Computation of Context as a Cognitive Tool

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Manon J. Sanscartier

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

In the field of cognitive science, as well as the area of Artificial Intelligence (AI), the role of context has been investigated in many forms, and for many purposes. It is clear in both areas that consideration of contextual information is important. However, the significance of context has not been emphasized in the Bayesian networks literature. We suggest that consideration of context is necessary for acquiring knowledge about a situation and for refining current representational models that are potentially erroneous due to hidden independencies in the data.

In this thesis, we make several contributions towards the automation of contextual consideration by discovering useful contexts from probability distributions. We show how context-specific independencies in Bayesian networks and discovery algorithms, traditionally used for efficient probabilistic inference can contribute to the identification of contexts, and in turn can provide insight on otherwise puzzling situations. Also, consideration of context can help clarify otherwise counter intuitive puzzles, such as those that result in instances of Simpson's paradox. In the social sciences, the branch of attribution theory is context-sensitive. We suggest a method to distinguish between *dispositional causes* and *situational factors* by means of contextual models. Finally, we address the work of Cheng and Novick dealing with causal attribution by human adults. Their *probabilistic contrast model* makes use of contextual information, called *focal sets*, that must be determined by a human expert. We suggest a method for discovering complete focal sets from probabilistic distributions, without the human expert.

Acknowledgements

For giving me permission to pursue my interdisciplinary aspirations, and trusting my intuition and results in the area of cognition, I would like to express my genuine gratitude to my supervisor Dr. Eric Neufeld. He took me on as a student knowing I had very particular research interests that I was unwilling to give up, and offered some starting points in the areas that interested me. For that, I am thankful. He was also a great source of encouragement at the start of my program, when my confidence in my abilities often lacked.

I would also like to thank Dr. Julita Vassileva, Dr. Winfred Grassmann, and Dr. Anthony Kusalik, for they have accepted to be on my thesis committee. Their astute comments, insights, and advice have helped me improve my thesis. I would also like to thank the cognate member, Dr. Ivan Kelly, and the external examiner, Dr. Choh Man Teng, for having read my thesis so carefully, for offering some great suggestions, and for showing some enthusiasm for the topic. I feel my thesis is a more complete ensemble because of their input.

I am thankful to the Faculty of Graduate Studies and Research as well as the Department of Computer Science for the financial support they have provided for the first 16 months of my Ph.D. studies, which allowed me to concentrate on my research. I am also thankful to the Natural Sciences and Engineering Research Council (NSERC) for awarding me a PGS-D3 scholarship, thus relieving me from TA and marking duties, allowing me to put more efforts towards my research, lecturing classes, sitting on several committees, and occupying a large role with the Computer Science Graduate Course Council (CSGCC).

Finally, I would like to thank the people around me who have contributed to making this experience enjoyable. I would like to thank the people in the Department of Computer Science, my friends, and my family. I would especially like to thank my partner Kevin, who has offered more support than I could even give myself at times, and for refusing to believe me when I said I truly, genuinely wanted to quit on more than one occasion $\ddot\smile$. He believed in me throughout, and helped me to the finish line! Finally, I would like to thank my dear friend Mike for standing by my side the entire length of my degree, offering phenomenal advice on every component of the degree, be it the learning, the teaching, the research, the handling of seemingly impossible situations, and the challenge of remembering throughout that being a student doesn't completely overwrite the fact that one is also a person.

Wiggles Biggles

Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
BN	Bayesian Network
CI	Conditional Independence
CPD	Conditional Probability Distribution
CSI	Context-Specific Independence
CWI	Contextual Weak Independence
DAG	Directed Acyclic Graph
HMM	Hidden Markov Model
JPD	Joint Probability Distribution
KR	Knowledge Representation
MPD	Marginal Probability Distribution

Chapter 1

Introduction and Motivation

Making a decision, solving a problem, or attributing a cause to an effect without first considering the context in which it takes place, is like reading a book with missing pages; your conclusions may be erroneous due to missing information. A classic example by Saxe [1] highlights the importance of context in the cognitive science literature. Ten Brazilian boys sell candy to passers-by. They have no formal education; they use an inflated monetary system, which makes mathematical manipulation more complicated, and they need to have a sophisticated understanding of mathematics and ratios. They reason well with ratios and rebates, such as selling one box of candy for 20,000 cruzeiros, 2 candy bars for 500, 5 bars for 1,000, etc. However, when required to access the same mathematical skills on a standardized math test, they perform poorly. Problem solvers operate in their natural environment. The same goes with ordinary people purchasing olives at the grocery store. An experiment by Kirshner and Whitson [2] showed that ordinary people succeed in figuring out which brand of olives is cheaper by volume in a grocery store, even though they would fail to understand the same type of problem on a standardized math test.

Although the role of context has been studied and recognized as being important in the cognitive science and Artificial Intelligence literature, the semantic role of context does not appear to have played a significant role in either the Bayesian networks literature or the cognitive science literature. The literature on context does not focus on acquiring knowledge about a situation and refining current representational models that are potentially erroneous due to hidden independencies in probabilistic distributions. On the other hand, in the Knowledge Representation literature (KR), context has been exploited to find solutions to problems in categorical reasoning. The same types of problems have been addressed in belief revision and in the study of reference classes, where the reference class refers to the context. In uncertain reasoning, formal methods for expressing contextual information in probabilistic independencies have been investigated [3, 4]. Methods of inference to take advantage of contextual data for faster query processing have been presented in the literature [5, 6]. Also, to improve inference with context, techniques for the discovery of context have been offered [7, 8].

On the cognitive side, context has been studied in everyday learning. Recent studies [9, 10, 11, 12, 13, 14, 15, 16, 17, 18] include knowledge transfer, mental reasoning about causal relations, probabilistic reasoning by children, language processing, and attribution.

In this thesis, we contribute mainly to the automation of contextual consideration by discovering useful contexts from probabilistic distributions. With the discovered particularities about subsets of the available information, we build contextual models that better represent the causal relations hidden in a particular situation, which facilitates better decision making, problem solving, and attribution. We also show how the algorithmic aspects of context suggest that it may be what helps human reasoners be efficient in the context of vast knowledge bases. We provide a treatment of context in the setting of BNs that gives a useful account of AI and attribution problems by using context to refine cognitive causal models. We contribute towards a cognitive account of context, largely by linking context-specific independence (CSI) with the idea of focal sets from Cheng's [19, 20, 21] work, which provides an existing model for attribution of causal judgment by human adults. Finally, we show how contextual discoveries can be used to provide a computational mechanism that can discover focal sets in data.

In the following two sections, we present an overview of previous work in Artificial Intelligence and in cognitive science that have dealt with the role of context. Although the importance of context is recognized in this previous research, the discovery of context from probability distributions, and the potential usefulness of building contextual models has not been addressed.

1.1 Previous Work Dealing with the Role of Contextual Information

In this section, we provide a brief overview of contextual consideration in the Artificial Intelligence and cognitive science literature.

1.1.1 Context and Artificial Intelligence

Even though context is widely used in a variety of AI settings, it has not received a consistent treatment, or understanding, in AI. A striking example is the idea of Hidden Markov Models (HMM) [22], used widely in speech, text, and character recognition. Recognizing symbols, strings, or sounds in isolation is almost impossible. A vertical stroke might be a number one, a small 'l', a capital 'i', a conditioning bar, and so on. However, a small amount of context, for example, knowing the neighbouring character, greatly improves recognition algorithms at a small cost. Another example is that of reactive planners [23, 24]. Traditional AI planners did not scale up well in real world settings, possibly because of the enormous quantities of knowledge required to maneuvre through even simple domains. Reactive planners took a different approach. Rather than having to know the entire environment in advance, they continuously monitored the environment and changed plans whenever the situation deemed it appropriate, similar to control devices in engineering.

Knowledge Representation (KR) is one sub area of AI where context has been studied. Reiter [25] was one of the first mainstream Artificial Intelligence researchers to observe the role of context in categorical reasoning. He observed that people are able to maintain knowledge bases similar to the following:

Birds fly.

Penguins are birds and don't fly.

Tweety is an penguin.

Represented as sentences of logic, the above sentences are inconsistent if any penguins exist. Not only do humans maintain such apparently inconsistent databases, they reason with them effectively. In the context of Tweety being a bird, humans conclude that Tweety can fly, and in the context of Tweety being a penguin, humans conclude instead that Tweety can't fly. Reiter called this type of reasoning *default logic*, and modeled it with a knowledge base consisting of a set of logical sentences (sentences that were always true), a set of defaults (sentences that looked like logical sentences but were only typically true, i.e. in most contexts), and a set of default inference rules. This development attracted the interest of many researchers, and many variations were produced including nonmonotonic logic [26], predicate circumscription [27], defeasible logic [28], and Theorist [29].

Whereas traditional logic bases are monotonic, that is, truth values of existing propositions do not change with the receipt of new knowledge, default logic bases are nonmonotonic, so the value of predicates, for example, $fly(Tweety)$, can change as knowledge is added. Most of these researchers, however, did not carefully formalize the notion of context as a knowledge state, that is, where specific knowledge might affect other pieces of knowledge, e.g. generalizations.

The problem of Knowledge Representation has also been studied by Gardenfors et al. [30], who studied belief revision. They observed that new knowledge may change the truth values of certain predicates. However, different researchers provided different arguments as to the extent of belief revision concomitant to receipt of new information.

