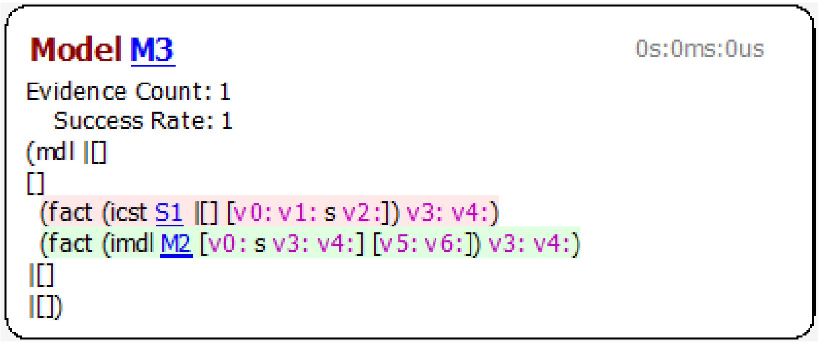
Now we can look at the models and how they use these composite states. On figure 7.5, the left-hand side (or LHS) of model M3 consists of a pattern where composite state S1 has been instantiated. In other words, there is a hand, and an object that is not a hand, and they both have the same position. Now, the right-hand side (or RHS) indicates that model M2 should be instantiated with specific values originating from composite state S1.

Model M2 (figure 7.4) has a LHS pattern that activates if the command was given to grab whatever is there with object v0 (presumably the hand) at time between v4 and v5. It will then predict that following this, at times between v6 and v7, it will be attached to object v1.



```
Model M2                                      0s:0ms:0us
Evidence Count: 1
  Success Rate: 1
(mdl [v0: v1: v2: v3:]
[]
 (fact (cmd grab [v0:]) v4: v5:)
 (fact (mk.val v0: attached v1:) v6: v7:)
[]
 v6:(add v2 0s:100ms:0us)
 v7:(add v3 0s:100ms:0us)
[]
 v4:(sub v6 0s:80ms:0us)
 v5:(sub v7 0s:100ms:0us)
 v2:(sub v6 0s:100ms:0us)
 v3:(sub v7 0s:100ms:0us))
```

Figure 7.4: An example of an AERA model



```
Model M3                                      0s:0ms:0us
Evidence Count: 1
  Success Rate: 1
(mdl |[]
[]
 (fact (icst S1 |[] [v0: v1: s v2:]) v3: v4:)
 (fact (imdl M2 [v0: s v3: v4:] [v5: v6:]) v3: v4:)
|[]
|[])
```

Figure 7.5: Another example of an AERA model

7.2 How to Pose Questions to AERA

AERA's models are causal and thus, relate a cause variable (LHS) to another effect variable (RHS). In this way, if a fact (either an observation or a simulated fact) matches the RHS of the model, the system can trace back to the potential causes and eventually select the most important one. Thus, AERA system is capable of producing causal explanations of observed events, simulated facts or taken actions by backward chaining (a.k.a abduction).

It is therefore quite conceivable that AERA could learn to explain itself, given a goal to do so.

7.3 Causal Explanations in AERA

All of this data, the models created or used, any composite states injected, as well as predictions made and actions taken, are readily available to the user, third party applications like the visualiser, and even more importantly, to AERA itself. This availability opens up a lot of possibilities when it comes to explanation. In the visualiser itself, it is possible to get shallow explanations for predictions. On figure 7.8 we see an event: The command was given to release. This results in a prediction (seen on figure 7.7) that the hand will be attached to nil, i.e. be empty. Now, say we want to know what caused this prediction, the visualiser can point to what made it - the explanation

