Chapter 1: Introduction

"The complexities of cause and effect defy analysis."

— Douglas Adams

The goal of this research is to investigate and demonstrate the suitability of action languages and answer set programming (ASP) to design and realize a novel framework for reasoning about and explaining *actual causation*. Also referred to as causation in fact, *actual cause* is a broad term that encompasses all possible antecedents that have played a meaningful role in producing a consequence [1]. Reasoning about actual cause concerns determining how a particular consequence came to be in a given scenario, and the topic been studied extensively in numerous fields including philosophy, law, and, more recently, computer science and artificial intelligence.

Attempts to mathematically characterize actual causation have largely pursued counterfactual analysis of structural equations (e.g., [2–6]), neuron diagrams [7, 8], and other logical formalisms (see e.g., [9]). Counterfactual accounts of actual causation are inspired by the human intuition that if X caused Y, then not Y if not X [10]. At a high level, this approach involves looking for *possible worlds* in which Y is true and X is not. If such a world is found, then X is not a cause of Y. It has been widely documented, however, that the counterfactual criteria alone is problematic and fails to recognize causation in a number of common cases such as overdetermination (i.e., multiple causes for the effect), preemption (i.e., one cause "blocks" another's effect), and contributory causation (i.e., causes must occur together to achieve the effect) [11, 12]. Subsequent approaches have addressed some of the shortcomings associated with the counterfactual criterion by modifying the existing definitions [13, 14], introducing supplemental definitions [9, 15, 16], and by modeling time [17] with some improved results. In spite of these improvements, there is still no widely accepted definition of actual cause.

Departing from the counterfactual intuition and reasoning about possible worlds, our framework favors reasoning about the underlying causal mechanisms of the scenario itself in order to

explain actual causation of an outcome of interest. Our framework uses techniques from Reasoning about Actions and Change (RAC) to support reasoning about domains that change over time in response to a sequence of events. The action language $\mathcal{AL}$ [18] enables us to represent a scenario as the evolution of state over the course of the scenario's events. Moreover, the elements of $\mathcal{AL}$ semantics can be used to define notions of direct and indirect cause, and the language's solution to the frame problem can be leveraged to detect the "appearance" of an outcome of interest in a scenario. Our position on reasoning about actual cause is supported by recent work in the area [19, 20] that shares similar intuition to our own about the appearance of outcomes. Finally, $\mathcal{AL}$ lends itself naturally to an automated translation in ASP, using which, reasoning tasks of considerable complexity can be specified and executed.

1.1 Explaining Actual Causation

Sophisticated actual causal reasoning has long been prevalent in human society and continues to have an undeniable impact on the advancement of science, technology, medicine, and other fields that are critical to the success of modern society. From the development of ancient tools to modern root cause analysis in business and industry, reasoning about causal influence in a historical sequence of events enables us to diagnose the cause of an outcome of interest and gives us insight into how to bring about, or even prevent, similar outcomes in future scenarios. Consider problems such as explaining the occurrence of a set of suspicious observations in a network security system, reasoning about the efficiency of actions taken in an emergency evacuation scenario, or investigating how an automatically generated workflow produced some unexpected results. It is easy to imagine that analyzing such (potentially very complex) scenarios requires the ability to represent and reason about how the state of the world has been changed by the scenario's events to produce some outcome of interest.

Consider the behavior of an advanced cyber-physical system such as a self-driving car, reasoning about causation (e.g., blame or praise) becomes significantly more complex – the car likely contains a large number of software and hardware modules (possibly from different vendors), there may be other cars and pedestrians involved in the scenario of interest, and there may have been

wireless communication with other vehicles or a central server, all of which may influence the actions taken by the car's control module over the course of its drive. To reach an intuitively satisfactory explanation of why some outcome of interest came to be in such a domain, the insights that have been produced by the decades-long study of actual causation seem indispensable.

Modern work on actual causation originated in philosophy with the seminal paper by Lewis [10]. His work, like that of other philosophers following him, was primarily theoretical and not intended to be put to practical use. The famous Halpern-Pearl (HP) paper [3] initiated interest in this concept within the field of AI and it constitutes a first milestone on the way towards applications of the concept of actual causation. However, neither the HP paper nor the many that have followed it (e.g., see Chapter 7) have yet reached the point where their results could be directly applied, for example, in the context of a self-driving car as proposed here, or in other similarly complex scenarios. We believe that this is due, at least in part, to a lack of distinction between the laws that govern individual states of the world and events whose occurrence cause state to evolve.

We also believe that our approach to reasoning about *what actually happened* rather than *what could have happened* sets us apart from the majority of the work in this research area (discussed in greater detail in Chapter 7), and our choice of $\mathcal{AL}$ as a formalism positions the approach for potential use in practical settings via translation to an implementation in ASP.

The primary contributions of this dissertation are as follows:

1. A novel theoretical framework for reasoning about actual causation in terms of the semantics of the action language $\mathcal{AL}$.

2. A sound and complete implementation of the framework in ASP.

3. Empirical studies of the practical feasibility of the implementation on increasingly large and complex novel causal scenarios, with respect to time.

1.2 Organization

We now present the organization of the dissertation.

- Chapter 2 contains background information about the environments of interest for which the framework is defined, the action language $\mathcal{AL}$, and ASP.

- Chapter 3 presents the theoretical framework for explaining actual causation. The chapter contains a novel running example which we use to aid our discussion of the framework's definitions. In addition to presenting the definition of direct cause, we present a simple notion of indirect causation and identify two important drawbacks to the definition. Next, we present an improved definition that addresses the identified shortcomings and provides additional information about indirect causes. Finally, we draw initial conclusions about the framework's ability to handle traditionally challenging cases of causation.

- Chapter 4, we use the framework to reason about examples from the literature that have been used to challenge the counterfactual definition of actual cause, as well as a novel example inspired by the self-driving car scenario outlined above.

- Chapter 5 provides implementations of the theoretical framework in ASP and presents theoretical results about soundness and completeness of the program for computing direct cause and the improved definition of indirect cause. We also present ASP translations of a subset of the examples from Chapter 4.

- Chapter 6 presents experimental results from empirical studies aiming to assess the practical feasibility of the approach with respect to time. To the best of our knowledge, there is no established set of benchmarks for the type of reasoning presented in this dissertation, and so we have generated a set of novel problem instances that allow us to examine and make initial conclusions about the performance of the implementation on a number of interesting and increasingly complex causal scenarios.

- Chapter 7 provides in-depth discussions about related approaches. We will present an overview of technical approaches and comparative discussion for the most well-known and widely studied approach in the field [3], the work that led us to our intuition about representing scenarios as the evolution of state in response to events [21], and the work that leverages

similar intuition to our own in reasoning about scenarios [19].

- Chapter 8 presents our conclusions and suggests avenues for future work.