Some uncertain reasoning mechanisms have a better formalization of context. Traditional conditional probability can be viewed as a context-based inference mechanism. For example, the conditional probabilities $p(fly|bird)$ and $p(fly|penguin)$ can be assigned probability values almost independently. If we only know bird to be true, we select the first probability value, $p(fly|bird)$, to determine our belief in fly, and if we know *penguin* to be true, we select the second probability value, $p(fly|penguin)$. The quantity to the right of the conditioning bar is the context, or reference class, and we use our knowledge about a current situation to index into the correct class. Reichenbach [31] provided a succinct explanation of why we make this choice. He argued that given a body of generic knowledge (mostly probability statements), and some context, we make an inference in our context using the narrowest reference class for which we have adequate statistics. Kyburg [32] expanded greatly on this idea. His work elaborated on this, but to use the simpler cases we should prefer an inference based on a narrower class. This narrower class is more likely to account for exceptions that get averaged away in the larger class. However, as reference classes get smaller, sample sizes get smaller and statistics lose power. The idea of adequacy addresses the tradeoff between statistical power and narrowness in the context of interval-based probabilities. However, in the case where probabilities are point valued, we can always choose the narrowest reference class, matching our context, for which probabilities are known.

Kyburg's theory supports intuitive reasoning for preferring context. In Kyburg's formalism, context only changes the inferred probability values. In ours, context may change the entire model of a scenario. While KR seeks to find representations and modeling methods with wide applicability and great inferential power, our usage of context suggests that models are relatively small and may change from one setting to another.

Context has also been studied formally in uncertain reasoning with Bayesian networks. Pearl's [33] work has made the storing of a joint probability distribution (JPD) for a large data set unnecessary since all probabilistic conditional independencies (CIs) in the Bayesian network allow for the JPD to be stored in smaller conditional probability distributions (CPDs), each representing a portion of the network. CI allows for those smaller portions of the network to be multiplied together without loss of information and without having invalid information penetrate the network by the multiplication of distributions. Boutilier et al. [3] generalized the idea of CI to achieve even smaller CPDs and to reduce the number of multiplications and additions required. Based on this idea, inference was improved [5] by using context-specific independence (CSI), a generalized form of CI.

A Bayesian network is built from known CIs. However, when it comes to the more general CSI, we must discover them to use them. In previous research, two discovery methods have been proposed to discover CSI and therefore increase query processing speed [5, 7]. Due to the inferencial benefits of CSI, it has been generalized even further, to *contextual weak independencies* (CWI) [4], where inference methods [6], and discovery methods [8] are similar to those of CSI.

Largely, research in AI related to the idea of context has not addressed the semantics or cognitive significance of context. In the present research, we use CSI and CSI discovery methods to study the meaning of contextual independencies, rather than the well-known algorithmic portion and the inferencial benefits generalized contextual independencies provide [5, 7, 4, 6, 34, 35].

1.1.2 Context and Cognition

Context appears in the recent cognitive science literature, but, as in the AI literature, in a variety of ways. In the cognitive science literature, context has been used in many situations as a theoretical tool where consideration of the environment is believed to have an impact on the way people learn.

In a very recent study [9], Wagner offers a new explanation of knowledge transfer that is highly context-sensitive. His contextual variation emphasizes the importance of context consideration in building mental representations of situations. He conducted a case study analysis of an undergraduate student's strategies for solving a variety of problems for different domain applications, but masked a sole principle of elementary statistics. The study indicated that the student initially thought the problems were very different from each other, but slowly identified the problems as different instances of a basic statistical concept. Once that context was established, the student was much more successful at solving the problems. However, Wagner's theory does not offer methods to refine existing representations of situations, but rather demonstrates how solving problems that may seem completely different at first glance may become similar in terms of solving methodology once context is taken into account. This result supports the validity of a consideration of context in problem solving.

Recent work about the use of probabilities in mental reasoning about causal structures reveals that the focus (which can be viewed as the context) of information may be more important than the size of the data set [16]. When testing human causal attribution, it is common to use verbal vignettes with covariation information included in the verbal description of the situation. However, it is typically believed that the amount of information given affects attribution directly. In Majid et al.'s study [16] consisting of four experiments, they show that focus plays an important role in attribution and that there seems to be a confound between quality and size of the information. That is, when presenting subjects with information about a situation for which they need to attribute a cause, providing the subject with fewer but more related facts, may be more effective than a large amount of perhaps unrelated information for an accurate attribution.

Also in the realm of human computation of probabilities, Teigen and Keren [17] have studied the *surprise* effect in terms of outcome expectations. In general, surprises are created by low probability outcomes simply because the lower the probability of occurence, the more surprising it is if the event actually occurs. However, the authors believe that not all low probability outcomes are equally surprising. They propose a *contrast hypothesis* to investigate the surprise associated with an outcome. The hypothesized belief is that the level of surprise corresponding to an outcome is primarily determined by how much it contrasts with the more expected alternative. The results suggest that high contrast between outcomes is highly associated with high surprise rates. In addition, different categories of contrasts were tested against the expected alternative. The categories consisted of factors such as contrasts formed by novelty and change, contrasts due to relative probabilities, and contrasts due to perceptual or conceptual distance between the expected and the obtained outcome. This categorization suggests, once again, the importance of context in predicting outcomes.

Another cognitive study dealing with reasoning with probabilities was conducted by Zhu and Gigerenzer [12]. They studied the use of probabilities by older children with an implicit consideration of context. In their paper, Zhu and Gigerenzer argue that children can reason with probabilistic information in certain contexts only. They reported in their experiments that when information was presented to grade four, five, and six children as actual normalized probabilities, all children were unable to estimate posterior probabilities. However, when the same information was presented in natural frequencies, the reasoning was significantly more successful, and showed a steady increase from grade four to grade six (19, 39, and 53% successful respectively), which is a much more intuitive observation. Without a consideration of context on the representation chosen as input, the conclusion would likely have been that children simply cannot use probabilistic information.

A similar conclusion might be drawn in language processing. Kaiser and Trueswell [11] discuss the role of context in verbal word recognition. They claim that the context in which listeners hear certain words will change the speed of recognition of the word. The more appropriate the word is for the context, the faster the subject will recognize it, e.g., before the experimentor finishes uttering the word. In a different experiment by Treiman, Kessler, and Bick [14], the context of consonants was studied in an investigation on pronunciation. The authors found that pronunciation of vowels in nonwords is dependent upon the consonants surrounding the particular vowel. This suggests that a study aimed at understanding solely the role of context may provide clues about possible trends in pronunciation, which may be beneficial in applications such as speech pathology. Finally, a recent study [10] suggests that context plays a crucial role in syntactic processing. Results show that theory based on context not only applies to resolving ambiguity, but also in processing unambiguous sentences. This idea is strongly related to the use of Hidden Markov Models (HMMs) for speech, text, and character recognition [22].

In this section, we have outlined experiments conducted in different fields of cognition, where an explicit consideration of context has improved the overall conclusions about a particular situation. In knowledge transfer, context can help us recognize when seemingly different mathematical problems can be solved similarly. Also, in probabilistic reasoning about causal structures, we saw that the focus, or context, can help the reasoner make better conclusions the validity of causal statements. Also in the realm of probabilities, we saw how the representation of numbers may dictate whether or not children understand a problem. Finally, in language processing, the context in which a symbol is located can have a large impact on the particular character's recognition. This leads directly into a discussion of our current consideration of context in the thesis.

1.2 Present Consideration of Context

The objectives of this thesis have materialized in several forms in the present research. For ease of understanding, we divide the work into three themes. The first deals with AI and problem solving. We address the tools in AI that we use to discover context in probabilistic distributions.

We also show how problem solving can be greatly improved by using discovered contexts to address particular situations. We show how context consideration from probabilistic distributions can alleviate erroneous inferences, and in the extreme case, avoid instances of Simpson's paradox.

The second theme addresses issues in attribution theory. We discuss existing theories of attribution, and corresponding attributional biases meant to deal with *exceptions*. By exceptions, we mean situations where the conclusions go against the outcome the theory would predict. We suggest a contextual consideration that could potentially eliminate the need for biases, since the answers to many of those seeming exceptions are in particularities about a specific situation. The complete elimination of bias is out of the scope of this thesis but for the moment, we suggest a method of distinguishing between *dispositional causes* and *situational factors* by means of contextual models. Dispositional causes and situational factors are further discussed in Chapter 3.

The last theme deals with an existing model for causal attribution by human adults developed by Cheng and Novick [19, 20, 21]. Their *probabilistic contrast model* makes use of contextual information in the form of *focal sets*, contextually selected sets of events over which covariation is computed. In the probabilistic contrast model, the focal sets must be determined by a human expert. We suggest a method for discovering complete focal sets from probability distributions by considering each element making up a focal set separately.

The next three subsections give a brief overview of these three themes.

1.2.1 Tools in Artificial Intelligence and Problem Solving with Contextual Consideration

Bayesian networks [33] are a widely used tool used for uncertain reasoning in AI. They facilitate the indirect acquisition of the joint probability distribution due to *conditional independence* (CI) assumptions. From an AI view point, this allows for a more compact representation of the probability distribution and makes the inference process feasible in some applications. However the notion of conditional independence is too restrictive to capture independencies that only hold in certain contexts. A generalization of CI, namely *context-specific independence* (CSI) [3] has been formalized to allow more efficient inference in query processing. CSI allows us to decompose a distributions where an independence holds only in certain contexts, and not for all subsets of values of a variable. This decomposition results in having to specify fewer values when building a probabilistic distribution.