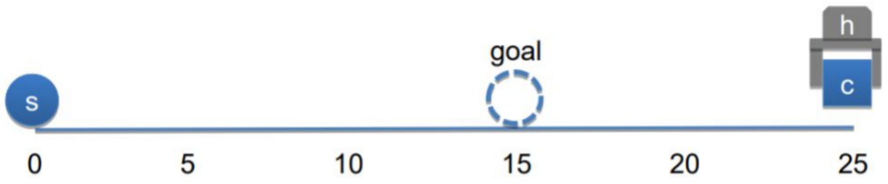
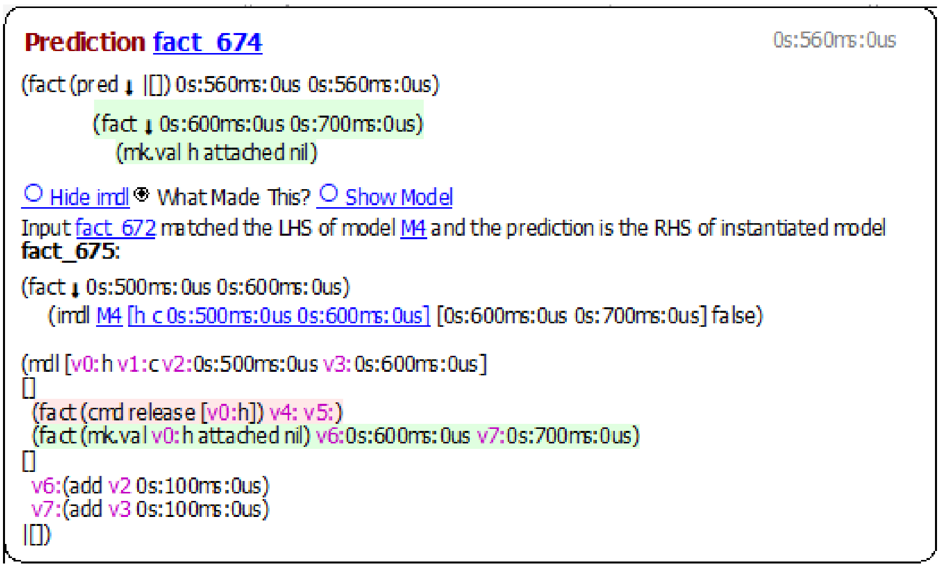


Figure 7.6: An illustration of the hand-grab-sphere task

can be seen below on figure 7.7: The event on figure 7.8 matched the left hand side of model M4 (figure 7.9), which resulted in the model injecting a prediction for its RHS.



Prediction fact_674 0s:560ms:0us

```
(fact (pred ↓ |[]) 0s:560ms:0us 0s:560ms:0us)
        (fact ↓ 0s:600ms:0us 0s:700ms:0us)
          (mk.val h attached nil)
```

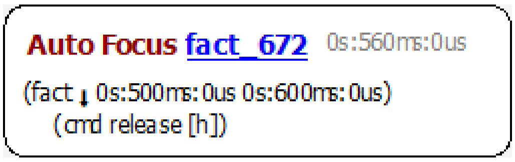
○ Hide imdl ◉ What Made This? ○ Show Model

Input fact_672 matched the LHS of model M4 and the prediction is the RHS of instantiated model **fact_675:**

```
(fact ↓ 0s:500ms:0us 0s:600ms:0us)
   (imdl M4 [h c 0s:500ms:0us 0s:600ms:0us] [0s:600ms:0us 0s:700ms:0us] false)

(mdl [v0:h v1:c v2:0s:500ms:0us v3:0s:600ms:0us]
[]
 (fact (cmd release [v0:h]) v4: v5:)
 (fact (mk.val v0:h attached nil) v6:0s:600ms:0us v7:0s:700ms:0us)
[]
 v6:(add v2 0s:100ms:0us)
 v7:(add v3 0s:100ms:0us)
|[])
```