Throughout the development of CSI, no emphasis was put on the natural "grouping" of the values of variables, or on the semantic content of the distributions where CSIs are found, why these CSIs hold in some cases, etc. The main focus has been on complexity reduction, achieving fewer variables per distribution and requiring fewer operations for inference. Once the issues of *representation* of CSI, *inference* with CSI, and *discovery* of CSI were answered, research focused on further generalizations of CSI; ways for making distributions even more compact, and inference even faster were investigated. One such generalization is contextual weak independence (CWI) [4]. CWI is an even more generalized version of CSI, where context is found within a CSI. It is a further generalization of context for more specific subgroups of data.

In the quest for faster inference, context-specific independence (CSI) was merely a step forward, and thus has not been studied for purposes of human reasoning or for investigating how knowledge of CSI may shed light on the situation being modeled. Instead, by taking advantage of probabilistic conditional independence, an indirect representation of the *joint probability distribution* (JPD) was possible with *Bayesian networks* (BNs). With this new sound formalism to obtain the JPD, the uncertain reasoning community investigated generalizations of this indirect representation to achieve better and faster inference and more compact representations of the JPD. CSI fulfilled this goal of providing improved query processing with more efficient inference and a more compact representation. However, the potential for cognitive interpretations of this type of independency was never considered, although CSI could have a major impact on how problems are solved in general. This is the problem we address here.

If CSI can be discovered in data and provide correct inference, then intuition suggests there must be something particular about the data in which these independencies hold. Our investigation deals with the type of useful relationships that can be uncovered and used to increase our knowledge of a particular situation under study. The hypothesis motivating this idea is that if a different set of independencies holds within the same distribution, but for different subsets of the distribution, these subsets must be important in some structural way, and we may have to treat them differently. Our research results demonstrate how such consideration of context can, in a variety of situations, provide a new way to separate an initial seemingly correct, but possibly erroneous, causal model into two or more new models that take into account independencies that were too specific for the initial model. This decomposition of the causal model allows for a more accurate representation of the situation being modeled.

For example, consider a group of students who all have trouble reading. Let the variables in the distribution under investigation be *Age*, *Reading Skills* of the student, *First Language* spoken at home, *Occupation of Parents*, and *Mathematics Skills*. Considered as a whole, we may not be able to find an independence that holds for all values, meaning no variable could be removed from the distribution, no CIs hold and any information hidden within the variables will remain unused and we will have to conclude that the group simply consists of people with "learning difficulties." There is no commonality between any of the variables remaining in the distribution. However, say we discover a CSI in the context *First Language (FL) = non-english (ne)*. We notice that given $FL = ne$, variables *Reading Skills* and *Mathematical Skills* become independent, while in context $FL = english\ (e)$, the variables are highly dependent. In terms of computational impact, we can reduce the portion of the distribution where $FL = ne$, but the impact is much more important than a simple reduction. Upon further study, we come to the conclusion that generally students with learning difficulties tend to do poorly in most academic subjects. If there is no dependence between *Mathematical Skills* and *Reading Skills* for *non-english* students, there may be something about those students that goes beyond learning difficulties. It may be that the language barrier is preventing them from acquiring reading skills as quickly as other students, whereas the effect is less predominant in mathematics since the symbols and the reasoning is standard across languages. From this simple example, we would have to reject the possibility that a categorization of "learning difficulty" is an adequate choice. Without considering context, we would have made an erroneous attribution.

As a result of our investigation, we show how a formal consideration of context can help correct erroneous assumptions used in formalizing some seemingly paradoxical scenarios. When relevant independencies hold within variables, erroneous inference is almost inevitable. In the extreme case, that type of error may lead to seemingly paradoxical scenarios. To emphasize this problem and show how CSI may help solve it, we discuss a well-known paradox, namely *Simpson's Paradox*, and describe how it can be formalized by means of CSI [36] 5.2.

1.2.2 Towards Elimination of Bias in Attribution Theory

Attribution theory is the field of social psychology that deals with lay, or common sense, explanations of behaviour. The theory assumes that people try to understand why others do the things they do by attributing causes to that particular behaviour. In general, the conclusions we make about an individual's behaviour determine our reactions to the particular individual. Even more generally, our interactions with other people have roots in the attributions or explanations we make of what they say and do. Attribution theory explores how people associate causes to events and how their subsequent actions, namely the event they choose as causal, will be affected by this cognitive perception.

In theorizing about attribution, we are interested not only in the true cause of behaviour but also in how human adults, assign a cause to another person's behaviour, whether the inferred cause is true or not. However, attribution is equally interested in understanding how other individuals would attribute the behaviour. Two fundamental principles [37] play a key role in determining the attributions people make, namely the *discounting principle*, and the *covariation principle*. The discounting principle can be thought of as the principle of the most easily observed cause. According to the discounting principle, as the number of possible causes for an effect increases, our confidence in our knowledge of the true cause should decrease. People tend to accept the more frequently observed scenario as being causal, and disregard all other possibilities. The covariation principle is applicable when two events are associated over a series of instances. If event B always happens when event A occurs, and does not take place when A is absent, people often infer that one causes the other.

From the principles of discounting and covariation, people generally use two categories of causes to understand other people's behaviour, namely *situational* and *dispositional* causes. A cause that explains actions in terms of a social setting or environment is defined as *situational*, while *dispositional* causes are based on characteristics of the person in question.

Based on the principles of discounting and covariation and the causes for explanation of behaviour, three theories have proved most influential [38]. The first is Jones and Davis's model of correspondent inferences [39], which concerns a single social interaction. The second is Kelley's covariation model [40], which consists of a relationship over time, and finally the third is Weiner's model of achievement attributions [41], which deals with situations involving success or failure.

One problem with most psychological accounts for the determination of attribution is that attribution theory is based on the assumption that humans are rational thinkers, and therefore always use available information in a rational way to make decisions. Based on that belief, the human subject is a "naive scientist", since we believe in the systematic search for relevant information followed by a *rational* logical explanation of behaviour. However, research shows that we often make attributions that are *not* based on rational conclusions, thus the "naive scientist" is fallible. These irrational attributions create *biases* in the attribution process, which can have a significant impact on our conclusions. Some well-documented biases are those of *Correspondence Bias*, *Self-Serving Bias*, *Defensive Attributions*, and the *Illusion of Control*.

The *correspondence bias* [42] relates to events where even when logic and evidence suggest otherwise, people have biases that lead them to conclude that the person who performed an act were predisposed to do so. The *self-serving bias* [43] in rational attribution is one that subconsciously helps us protect our ego and self-esteem. Human adults tend to attribute their successes to internal factors and to detach themselves from their failures by attributing them to external factors. The *defensive attributions bias* deals with our need to feel secure. It has been suggested [44] that we act defensively to disassociate ourselves from the possibility of a threatening event. Finally, the *illusion of control bias* addresses our exaggerated belief in our own capacity to determine what happens to us in life [45].

Theories of attribution are used to determine how people attribute causes to events in everyday life, considering that humans live in an environment with many variables. Therefore, they must account for the different situations that people are in to determine the cause of their behaviour. This consideration leads to the above biases. In the present research, we address how causal models can be refined based on the environment or the circumstances surrounding a situation being modeled through a consideration of contextual data. We investigate ways to use independencies in AI in contexts of the data to find particularities about subsets where independencies hold. We consider subsets where this is true in more representative models adapted to the context.

Finally, we show how contexual independencies can help discover hidden dispositions and situational factors in causal relations [46], and we present a decomposition of the causal model that considers situational and dispositional factors separately, letting variables be omitted in the case where they are irrelevant, and emphasized otherwise. Once again, we emphasize here that without consideration of CSI, the model we present would seem intuitively correct, and false conclusions would be drawn from it.

1.2.3 Discovering Focal Sets in Cheng and Novick's Probabilistic Contrast Model

When seeking how human adults induce the causes of events in everyday life, we must make an attempt to recover the causal structure of the world. This is the primary goal of causal induction, along with making predictions about future events. Based on this, Cheng and Novick, in their early work, insisted that covariation "has generally been regarded as a necessary (although insufficient) criterion of normative induction" [20]. Recall that covariation refers to the change in the probability of an effect given the presence versus the absence of a potential cause. Cheng and Novick also claim that their reassessment of attributional bias is the solution to the problem of incomplete information. We discuss these claims and their implications in this section.

Cheng and Novick studied the biases found in causal attribution. They strived to answer "why the biases would appear under certain sets of conditions but not under others." [19] This last statement can be rewritten as: biases seem to be present in some contexts but not in others. In this sense, much of Cheng and Novick's work is an attempt at taking into account the power of contextual information at affecting, or disrupting seemingly correct attributions.

Prior expectations formed in our minds tend to override some perhaps more objective data-based processing [47, 48, 49, 50]. Another explanation, by Hilton and Slugoski [51], suggested that subjects may not even use covariational information to make their judgments or attributions but rather something completely different, something they call an "abnormal condition". This suggests that context is being treated as an exception, or an unnatural occurence, that needs to be "handled", rather than be treated as a natural phenomenon that simply needs to be discovered. If we have ways to discover in what context certain attributions are made, we could then consider these contexts in isolation. Although Cheng and Novick do not discover contexts explicitly, they do make use of them. They call such subsets *focal sets*. We discuss focal sets and situate them in the realm of the present work in Chapter 4. For the moment, we discuss briefly how Cheng and Novick initially attempted to tackle the problem of biases.

Cheng and Novick's initial proposition to explain biases in attribution was an information-based proposal that causal induction was in fact "based on an assessment of covariation" [20]. They noticed that the information we presume available to the subject to make causal attributions is assumed to be the information the subjects actually use when they make their causal attribution. That statement implies that humans are entirely rational thinkers and that the information people use when making causal attributions is known, which would eliminate the possibility that past experiences and beliefs may come into play when we attribute a cause to an effect. Cheng and Novick do not believe that statement is true. They believe that people think beyond the given information when arriving at an analysis of a situation. This belief is shared by others in the cognitive psychology literature [52, 53, 54]. In addition, Hilton and Slugoski [51] have reported that human causal attributions were influenced by their implicit knowledge of norms. Since our knowledge of norms and our a priori knowledge is not shared by all subjects or known by the experimenter, Cheng and Novick came to the conclusion that the biases arise not from inferential process caveats but rather the data on which the inference rules operate. They claim that the problem is one of incomplete information. This statement is what gives rise to a portion of the present research.

Techniques in AI allow us to discover independencies in the data and thus make better, more accurate use of the available data [7, 8]. In addition, we have investigated integrating independencies discovered in the data with existing AI algorithms to build more specific, descriptive, and accurate causal models from larger, more general, sometimes misleading, causal models. This work draws a clear distinction between the inference process and the data used to infer causation. We believe that problems arising from poor consideration of context, and leading to the integration of bias, are in the data, not in the inferential process. Also, the important issue of bias can be addressed using this distinction.

Work done in causal attribution before Cheng and Novick's probabilistic contrast model typically has *not* distinguished between data and process, or has *not* accurately defined what information people actually use to make causal attributions, thus leading to incorrect attribution. Cheng and Novick explored the possibility that people decide on causal relationships based on more than the facts that are provided in a controlled experiment setting, such as personal beliefs and past experiences, etc., which are rational human biases. They believe the way people perceive situations and facts is a contextual matter, and those contexts are addressed under the umbrella of biases in attribution. According to their model, biases can be understood as contexts in which the exception becomes the norm. They believe that an understanding of those biases can help us determine what data is relevant to causal attribution, but can only be considered or detected if a distinction is made between data and process.

From this awareness of variability based on context, Cheng and Novick concluded that the problem with covariation wasn't one of incompatibility with the way humans process information (i.e. not thinking rationally, or in a normative manner), but rather one of insufficiency for determination of covariation due to assessment of bias. Whereas the previous models of covariation suggest that deviations may arise because humans do not think rationally, Cheng and Novick counter that model, and argue that humans do make rational inferences, and discrepancies come from an incorrect assessment of bias. They suggested that a model must be normative if we are to qualify deviations from the model as "biases". They suggested that observed deviations from existing models of covariation, may, after all, be rational inferences, and proposed their initial probabilistic contrast model: a covariational model based on estimated differences in the probabilities of the effects conditional on the presence versus absence of potential causal factors.

From studying bias and incomplete information, Cheng and Novick realized that covariation was in fact not sufficient to explain causal attribution, since covariation alone doesn't necessarily imply causation. The main factor that led Cheng and Novick to believe that covariation could not account completely for human causal induction was that covariation alone is unable to explain why even untutored reasoners do not equate covariation with causation.

In the psychology literature, there is an opposing approach to explaining causal inferences, namely the *causal power* approach. This approach has attempted to address reasoners' intuitive understanding of this fundamental inequality, but has been unsuccessful at specifying the process that transforms information from the available noncausal input to a causal judgment. To address this issue, Cheng [21] formulated a revised version of Cheng and Novick's [19] probabilistic contrast model. This improved model, the power PC theory, demonstrates that an integration of the covariation and power approaches can overcome the problems confronting each approach studied in isolation.

The two approaches have distinct roots in philosophy, which makes the potential for their combination very interesting, as philosophical theories need to be widely accepted to be deemed a philosophical account. Cheng and Novick came to the conclusion that neither covariation alone nor causal power alone can explain the inferences humans make about causal relations. The main question about causality in the philosophy literature deals with: "How does a reasoner come to know that one thing causes another?" Covariation traces its roots to the philosopher David Hume [55], while causal power stems from the philosophy of Kant [56].

Contextual consideration is a mandatory prerequisite to Cheng and Novick's probabilistic contrast model as well as the fully integrated Power PC model, which considers covariation and causal power in conjunction. Their consideration of context, termed *focal sets* classifies contextually selected sets of events over which contrasts are computed. Cheng and Novick rely on a human expert when building focal sets. We provide an automated method for discovering focal sets from the available information, rather than to rely solely on the human experts. This discovery method may also discover interesting and legitimate contexts that may have remained unnoticed with the expert alone (see Section 5.4).

1.3 Organization of Thesis

The remainder of the thesis is organized as follows. Chapter 2 deepens the above discussion regarding the first theme, namely tools in AI and solving problems with context. Then, Chapter 3 focusses on the second theme, addressing existing theories of attribution and the biases documented in the literature, to account for erroneous attributions. In Chapter 4, the third theme, a model of causal induction by human adults, is presented in more detail. In this chapter, we discuss Cheng and Novick's initial covariational probabilistic contrast model, followed by the resoning behind the required adjustments to their model, and finally, their improved contrast model, Power PC. In Chapter 5, we revisit the three themes and present our contributions pertaining to contextual considerations regarding decision making and problem solving, attribution theory, and Cheng and Novick's probabilistic contrast model. Finally, we offer some conclusions and suggest future work in Chapter 6.

Chapter 2

Uncertain Reasoning in Artificial Intelligence

The contributions suggested here rely heavily on a generalized form of *probabilistic conditional independence* (CI) used in uncertain reasoning for a more compact representation of probability distributions and efficient query processing. We will later argue that *context-specific independence* (CSI) can be used for much more than reducing the number of mathematical operations required in inference, and shows promise as a cognitive tool.

For the moment, we situate CSI in the vast topic of uncertain reasoning in AI and discuss its role within that framework. We first justify the use of probability theory as a representation for uncertain information. We then discuss how Bayesian networks (BNs) made possible a probabilistic representation of uncertain reasoning.

2.1 Choosing a Representation

In choosing a representation for uncertain reasoning in AI, several approaches to probability have been considered. For instance, Mises [57] presented a frequentist approach, while Carnap [58] presented a logical approach. In this section, we address the subjective approach to probability theory.

Probabilistic expert systems have been used to deal with uncertainty for several reasons. In using a probabilistic representation, every uncertainty statement is subjectively expressed in the form of a probability. Also, combinations of uncertainties are grouped using the rules of probability. Furthermore, calculation of probabilities is appropriate to handle any situation involving uncertainty, once each configuration is assigned a probability value, which translates in probabilistic terms to the *joint probability distribution* (JPD), which we define formally in the next section. Once we have the complete JPD, we can answer any query about the variables inside the JPD and we can update it when we receive new information. This last statement leads to the most important criticisms in choosing probability theory as a method for managing uncertain information. Criticisms of probability theory were based on [59]:

(i) the exponential number of parameters required, and

(ii) the impossibility of accurate estimation of individual probabilities.

Once we *have* a probability value for every configuration in a joint probability distribution, we can answer any query about any number of combinations of the data in that particular domain. Take for example rolling a die. With a regular (6 face), fair (equally likely to land on any face) die, we know that the probability of landing on any one of the 6 faces is 1/6. With these values in place, we can answer any query (e.g. what is the probability of landing on an even number...1/6+1/6+1/6 = 3/6 = 0.5), and we can *update* the distribution (e.g. we find out the dice is not fair and will never land on the face of the number 5).

However, *obtaining* this joint distribution directly was not a computationally feasible task as the number of parameters increased exponentially as a function of the number of variables in the distribution. Consider the following example below.

A medical diagnosis application involves 50 binary variables. In order to have a complete JPD, $2^{50} - 1$ probability values must be specified. Not only does it seem outrageous to specify $2^{50} - 1$ values but it may become impossible for the domain expert to specify even one value for one configuration when 50 variables must be considered. This leads us to confirm the validity of the two above-mentioned criticisms (i) and (ii), which conclude that it is impossible to specify the values in the distribution directly in a real life application.

However, probabilistic network models, such as *Bayesian networks*, which we discuss in Section 2.3, have managed to overcome the criticisms of probability theory for uncertain reasoning in cases where the network is sparsely connected, since it can, with the notion of *probabilistic conditional independency* (CI), obtain the JPD values *indirectly* .

2.2 Probabilistic Terminology

Here, we give a standard account of probability calculus [60], as we will later use short hand notation, making some assumptions of properties understood by context.

Let Ω be a finite probability space, that is, a finite set of points and let pr be a strictly positive real function on Ω such that

$$\sum_{\omega \in \Omega} pr(\omega) \quad = \quad 1.$$

An *event* a is a nonempty subset of Ω; $\neg a$ denotes the complement of a so a and $\neg a$ form a partition of Ω. The (prior) *probability of* a is

$$p(a) \quad = \quad \sum_{x \in a} pr(x).$$

A *random variable* (or just *variable*) is any function R from Ω to $\Re$, the real numbers. Then the *expected value* or *expectation* of a random variable R is

$$E(R) \quad = \quad \sum_{\omega \in \Omega} R(\omega) pr(\omega).$$

Corresponding to every event a is a discrete random variable A which is the *characteristic function* of a. That is, $A(\omega) = 1$ if $\omega \in a$ and $A(\omega) = 0$ otherwise, for all $\omega \in \Omega$. Note that

$$p(a) = E(A).$$

The set of points ω such that $A(\omega) = 1$ corresponds to the event a, and the set of points such that $A(\omega) = 0$ corresponds to the event $\neg a$. The partitioning events a and $\neg a$ are *outcomes* of the variable A, and more generally a variable is a partition of Ω. We could generalize the idea of outcomes beyond binary variables, but don't for the present purpose.