Figure 7.7: An example of a prediction

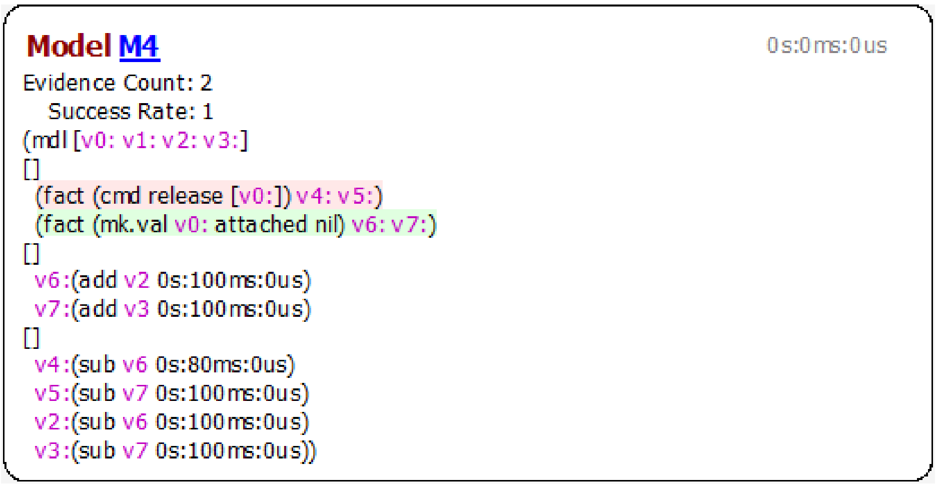


Auto Focus fact_672 0s:560ms:0us

```
(fact ↓ 0s:500ms:0us 0s:600ms:0us)
   (cmd release [h])
```

Figure 7.8: An example of an injected event

In the hand-grab-sphere example, the cause of an event is the action that the AERA agent takes by applying related commands. Therefore, to know why a particular command is generated, the agent examine the simulation process that led to its generation. This process starts by backward chaining from a high-level goal to potential actions that can be taken to achieve the goal. A simulation occurs at every time frame, and in the end of a frame the agent commits to a set of inference steps and thus a specific action. In other words, the system creates several simulation branches and selects the most promising one in every frame. The figure 7.9 shows how the high level goal of "sphere s being at position 15" is injected, after which the subgoal of "hand h being at position 0" is made, and this leads to producing the command “move hand by -25". By tracing back the inference steps from the high level goal to subgoals and then to applied commands, the system can explain its decisions in every time frame.



```
Model M4                                        0s:0ms:0us
Evidence Count: 2
   Success Rate: 1
(mdl [v0: v1: v2: v3:]
[]
 (fact (cmd release [v0:]) v4: v5:)
 (fact (mk.val v0: attached nil) v6: v7:)
[]
 v6:(add v2 0s:100ms:0us)
 v7:(add v3 0s:100ms:0us)
[]
 v4:(sub v6 0s:80ms:0us)
 v5:(sub v7 0s:100ms:0us)
 v2:(sub v6 0s:100ms:0us)
 v3:(sub v7 0s:100ms:0us))
```