The probability of joint event a and b, or *joint probability* of a and b is $p(a \cap b)$ and the *joint probability distribution* of any two variables is the set of joint probabilities of all outcomes of the variables. The *conditional probability* of a given b is

$$p(a|b) = \frac{p(a \cap b)}{p(b)}$$

where $p(b)$ must be nonzero. The *conditional probability distribution* (CPD) of any two variables is the set of conditional probabilities of all outcomes of the variables. The definition of conditional probability implies that the joint probability of a and b can be rewritten as

$$p(a \cap b) = p(b) \cdot p(a|b).$$

Let $p(A_1, A_2, \ldots, A_n)$ denote the joint distribution of the variables in $D = \{A_1, A_2, \ldots, A_n\}$, a finite set of discrete random variables. We define the *marginal probability* $p(a_i)$ of the i^{th} variable as [60]

$$p(a_i) = \sum_{A_1,\ldots,A_{i-1},A_{i+1},\ldots,A_n} p(A_1, \ldots, A_n),$$

where $A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n$ is simply the sum over all possible values of the variables. The *marginal probability distribution* (MPD) of any two variables is the set of marginal probabilities of all outcomes of the variables. We call the process of computing the marginal probability distribution *marginalization.*

Finally, for completeness, the *conditional expectation* of a random variable R given m is defined analogously to the definition of conditional probability:

$$E(R|m) = \frac{\sum_{\omega \in \Omega} R(\omega) pr(\omega)}{p(m)}.$$

Again observe that

$$E(R|m) = p(r|m)$$

when R is the characteristic function of the set r. For the remainder of this document $p(a \cap b) = p(a, b)$ when $\cap$ is understood by context. From this point on, for simplicity, we can assume all variables to be binary unless otherwise specified.

2.3 Bayesian Networks

Pearl [33] formalized the notion of *Bayesian networks* (BNs). The general idea behind BNs is that although it may not be efficient to *directly* specify a joint probability distribution, BNs allow for a JPD to be specified *indirectly*. This is achieved due to the *conditional independence* (CI) assumptions encoded in the Bayesian network. BNs will be discussed in greater detail and defined formally after we discuss CI.

2.3.1 Probabilistic Conditional Independence (CI)

Let A denote the domain in which outcome a appears. We say that variables Y and Z are *conditionally independent* given X, denoted $I(Y, X, Z)$, if, given any $x \in X$, $y \in Y$, then for all $z \in Z$,

$$p(y|x,z) \quad = \quad p(y|x), \qquad \text{whenever } p(x,z) > 0. \tag{2.1}$$

or equivalently,

$$p(Y,X,Z) \quad = \quad \frac{p(Y,X) \cdot p(X,Z)}{p(X)}. \tag{2.2}$$

From the first definition, we can see that the set Z of variables does not change the probability of Y once we know the value of X. Therefore the value of Z provides no information in the CPD $p(y|x,z)$. The second definition is also interesting because it explicitly shows how the distribution can be decomposed into smaller distributions based on the conditional independence. We are *reducing* the size of the MPDs.

To illustrate the idea more clearly, consider the CPDs in Table 2.1. The 4 CPDs contain conditional independencies, which will make it possible to decompose them into smaller distributions. By Equation (2.1), the CIs found in the CPDs in Table 2.1 are as follows:

$$\begin{aligned} p(B|A) &= p(B), & (2.3)\\ p(C|A,B) &= p(C|A), & (2.4)\\ p(D|A,B,C) &= p(D|A,B), & (2.5)\\ p(E|A,B,C,D) &= p(E|A,C,D). & (2.6) \end{aligned}$$

The CIs in Equations (2.3) - (2.6) are denoted as follows: $I(\{B\}, \emptyset, \{A\})$, $I(\{C\}, \{A\}, \{B\})$, $I(\{D\}, \{AB\}, \{C\})$, $I(\{E\}, \{ACD\}, \{B\})$, respectively. The resulting CPDs are presented in Table 2.2.

To discover those independencies simply by inspecting the CPDs, we ask: "Does knowing the value of one of the variables change the likelihood of the configuration?" If not, there is no need to store it.

Table 2.1: The conditional probability distributions (CPDs) $p(B|A)$, $p(C|A,B)$, $p(D|A,B,C)$, and $p(E|A,B,C,D)$ containing CIs.

AB	$p(B\|A)$
00	0.3
01	0.7
10	0.3
11	0.7

ABC	$p(C\|A,B)$
000	0.2
001	0.8
010	0.2
011	0.8
100	0.3
101	0.7
110	0.3
111	0.7

$ABCD$	$p(D\|A,B,C)$
0000	0.3
0001	0.7
0010	0.3
0011	0.7
0100	0.3
0101	0.7
0110	0.3
0111	0.7
1000	0.6
1001	0.4
1010	0.6
1011	0.4
1100	0.8
1101	0.2
1110	0.8
1111	0.2

$ABCDE$	$p(E\|A,B,C,D)$
00000	0.1
00001	0.9
00010	0.1
00011	0.9
00100	0.8
00101	0.2
00110	0.8
00111	0.2
01000	0.1
01001	0.9
01010	0.1
01011	0.9
01100	0.8
01101	0.2
01110	0.8
01111	0.2
10000	0.6
10001	0.4
10010	0.3
10011	0.7
10100	0.6
10101	0.4
10110	0.3
10111	0.7
11000	0.6
11001	0.4
11010	0.3
11011	0.7
11100	0.6
11101	0.4
11110	0.3
11111	0.7

Table 2.2: By utilizing the CIs in Equations (2.3) - (2.6), the initial CPDs in Table 2.1 can be simplified as shown.

B	$p(B)$
0	0.3
1	0.7

AC	$p(C\|A)$
00	0.2
01	0.8
10	0.3
11	0.7

ABD	$p(D\|A,B)$
000	0.3
001	0.7
010	0.3
011	0.7
100	0.6
101	0.4
110	0.8
111	0.2

$ACDE$	$p(E\|A,C,D)$
0000	0.1
0001	0.9
0010	0.1
0011	0.9
0100	0.8
0101	0.2
0110	0.8
0111	0.2
1000	0.6
1001	0.4
1010	0.3
1011	0.7
1100	0.6
1101	0.4
1110	0.3
1111	0.7

In the CPD $p(B|A)$, knowing the value of the variable A does not change the belief in B, i.e. $p(B|A) = p(B)$. We say that given the empty set, variables A and B are independent. For CPD $p(C|A,B)$, knowing B does not change the belief in C when the value of A is known so variables A and C are independent given B, i.e. $p(C|A,B) = p(C|A)$. Following the same argument for the two remaining CPDs, variables C and D are independent given variables A and B in $p(D|A,B,C)$, i.e. $p(D|A,B,C) = p(D|A,B)$, and finally variables B and E are independent given variables A, C, and D in $p(E|A,B,C,D)$, i.e. $p(E|A,B,C,D) = p(E|A,C,D)$.

The number of values to be specified is reduced when we use CIs. Instead of specifying 60 values like in Table 2.1, only 30 values need to be specified in Table 2.2 when CI is considered. Now, to represent a distribution in terms of small CPDs containing conditional independencies, we must review the chain rule of probability. By the chain rule of probability, we may write the following identity

$$p(A,B,C,D,E) = p(A) \cdot p(B|A) \cdot p(C|A,B) \cdot p(D|A,B,C) \cdot p(E|A,B,C,D). \tag{2.7}$$

By the definition of conditional probability (see Section 2.2), we can rewrite Equation (2.7) as

$$\begin{aligned} & p(A,B,C,D,E) \\ & = p(A) \cdot \frac{p(A,B)}{p(A)} \cdot \frac{p(A,B,C)}{p(A,B)} \cdot \frac{p(A,B,C,D)}{p(A,B,C)} \cdot \frac{p(A,B,C,D,E)}{p(A,B,C,D)}. \end{aligned} \tag{2.8}$$

Canceling out the common terms in Equation (2.8), shows that the right side of the equation, $p(A,B,C,D,E)$, is identical to the left side. Note that the CPDs in Equation (2.7) are found in

Table 2.1, with the exception of $p(A)$. The reason being that $p(A)$ does not contain any nontrivial CIs and therefore will not be decomposed into a smaller distribution.

By substituting the CIs in Equations (2.3) - (2.6) into Equation (2.7), the following simplified factorization, termed *factorized JPD*, is obtained

$$p(A, B, C, D, E) \quad = \quad p(A) \cdot p(B) \cdot p(C|A) \cdot p(D|A, B) \cdot p(E|A, C, D). \tag{2.9}$$

Note that the CPDs in Equation (2.9) are found in Table 2.2, once again with the exception of $p(A)$. For more details on reading CIs, refer to Appendix A.

2.3.2 Directed Acyclic Graphs (DAGs)

Because it is inconvenient to explictly state all of the CIs that hold in a distribution, we use a graphical structure called a *directed acyclic graph (DAG)*. The DAG together with the corresponding CPDs define a *Bayesian network* (BN). To each variable A_i with parents $Y_1, \ldots, Y_j$ in the DAG, there is an attached conditional probability table $p(A_i|Y_1 \ldots Y_j)$. Also, in the DAG, every child is independent of its non-descendents, given the state of its parents. The DAG in Figure 2.1 corresponds to the CPDs formed from the inspected CIs in Table 2.2.

Bayesian networks have also been used as a representation of causality. Pearl and Verma [61] provide a causal semantics for BNs. They claim that a *causal model* of a set of random variables R can be represented by a directed acyclic graph (DAG), where each node corresponds to an element in R and edges denote direct causal relationships between pairs of elements of R.

The direct causal relations in the causal model can be expressed in terms of *probabilistic conditional independencies* (CIs) [33]. For the remainder of this document, the terms Bayesian network and causal model will be used interchangeably. Before addressing the issue of probabilistic inference, we describe a method for validating *non-local* CIs from a Bayesian network.

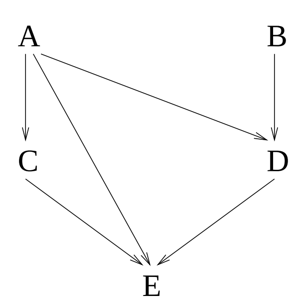

Figure 2.1: A Bayesian network for $p(A, B, C, D, E)$.

2.3.3 Methods for the Verification of Non-Local CIs

As mentioned previously, each CPD in a Bayesian network represents the probability values of the configurations of one node (variable) in the DAG given the state of its parents. Each CPD quantifies the relationship between a node and its parents in the DAG without considering the other nodes in the DAG. However, there exists a method to verify the validity of independencies between any two sets of nodes in the DAG. We call these independencies *non-local independencies* and the method *d-separation.* The d-separation method is used to test CI statements in a DAG. First, we define *local* and *non-local* independencies, and then, we define *d-separation.*

Definition 1 [3] A *local independency* is one that involves variables from a single CPD (i.e. a given node and its parents). A *non-local* independency is one that involves any other sets of nodes in the Bayesian network.

A DAG represents local CIs explicitly and non-local CIs implicitly.

Definition 2 [33] Let X, Y, Z be disjoint subsets of variables in a DAG $\mathcal{D}$. Set X *d-separates* sets Y and Z, $I(Y, X, Z)$, if along every path (direction of arrows in DAG not important) between a node in Y and a node in Z, there exists a node N in the path satisfying one of the following two conditions:
(i) N has converging arrows, and none of its descendents (including N) is in X.
(ii) N does not have converging arrows and N is in X.

The d-separation method is sound and complete [33].

Example 1 *Consider the DAG in Figure 2.2. Using Definition 2, we verify the existence of the CI I(*$\{B\}$*,*$\{A\}$*,*$\{C\}$*).*
There are three paths from $\{B\}$ *to* $\{C\}$*:*
1. $\langle B, A, C\rangle \rightarrow A$ *is not converging and* A *is in* X*, so* A *satisfies (ii).*
2. $\langle B, D, C\rangle \rightarrow D$ *is converging and* D *is not in* X*, so* D *satisfies (i).*
3. $\langle B, E, C\rangle \rightarrow E$ *is converging and* E*,* F *are not in* X*, so* E *satisfies (i).*
Therefore, $\{A\}$ *d-separates* $\{B\}$ *and* $\{C\}$*.*

Example 2 *Consider again the DAG in Figure 2.2. Using Definition 2, we show the non-validity of the CI I(*$\{B\}$*,*$\{AF\}$*,*$\{C\}$*).*
We know from Example 1 that there are three paths from $\{B\}$ *to* $\{C\}$*:* $\langle B, A, C\rangle$*,* $\langle B, D, C\rangle$*, and* $\langle B, E, C\rangle$
$\rightarrow$ *For path* $\langle B, E, C\rangle$*,* E *has converging arrows, and one of its descendents* F *is in* X*, so* E *does not satisfy condition (i).*

→ Also for path $\langle B, E, C \rangle$, neither B, E, or C are in X, therefore condition (ii) is not satisfied. Since conditions (i) and (ii) fail for at least one path from B to C, namely $\langle B, E, C \rangle$, $\{A, F\}$ does not d-separate $\{B\}$ and $\{C\}$.

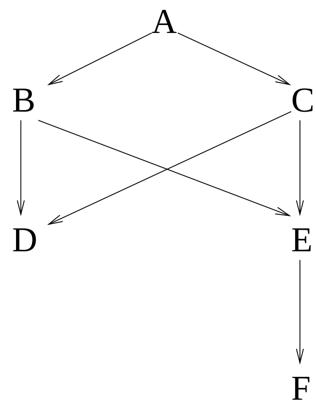

Figure 2.2: A DAG for $p(A, B, C, D, E, F)$.

2.3.4 Inference on Factorized Distributions

In terms of inference, with a factorized distribution it may be possible to answer queries with local computations. It may not be necessary to compute the whole JPD in order to answer a query that involves only a fraction of the variables. We use the multiplication and marginalization operations to perform this task.

Given a factorized JPD, to *locally* marginalize out a variable a_i from the distribution, we must first remove from the factorization all functions that involve a_i and compute the product of the resulting functions. Then we marginalize out a_i from the product, and put the resulting function back into the factorization.

As an example, let's compute $p(A, B, C, E)$ from the factorization in Equation (2.9). Since the distribution is over A, B, C, D, and E, we must marginalize out variable D from the distribution. This simple idea is the basis for all inference in BNs. Different algorithms try to use CIs in specialized ways. We compute $p(A, B, C, E)$ as follows:

$$\begin{aligned} p(A, B, C, E) &= \sum_D p(A, B, C, D, E) \\ &= \sum_D p(A) \cdot p(B|A) \cdot p(C|A) \cdot p(D|A, B) \cdot p(E|A, C, D) \\ &= p(A) \cdot p(B|A) \cdot p(C|A) \cdot \sum_D p(D|A, B) \cdot p(E|A, C, D). \end{aligned} \tag{2.10}$$

Although the above manipulation may seem trivial, the computational savings are quite large. Instead of multiplying all the variables together and doing the marginalization over the resulting large set, the marginalization is performed only over the CPDs involving variable D. The other CPDs remain untouched.

Using the CPDs $p(D|A, B)$ and $p(E|A, C, D)$ in Table 2.2, computing the product $p(D|A, B) \cdot p(E|A, C, D)$ requires 32 multiplications. Marginalizing out variable D from this product requires 16 additions. The resulting distribution can be multiplied with $p(A) \cdot p(B) \cdot p(C|A)$ to obtain our desired distribution $p(A, B, C, E)$.

It is important to note that although we can do inference *locally*, i.e., without computing the whole JPD, there may be some cases where the distribution is still too large to make the computation feasible.

2.4 A More Compact Representation and Improved Query Processing

Bayesian networks may make indirect acquisition of the joint probability distribution feasible due to conditional independence assumptions. This allows for a more compact representation of the distribution and makes the inference process feasible in many applications. However the notion of conditional independence is too restrictive to capture independencies that only hold in certain contexts. In this section, we review this type of contextual independence that has been formalized as *context-specific independence* (CSI) [3]. We show how inference with CSI is possible and how it can speed up the inference process. Next, we discuss methods for capturing CSI. We first discuss a method by Boutilier et al. [3], which facilitates the acquisition of CSI from a human expert. Finally, we discuss an algorithm that allows us to detect CSI from the conditional probability distributions in the case where no expert is available [7].

2.4.1 The Role of CSI in Uncertain Reasoning with Bayesian Networks

Although Bayesian networks have rendered probabilistic inference computationally feasible in applications where each conditional probability distribution involves only a fraction of the variables, conditional independence alone remains restrictive. It is only possible to take advantage of the conditional independence and benefit from a decomposition of the CPDs if a certain CI holds for *all* values of a variable in the distribution (see Equation 2.1). With context-specific independence, we can recognize CIs that hold for a subset of values of a variable in a distribution. Thus, CSI is a CI that need only hold in a specific context and not *all* contexts like its CI counterpart. It allows us to further decompose the distributions, which means fewer values need to be specified.

For example, consider a CPD consisting of four variables, *Income*, *Profession*, *Weather*, and *Computer Skills*. When the value for *Profession* is *office-clerk*, the variables *Income* and *Weather* are independent. That is, the probability value of *Income* is the same no matter what value *Weather* takes on. Therefore, in theory it is useless to keep the *Weather* variable in the distribution, as it does not give us any clue about the probability of *Income*. However, without the notion of CSI, we cannot eliminate *Weather* from this distribution since the same cannot be said for the value *Profession = farmer*. If the *Weather* variable is manipulated, the probability values for the *Income* variable will fluctuate greatly. Ideally, we would like to keep the *Weather* variable in the subset of the distribution where *Profession = farmer*, and eliminate it in the subset where *Profession = office-clerk*.

Furthermore, we notice that a similar scenario exists for the variable *Computer-Skills*, except the independency holds for *Profession = farmer* but not for *Profession = office-clerk* in this case. The level of computer competence of the *office-clerk* will play a role in the determination of their *Income*, while the farmer's *Income* will not be affected by his/her *Computer-Skills*. In this second scenario, ideally we would remove the variable *Computer-Skills* in the context *Profession = farmer*, and keep it when the value *Profession = office-clerk*. This kind of independence is called *context-specific independence* (CSI).

Let p be a JPD over a set R of variables, let $\{X, Y, Z, C\}$ be pairwise disjoint subsets of R, and let $x \in X$, $y \in Y$, $z \in Z$, and $c \in C$, where A represents the domains in which outcome a appears. We say that Y and Z are *conditionally independent* given X in *context* $C = c$ [3], denoted $I_{C=c}(Y, X, Z)$ if,

$$p(y|x, z, c) \quad = \quad p(y|x, c), \qquad \text{whenever } p(x, z, c) > 0. \tag{2.11}$$

This definition is similar to the definition of CI. The difference is it explictly states the context c in which the independence holds.