Figure 7.9: The model that made the prediction on figure 7.7

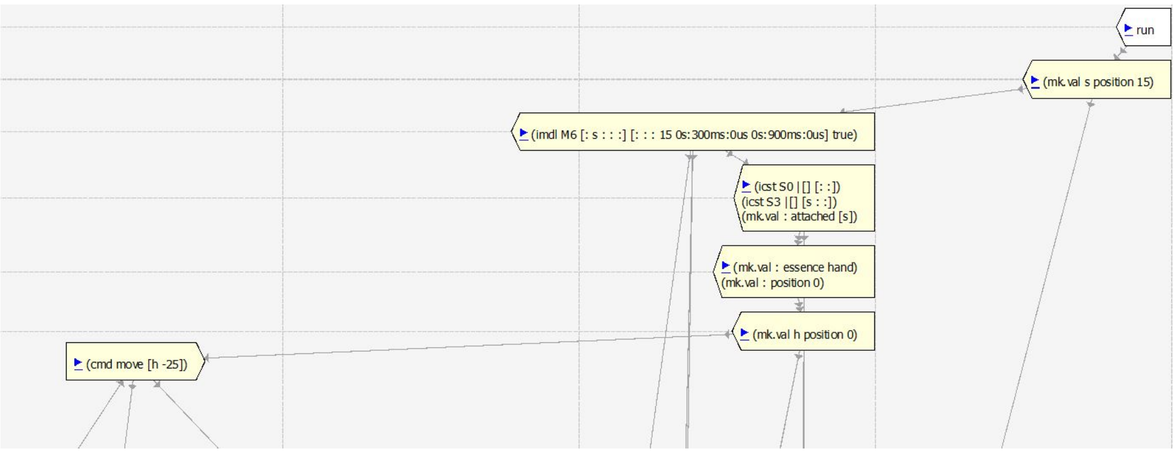


Figure 7.10: Simulation process in AERA

7.4 Knowledge representation in NARS

To explain NARS's knowledge representation and aquisition, one should understand the syntax of the Narsese language. The most basic form of a Narsese statement is as follows: *subject term — (inheritance relation) —> predicate term.* Example:

<robin –> bird>.

This example consists of four components to be explained: Firstly, the subject term robin, for which a relation is being defined. Second, the relation itself, in this case it is the so-called inheritance relation, the basic relation underlying NARS's system for knowledge representation. It refers to the fact that the term "robin" is specialization of the term "bird," and bird a generalization of robin. This could be read as "all robins are birds," or alternatively "robin is a type of bird." The predicate term "bird," then represents the concept being related to the subject [41].

There is one additional important component which we do not see in the example - the truth value. Statements in Narsese are never absolute, as any derived knowledge is non-axiomatic: liable to change and improvement with additional new information. Therefore the truth value, consisting of the two subcomponents "frequency" and "confidence," tells us two important things: Firstly with what frequency is the relation between subject and predicate expected to occur, and secondly how confident, or trustworthy is this information, a greater number of observed data points always leading to a greater confidence value.

There is a default setting for these values if the are not specified by the input, but they can also be included in any statement.

In NARS's first order reasoning, the forms of representation are extended by relation variants and compound terms [41]. The following are examples of statements that contain compound terms or a similarity relation:

```
<robin –> (|,bird,swimmer)>.
<robin <-> swan>.
```

The subject and relation parts of the first example should be familiar, we are defining an inheritance relation of a robin. The predicate is slightly more complex than before. The compound term |,bird,swimmer means either bird or swimmer. Therefore, we can state that a robin is either a bird or a swimmer.

The second example uses a different type of relation, the similarity relation. What this essentially means, is "robin is like a swan," meaning that they are likely to share the same relations, for example, if robin is a type of bird, then a swan is likely also a type of bird.

In NARS's higher order reasoning, statements can be used as terms, that is, there are statements of statements. The negation, conjunctions and disjunctions can be used in the creation of a complex statement.

```
< <robin –> bird> ==> <robin –> animal> >.
```

Here we have three statements, the first acting as the subject and the second as predicate in the third, more complex statement. This relational operator of the third statement is called implication, indicating that if the subject is true, the predicate is implied: "if robin is a bird, then robin is an animal".

7.5 How to Pose Questions to NARS

A question is considered a NARS 'task.' It can provide an explanation about different terms and compound terms. Assume that NARS receives the following evidence:

```
<gull –> swimmer>.
```

Now, the question of "what is a swimmer" can be asked in this way:

```
<? –> swimmer>?
```

This question instructs NARS to discover potential substitutes for the question mark, in other words, to answer the question "what is a swimmer?". Given the above evidence, it should promptly respond that a gull is a swimmer.

The question mark can be substituted for any term, or it can replace the dot (that usually indicates a "judgment" or statement of fact, telling NARS to process the input as new knowledge). Consider the following example:

```
<robin –> bird>.
<gull –> bird>.
<gull –> swimmer>.
<robin <-> swan>.
<swan –> swimmer>.
<robin –> swimmer>?
```

Now we are simply asking NARS to infer the truth value of the statement “robin is a swimmer.” The above evidence would result in NARS finding this to likely be the case, but with more evidence of non-swimming birds this should be rectified.

7.6 Causal Explanations in NARS

NARS can represent causal chains by event-based statements, if the occurrence of a subject term implies the occurrence of the predicate term. In case the first event is a cause event for the second event, the what question can provide a causal explanation for the second event. For example:

```
<E1 =/> E2>.
<? =/> E2>.
```

Here an observation of E1 followed by E2 would be positive evidence for the first statement, while observing E1 without E2 would be negative evidence.

7.7 Discussion

According to Halpern and Pearl [56], for a causal explanation one must consider two category of variables: endogenous and exogenous. Endogenous variables are involved in the transition from a cause to its effect, whereas exogenous ones specify the context in which the process occurs but are not a part of a causal relation. For instance, in the statement *lighting a match (LM) causes a forest fire (FF)*, FF and LM are endogenous variables, while the existence of oxygen (OX) which is an obvious precondition for the existence of the fire is exogenous.

The AERA system creates composite states which contain both exogenous and endogenous variables. Instantiation of composite states determines a precondition (specific context) for employment of causal models learned by the system. The cause and effect variables in AERA are not exogenous variables, which is a feature that allows generation of valid causal explanations [56].

Narsese statements, on the other hand, are composed of terms, and thus the statement *match is lit* is considered a compound term rather than a variable.

7.8 Summary

AERA’s models are causal and thus relate a causing set of variables, called a composite state (LHS) to another effect state (RHS). In this way, if a fact (either an observation or a simulated fact) matches the RHS of the model, the system can trace back to the potential causes and eventually select the most promising one. Thus, the AERA system is capable of producing causal explanations of observed events, simulated facts or taken actions by backward chaining (a.k.a abduction). Abduction in NARS is a process on **two** Narsese statements, but this can be expanded infinitely in theory, since statements can be nested. Also, inferring a subject from a predicate does not lead the system to a causal explanation of the predicate (TBD: add example), since a Narsese statement is not necessarily causal. As implied above, a causal explanation in NARS is only possible if the statement is about a succession of temporal events, where the first event (subject) causes the second event (predicate).

Chapter 8

Conclusions & Future Work

The main focus of research in the field of explainable AI is machine learning and interpretation of the input and output of complex models that are otherwise unreadable to humans. While the rise of this field of interpretable AI has yielded many benefits in our use and understanding of these models, we believe it is just the beginning of a much larger work with an inordinate amount of potential for improvements in AI.

The field of narrow AI has seen great advancement in automating simple tasks - tasks like classification that do not consist of many steps or layers of causal chains. We posit that automation capabilities need to be expanded, to encompass as much as possible of all tasks, for this is the purpose of AI. An important step in this process of expanding automation capabilities is the automation of explanation, as explanations are what allows us to learn and grow. Explaining machines would allow us to learn what they have learned, and perhaps equally importantly, what they have not learned. In the case of reflective architectures, it may even allow an AI to learn from its own successes and mistakes, improving its learning.

It is important here to define and distinguish valid explanations from insights and interpretations provided by shallow analysis or statistical evaluations. A valid explanations must be grounded by causal reasoning, providing sufficient information to justify the explanandum to the explainee.

In complex systems, where actions of agents and the influence of events and processes interact on a subtle level, it is necessary to model the correct cause and effect relationships relevant to the goals to be achieved, and this holds particularly true for the goal of explanation. Maintaining something like a directed graph of causal links would allow for both prediction and backtracking to complete goals, and selecting relevant nodes from which to construct explanations.

Researchers that have approached the concept of explanation from other fields, such as psychology and social sciences, agree that one of the most important parts of explanation is causal attribution [11]–[15].

From a purely pragmatic perspective, getting anything done in a complex dynamic world is virtually impossible without some knowledge of what leads to what—in other words, without knowledge of causes. For general machine intelligence, therefore, such knowledge is inescapable.