Based on the above definition, we show how CSI may let us further decompose the CPDs. Consider again CPDs $p(D|A, B)$ and $p(E|A, C, D)$ from the factorization in Equation (2.9). Using the idea of context-specific independence, we can further decompose those two CPDs. Consider the CPD $p(D|A, B)$ redrawn in Table 2.3 (i). In that particular CPD, no conditional independencies hold, therefore we cannot decompose the distribution based on CI. However, we see that variables D and B are conditionally independent in the context $A = 0$. When the value of $A = 0$, the probability values of variable B do *not* change the probability of D. That is

$$p(D = d|A = 0, B = b) \qquad = \qquad p(D = d|A = 0).$$

Table 2.3 (ii), shows that when $A = 0$, variable B need not be stored, because the probability values will be the same with or without B. Therefore, if B is removed from that *portion* of the distribution, the number of probability values to be specified in the context $A = 0$ is reduced. Instead of storing the CPD $p(D|A, B)$, containing four probability values in context $A = 0$, in Table 2.3 (i), we store $p(D|A = 0)$, containing only two values, in Table 2.3 (iii) and $p(D|A = 1, B)$

Table 2.3: Variables D and B are conditionally independent in context $A = 0$.

ABD	$p(D\|A,B)$
000	0.3
001	0.7
010	0.3
011	0.7
100	0.6
101	0.4
110	0.8
111	0.2

(i)

ABD	$p(D\|A{=}0,B)$
000	0.3
001	0.7
010	0.3
011	0.7

ABD	$p(D\|A{=}1,B)$
100	0.6
101	0.4
110	0.8
111	0.2

(ii)

A D	$p(D\|A{=}0)$
0 0	0.3
0 1	0.7

(iii)

in Table 2.3 (ii). The total number of values to be specified drops from 8 to 6 by decomposing the CPD $p(D|A, B)$ with CSI.

Now, consider the CPD $p(E|A, C, D)$ in Table 2.4. Once again, although no conditional independencies hold over all values of a variable in the distribution, E and D are conditionally independent given C in context $A = 0$ while E and C are conditionally independent given D in context $A = 1$. That is,

$$p(E = e|A = 0, C = c, D = d) \quad = \quad p(E = e|A = 0, C = c)$$

and

$$p(E = e|A = 1, C = c, D = d) \quad = \quad p(E = e|A = 1, D = d).$$

When $A = 0$, we do not need to store variable D as it does not modify the belief in the occurence of E. Following a similar argument, we do not need to store variable C when the value of A is 1.

In this case, the number of probability values to be specified drops from 16 to 8 by decomposing the CPD with CSI. Instead of storing the CPD $p(E|A, C, D)$ in Table 2.4 (i), we store $p(E|A = 0, C)$ and $p(E|A = 1, D)$ shown in Table 2.4 (iii). The important point to remember is that although no CIs hold over all the values in the distribution, we may still be able to decompose parts of the CPDs with CSI.

2.4.2 Inference with CSI

In this section, we discuss how to perform inference with context-specific independence. To make inference with CSI achievable, we use the notions of *partial functions* and the *union-product operator* to obtain a *CSI-factorization*. From that factorization, we may sometimes be able to answer queries more efficiently than with the factorization obtained from CI alone.

Table 2.4: Variables E and D are conditionally independent given C in context $A = 0$, while E and C are conditionally independent given D in context $A = 1$.

(i)

$ACDE$	$p(E\|A,C,D)$
0000	0.1
0001	0.9
0010	0.1
0011	0.9
0100	0.8
0101	0.2
0110	0.8
0111	0.2
1000	0.6
1001	0.4
1010	0.3
1011	0.7
1100	0.6
1101	0.4
1110	0.3
1111	0.7

(ii)

$ACDE$	$p(E\|A=0,C,D)$
0000	0.1
0001	0.9
0010	0.1
0011	0.9
0100	0.8
0101	0.2
0110	0.8
0111	0.2

$ACDE$	$p(E\|A=1,C,D)$
1000	0.6
1001	0.4
1010	0.3
1011	0.7
1100	0.6
1101	0.4
1110	0.3
1111	0.7

(iii)

ACE	$p(E\|A=0,C)$
000	0.1
001	0.9
010	0.8
011	0.2

ADE	$p(E\|A=1,D)$
100	0.6
101	0.4
110	0.3
111	0.7

The Union-Product Operator

For our discussion, *partial functions* are functions *defined* for some, but not all, probability values in a distribution. The decompositions of Section 2.4.1, use partial functions. For example, $p(D|A = 0)$ is a *partial function* over A, B, and D, since it is *defined* when the value of A is 0 but it is *undefined* when $A = 1$. When the value of A is 1, we must consider variable B, $p(D|A = 1, B)$.

Definition 3 [5] A *partial function* of a set X of variables is a mapping from a proper subset of possible values of X. Thus, it is defined only for some, but not all possible values of X. The set of possible values of X for which a partial function is defined is called the *domain* of the partial function.

Definition 4 [5] A *full function* of a set X of variables is a mapping from the set of all possible values of X.

To manipulate partial functions, the standard multiplication operator "$\cdot$" needs to be generalized. Zhang and Poole [5] formalized this extension as the *union-product operator* $\odot$. The *union-product* $r(Y, X) \odot s(X, Z)$ of functions $r(Y, X)$ and $s(X, Z)$ is the function of variables in

Table 2.5: Three *partial* functions $p(D|A=0)$, $p(D|A=1,B)$, and $p(E|A=1,D)$.

AD	$p(D\|A{=}0)$
00	0.3
01	0.7

ABD	$p(D\|A{=}1,B)$
100	0.6
101	0.4
110	0.8
111	0.2

ADE	$p(E\|A{=}1,D)$
100	0.6
101	0.4
110	0.3
111	0.7

Table 2.6: The union-product $p(D|A=0) \odot p(D|A=1,B)$ of $p(D|A=0)$ and $p(D|A=1,B)$ in Table 2.5.

A	B	D	$p(D\|A=0)$	$p(D\|A=1,B)$	$p'(A,B,D)$
0	0	0	0.3	-	0.3
0	0	1	0.7	-	0.7
0	1	0	0.3	-	0.3
0	1	1	0.7	-	0.7
1	0	0	-	0.6	0.6
1	0	1	-	0.4	0.4
1	1	0	-	0.8	0.8
1	1	1	-	0.2	0.2

disjoint subsets $Y \cup X \cup Z$ defined as

$$r(y,x) \odot s(x,z) = \begin{cases} r(y,x) \cdot s(x,z) & \text{if both } r(y,x) \text{ and } s(x,z) \text{ are defined} \\ r(y,x) & \text{if } r(y,x) \text{ is defined and } s(x,z) \text{ is undefined} \\ s(x,z) & \text{if } r(y,x) \text{ is undefined and } s(x,z) \text{ is defined} \\ \text{undefined} & \text{if both } r(y,x) \text{ and } s(x,z) \text{ are undefined.} \end{cases}$$

The union-product operator $\odot$ is associative and commutative [5].

Example 3 Consider $p(D|A=0)$, $p(D|A=1,B)$, and $p(E|A=1,D)$ redrawn in Table 2.5. We compute $p(D|A=0) \odot p(D|A=1,B) \odot p(E|A=1,D)$. The symbol "-" indicates when the function is *undefined* for a particular configuration. The computation of $p(D|A=0) \odot p(D|A=1,B)$ is illustrated in Table 2.6. By removing the fourth and fifth columns, we obtain the resulting distribution $p'(A,B,D) = p(D|A=0) \odot p(D|A=1,B)$. The computation of the final union-product $p'(A,B,D) \odot p(E|A=1,D)$, is illustrated in Table 2.7. After removing the fifth and sixth columns, we obtain the resulting distribution $p'(A,B,D,E) = p(D|A=0) \odot p(D|A=1,B) \odot p(E|A=1,D)$.

The union-product operator allows for a single CPD to be horizontally partitioned into several CPDs, based on the contextual independencies. Returning to the factorization in Equation (2.9), the CPD $p(D|A,B)$ can be rewritten as

$$\begin{aligned} p(D|A,B) &= p(D|A=0,B) \odot p(D|A=1,B) \\ &= p(D|A=0) \odot p(D|A=1,B), \end{aligned} \tag{2.12}$$

Table 2.7: The union-product $p'(A, B, D) \odot p(E|A = 1, D)$, where $p'(A, B, D)$ is shown in Table 2.6 and $p(E|A = 1, D)$ is shown in Table 2.5.

A	B	D	E	$p'(A, B, D)$	$p(E\|A = 1, D)$	$p'(A, B, D, E)$
0	0	0	0	0.3	-	0.3
0	0	0	1	0.3	-	0.3
0	0	1	0	0.7	-	0.7
0	0	1	1	0.7	-	0.7
0	1	0	0	0.3	-	0.3
0	1	0	1	0.3	-	0.3
0	1	1	0	0.7	-	0.7
0	1	1	1	0.7	-	0.7
1	0	0	0	0.6	0.6	0.36
1	0	0	1	0.6	0.4	0.24
1	0	1	0	0.4	0.3	0.12
1	0	1	1	0.4	0.7	0.28
1	1	0	0	0.8	0.6	0.48
1	1	0	1	0.8	0.4	0.32
1	1	1	0	0.2	0.3	0.06
1	1	1	1	0.2	0.7	0.14

while $p(E|A, C, D)$ is equivalently stated as

$$\begin{aligned} p(E|A, C, D) &= p(E|A = 0, C, D) \odot p(E|A = 1, C, D) \\ &= p(E|A = 0, C) \odot p(E|A = 1, D). \end{aligned} \tag{2.13}$$

The CSI Inference Process

The union-product operator lets the functions obtained from the CSI decompositions of the CPDs represent a factorization of the entire JPD. We illustrate this idea by showing a CSI refinement of the factorization in Equation (2.9) obtained from the Bayesian network in Figure 2.1.