This work represents only a small beginning of a much larger research agenda that includes the classification of tasks, development of causal-based reasoning systems, and artificial general intelligence. Every tool has its set of applications, and we are only beginning to explore the tasks that can be performed by AI systems. In the future, we aim to explain more complex tasks and improve evaluations of AI agents, to determine what tasks which architecture is suitable for.

Chapter 9

Appendix A

Bibliography

[1] J. E. Bieger and K. R. Thórisson, "Requirements for general intelligence: A case study in trustworthy cumulative learning for air traffic control", in *Workshop on Architectures & Evaluation for Generality, Autonomy & Progress in AI International Joint Conference on Artificial Intelligence, Stockholm, July 15*, 2018.

[2] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?", *CoRR*, vol. abs/1712.09923, 2017. arXiv: `1712.09923`. [Online]. Available: `http://arxiv.org/abs/1712.09923`.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "?why should i trust you??: Explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ?16, San Francisco, California, USA: Association for Computing Machinery, 2016, 1135?1144, ISBN: 9781450342322. DOI: `10.1145/2939672.2939778`. [Online]. Available: `https://doi.org/10.1145/2939672.2939778`.

[4] G. Montavon, W. Samek, and K. Müller, "Methods for interpreting and understanding deep neural networks", *CoRR*, vol. abs/1706.07979, 2017. arXiv: `1706.07979`. [Online]. Available: `http://arxiv.org/abs/1706.07979`.

[5] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "The emerging landscape of explainable ai planning and decision making", *ArXiv*, vol. abs/2002.11697, 2020.

[6] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)", *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[7] K. R. Thórisson, "The 'explanation hypothesis' in general self-supervised learning", *Proceedings of Machine Learning Research*, vol. 159, pp. 5–27, 2021.

[8] S. Palacio, A. Lucieri, M. Munir, J. Hees, S. Ahmed, and A. Dengel, *Xai handbook: Towards a unified framework for explainable ai*, 2021. arXiv: `2105.06677 [cs.AI]`.

[9] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn", *Nature Communications*, vol. 10, no. 1, 2019, ISSN: 2041-1723. DOI: `10.1038/s41467-019-08987-4`. [Online]. Available: `http://dx.doi.org/10.1038/s41467-019-08987-4`.

[10] M. R. Cohen, *International Journal of Ethics*, vol. 43, no. 2, pp. 220–226, 1933, ISSN: 1526422X. [Online]. Available: `http://www.jstor.org/stable/2378336` (visited on 06/02/2022).

[11] T. Miller, *Explanation in artificial intelligence: Insights from the social sciences*, 2017. arXiv: `1706.07269 [cs.AI]`.

[12] J. Woodward, *Making things happen: A theory of causal explanation.* Oxford university press, 2005.

[13] T. Lombrozo, "The structure and function of explanations", *Trends in Cognitive Sciences*, vol. 10, no. 10, pp. 464 –470, 2006, ISSN: 1364-6613. DOI: `https://doi.org/10.1016/j.tics.2006.08.004`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S1364661306002117`.

[14] M. Strevens, "The causal and unification approaches to explanation unified-causally", *Noûs*, vol. 38, no. 1, pp. 154–176, 2004.

[15] J. Y. Halpern and J. Pearl, "Causes and explanations: A structural-model approach, part I: causes", *CoRR*, vol. cs.AI/0011012, 2000. [Online]. Available: `https://arxiv.org/abs/cs/0011012`.

[16] J. Cohen, "Teleological explanation", *Proceedings of the Aristotelian Society*, vol. 51, pp. 255–292, 1950, ISSN: 00667374, 14679264. [Online]. Available: `http://www.jstor.org/stable/4544486` (visited on 06/04/2022).

[17] A. Ney, "Reductionism", *Internet Encyclopedia of Philosophy IEP*, 2008. [Online]. Available: `https://www.iep.utm.edu/red-ism/`.

[18] J. Josephson and S. Josephson, *Abductive Inference: Computation, Philosophy, Technology*, ser. Computation, Philosophy, Technology. Cambridge University Press, 1996, ISBN: 978-0-521-57545-4. [Online]. Available: `https://books.google.is/books?id=uu6zXrogwWAC`.

[19] K. R. Thórisson, D. Kremelberg, B. R. Steunebrink, and E. Nivel, "About understanding", in *The Proceedings of the Ninth Conference on Artificial General Intelligence*, 2016, pp. 106–117.

[20] J. Pearl, "Bayesianism and causality, or, why i am only a half-bayesian", in *Foundations of bayesianism*, Springer, 2001, pp. 19–36.

[21] D. Doran, S. Schulz, and T. R. Besold, "What does explainable ai really mean? a new conceptualization of perspectives", *arXiv preprint arXiv:1710.00794*, 2017.

[22] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai", *Information fusion*, vol. 58, pp. 82–115, 2020.

[23] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)*, 2017. arXiv: `1711.11279 [stat.ML]`.

[24] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation", *PLOS ONE*, vol. 10, no. 7, pp. 1–46, Jul. 2015. DOI: `10.1371/journal.pone.0130140`. [Online]. Available: `https://doi.org/10.1371/journal.pone.0130140`.

[25] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks", *Digital Signal Processing*, vol. 73, pp. 1 –15, 2018, ISSN: 1051-2004. DOI: `https://doi.org/10.1016/j.dsp.2017.10.011`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S1051200417302385`.

[26] Y.-h. Sheu, "Illuminating the black box: Interpreting deep neural network models for psychiatric research", *Frontiers in Psychiatry*, p. 1091, 2020.

[27] W. Samek, T. Wiegand, and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models", *CoRR*, vol. abs/1708.08296, 2017. arXiv: `1708.08296`. [Online]. Available: `http://arxiv.org/abs/1708.08296`.

[28] K. R. Thórisson, D. Kremelberg, B. R. Steunebrink, and E. Nivel, "About understanding", in *Artificial General Intelligence*, B. Steunebrink, P. Wang, and B. Goertzel, Eds., Cham: Springer International Publishing, 2016, pp. 106–117, ISBN: 978-3-319-41649-6.

[29] A. Nguyen, J. Yosinski, and J. Clune, "Understanding neural networks via feature visualization: A survey", *CoRR*, vol. abs/1904.08939, 2019. arXiv: `1904.08939`. [Online]. Available: `http://arxiv.org/abs/1904.08939`.

[30] K. Simonyan, A. Vedaldi, and A. Zisserman, *Deep inside convolutional networks: Visualising image classification models and saliency maps*, 2013. arXiv: `1312.6034 [cs.CV]`.

[31] M. T. Keane and B. Smyth, "Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai)", in *International Conference on Case-Based Reasoning*, Springer, 2020, pp. 163–178.

[32] R. M. Byrne, "Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning.", in *IJCAI*, 2019, pp. 6276–6282.

[33] I. Stepin, J. M. Alonso, A. Catala, and P.-F. Martín, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence", *IEEE Access*, vol. 9, pp. 11 974–12 001, 2021. DOI: `10.1109/ACCESS.2021.3051315`.

[34] K. Sokol and P. A. Flach, "Counterfactual explanations of machine learning predictions: Opportunities and challenges for ai safety", in *SafeAI@ AAAI*, 2019.

[35] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness", *arXiv preprint arXiv:1703.06856*, 2017.

[36] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.", *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[37] J. E. Laird, *The Soar cognitive architecture.* MIT press, 2012, ISBN: 0262122960, 9780262122962.

[38] W. L. Johnson, "Agents that learn to explain themselves.", in *AAAI*, 1994, pp. 1257–1263.

[39] S. Franklin and F. Patterson Jr, "The lida architecture: Adding new modes of learning to an intelligent, autonomous, software agent", *pat*, vol. 703, pp. 764–1004, 2006.

[40] E. Nivel, K. R. Thórisson, H Dindo, G Pezzulo, M Rodriguez, *et al.*, "Autocatalytic endogenous reflective architecture", *RUTR SCS13002*, 2013.

[41] P. Wang, "The logic of intelligence", in *Artificial general intelligence*, Springer, 2007, pp. 31–62.

[42] J. Y. Halpern and J. Pearl, "Causes and explanations: A structural-model approach. part II: explanations", *CoRR*, vol. cs.AI/0208034, 2002. [Online]. Available: `https://arxiv.org/abs/cs/0208034`.

[43] D. J. Hilton and B. R. Slugoski, "Knowledge-based causal attribution: The abnormal conditions focus model.", *Psychological Review*, vol. 93, no. 1, pp. 75–88, 1986. DOI: `10.1037/0033-295X.93.1.75`. [Online]. Available: `https://doi.org/10.1037/0033-295X.93.1.75`.

[44] D. J. Hilton, "Conversational processes and causal explanation.", en, *Psychological Bulletin*, vol. 107, no. 1, pp. 65–81, 1990, ISSN: 0033-2909. DOI: `10.1037/0033-2909.107.1.65`. [Online]. Available: `http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.107.1.65` (visited on 08/12/2020).

[45] P. Lipton, "Contrastive explanation", *Royal Institute of Philosophy Supplements*, vol. 27, pp. 247–266, 1990.

[46] M. L. Ginsberg, *Counterfactuals.* Stanford, California: Stanford University, 1984.

[47] L. Eberding, M. Belenchia, A. Sheikhlar, and K. R. Thórisson, "About the intricacy of tasks", in *The Proceedings of the International Conference on Artificial General Intelligence*, 2021, pp. 65–74.

[48] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd. New York, NY, USA: Cambridge University Press, 2009, ISBN: 052189560X, 9780521895606.

[49] S. Franklin, T. Madl, S. D'mello, and J. Snaider, "Lida: A systems-level architecture for cognition, emotion, and learning", *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 1, pp. 19–41, 2013.

[50] P. Langley, J. E. Laird, and S. Rogers, "Cognitive architectures: Research issues and challenges", *Cognitive Systems Research*, vol. 10, no. 2, pp. 141–160, 2009.

[51] P. Wang, "On defining artificial intelligence", *Journal of Artificial General Intelligence*, vol. 10, no. 2, pp. 1–37, 2019.

[52] E. Nivel and K. R. Thórisson, "Towards a programming paradigm for control systems with high levels of existential autonomy", in *International Conference on Artificial General Intelligence*, Springer, 2013, pp. 78–87.

[53] K. R. Thórisson and E. Nivel, "Achieving artificial general intelligence through peewee granularity", in *The Proceedings of the Second Conference on Artificial General Intelligence*, 2009, pp. 222–223.

[54] K. R. Thórisson, E. Nivel, B. R. Steunebrink, H. P. Helgason, G. Pezzulo, R. Sanz, J. Schmidhuber, H. Dindo, M. Rodriguez, A. Chella, G. K. Jonsson, D. Ognibene, and C. Hernandez, "Autonomous Acquisition of Situated Natural Communication", *Computer Science & Information Systems*, vol. 9, no. 2, pp. 115–131, 2014, Outstanding Paper Award. (visited on 10/26/2014).

[55] J. Thompson and K. R. Thórisson, "Demonstrating model-based learning, prediction, and goal achievement in AERA", Presented at Second International Workshop on Self-Supervised Learning, 2021.

[56] J. Y. Halpern and J. Pearl, “Causes and explanations: A structural-model approach — part 1: Causes”, *CoRR*, vol. abs/1301.2275, 2013. arXiv: `1301.2275`. [Online]. Available: `http://arxiv.org/abs/1301.2275`.