With the CSI decompositions of $p(D|A, B)$ and $p(E|A, C, D)$, we can further refine the factorization of $p(A, B, C, D, E)$ into a more compact CSI-factorization. By substituting Equations (2.12) and (2.13) into the CI-factorization of $p(A, B, C, D, E)$ in Equation (2.9), the CSI-factorization of the JPD $p(A, B, C, D, E)$ is

$$\begin{aligned} p(A, B, C, D, E) &= p(A) \cdot p(B) \cdot p(C|A) \odot p(D|A = 0) \odot p(D|A = 1, B) \\ &\odot\ p(E|A = 0, C) \odot p(E|A = 1, D). \end{aligned} \tag{2.14}$$

In a CSI approach, the steps to marginalizing out a variable are similar to the ones presented previously for CI. The only difference is that in step (2), we compute the union-product instead of the product of the functions containing the variable to be marginalized out. The reason for the need of this modification is the introduction of partial functions. The modification is in fact a generalization of the steps presented previously.

Due to the CSI-factorization of $p(A, B, C, D, E)$, we obtain more CPDs and fewer variables in many of the CPDs. Therefore, computing $p(A, B, C, E)$ from Equation (2.14) involves

$$\begin{aligned} p(A, B, C, E) &= \sum_D p(A) \cdot p(B) \cdot p(C|A) \odot p(D|A = 0) \odot p(D|A = 1, B) \\ &\quad \odot\ p(E|A = 0, C) \odot p(E|A = 1, D) \\ &= p(A) \cdot p(B) \cdot p(C|A) \odot p(E|A = 0, C) \odot \sum_D p(D|A = 0) \\ &\quad \odot\ p(D|A = 1, B) \odot p(E|A = 1, D). \end{aligned} \tag{2.15}$$

Computing the union-product $p(D|A = 0) \odot p(D|A = 1, B) \odot p(E|A = 1, D)$ requires 8 multiplications. Next, 8 additions are required to marginalize out variable D. Eight more multiplications are required to compute the union-product of the resulting distribution with $p(E|A = 0, C)$. The resulting distribution can be multiplied with $p(A) \cdot p(B) \cdot p(C|A)$ to give $p(A, B, C, E)$.

2.4.3 Approximation of CSI

Although a sound formalism for inference with CSI is available, discovering CSIs can be challenging. One main difficulty is that very few CSIs can be discovered unless exact probability distributions are available. To solve that problem, CSI approximation methods have been proposed [62]. Input data with probability values close to zero were eliminated from distributions, making independencies more likely. Also, probability values that were significantly close to each other were given identical values, thus forcing CSI to be present and therefore making the representation more compact.

Finally, Poole and Zhang [34, 35] presented methods of allowing compact representations of the conditional probabilities of a variable given its parents. Such representations exploit contextual independence in terms of parent contexts. The authors hypothesize that the variables that act as parents may depend on the values of other variables.

For this thesis, we, like Pearl [63], assume the availability of exact probability distributions, where probabilities do not need to be approximated, and independencies are only discovered for identical probability values. In reality, such exact distributions are highly unlikely and approximation methods should be used as a preprocessing tool.

2.4.4 CSI Discovery

In this section, we discuss an acquisition method for context-specific independence, namely *CPD-trees*. Boutilier et al. [3] propose a method using *CPD-trees* to facilitate the acquisition of CSIs from a human expert. Then, we discuss an algorithm [7], which lets us detect the CSIs from a CPD using *CPD-trees* when no human expert is available.

Instead of viewing a CPD as a table, here we view a CPD as a tree structure, called a *CPD-tree* [3]. The CPD-tree representation is advantageous since it makes it particularly easy to elicit probabilities from a human expert due to its structured graphical representation. A second advantage of CPD-trees is that they allow a simple graphical method, which we call *CSI-detection*, for detecting CSIs [3]. We describe CSI-detection as follows.

Discovery by Human Expert

A CPD can be represented in a tree structure, called a *CPD-tree*, where variables in the CPD are represented by nodes in the CPD-tree, and the values of the variables in the CPD are represented by branches in the CPD-tree. Every path from root node to leaf node in the CPD-tree represents a unique configuration in the associated CPD with the probability value as the leaf node. Given a CPD-tree for a variable A and its parent set Π_A, i.e., a CPD-tree for the CPD $p(A|\Pi_A)$, the *label* of a path is defined as the values of the nodes on that path. A path is *consistent* with a context $C = c$ iff the labeling of the path is consistent with the assignment of the values in c. Given the CPD-tree depicting $p(Y|X, Z, C)$, we say that variable Y is independent of variable Z given X in the specific context $C = c$, if Z does not appear on any path consistent with $C = c$ [3].

Example 4 A human expert could specify the the CPD-tree in Figure 2.3 representing the CPD $p(E|A, C, D)$ in Table 2.2. Consider the context $A = 0$. Since variable D does not appear on any path consistent with $A = 0$, we say that variables E and D are independent given C in context $A = 0$. It can be verified that variables E and C are independent given D in context $A = 1$.

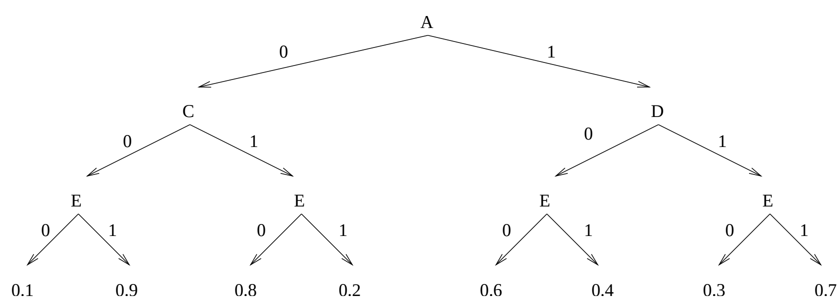

Figure 2.3: The CPD-tree given by a human expert representing $p(E|A, C, D)$ in Table 2.2.

Discovery from Probability Distributions

In this section, we discuss a method for detecting context-specific independencies from a CPD, since in many situations, no human expert is available and one must rely solely on data.

Example 5 Suppose there is no human expert available. The *initial* CPD-tree in Figure 2.4 is obtained directly from the CPD in Table 2.2. Although variables E and D are independent given C in context $A = 0$, while variables E and C are independent given D in context $A = 1$, the CSI-detection method does *not* detect any CSIs holding in this initial CPD-tree.

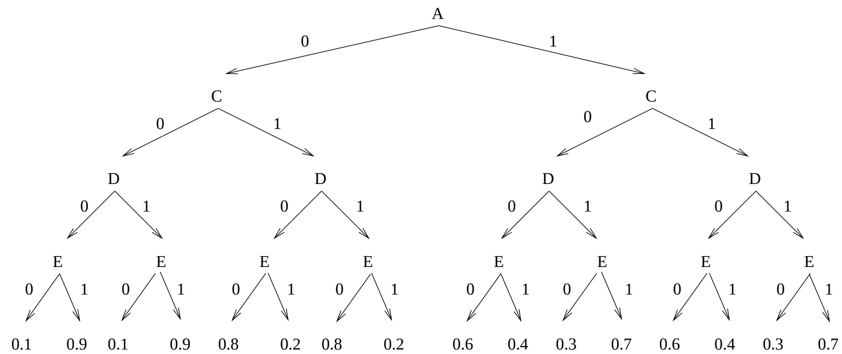

Figure 2.4: One *initial* CPD-tree for the given CPD $p(E|A, C, D)$ in Table 2.2.

Recall that for the purpose of this research, we assume exact distributions [63]. However, approximation methods are available when this is not the case [62]. Confidence intervals could also help approximate distributions.

The problem here is that the CSI-detection method is based on missing edges in the CPD-tree. On the contrary, the initial CPD-tree constructed directly from a given CPD will have all edges present. Thus, in order to take advantage of the CSI-detection method, we can use the following algorithm to remove the vacuous edges in the initial CPD-tree.

Algorithm 1 *REFINE CPD-TREE*

Input: an initial CPD-tree for a given CPD

Output: the refined CPD-tree obtained by removing all vacuous edges

begin

1. If all children of a node A are identical, then replace A by one of its offspring.

2. Delete all other children of node A.

end

Example 6 Consider again the initial CPD-tree in Figure 2.4. When $A = 0$ and $C = 0$, node D has identical children. Hence, node D can be replaced with node E. Similarly, for when $A = 0$ and $C = 1$. Moreover, when $A = 1$, node C has identical children. Node C can then be replaced by node D. The *refined* CPD-tree after these deletions is shown in Figure 2.3.

The following theorem, which appeared in [64] establishes the *soundness* of Algorithm 1.

Theorem 1 By removing the vacuous edges in the CPD-tree, Algorithm 1 correctly identifies CSIs.
Proof: Without loss of generality, consider the CPD $p(C|A, B)$ over three binary variables A, B, C, as shown in Table 2.8 (i). The *initial* CPD-tree for $p(C|A, B)$ using the node ordering $A - B - C$ is depicted in Figure 2.5. In order for Algorithm 1 to detect CSI, we must have a node with all identical children, as for node B in Figure 2.5. By the definition of Algorithm 1, node B will be replaced by one instance of node C, when $A = a_1$, as illustrated in the *resulting* CPD-tree in

Figure 2.6. Consider again the CPD in Table 2.8 (i). Variables C and B are in fact conditionally independent in the context $A = a_1$. Thus, the decomposition shown in Table 2.8 (ii, iii) is possible, where variable B is removed from the portion of the distribution where $A = a_1$. Thus, Algorithm 1 correctly identified the conditional independence of variables B and C in the context $A = a_1$.$\square$

Table 2.8: (i) The CPD $p(C|A, B)$ corresponding to the *initial* tree in Figure 2.5. (ii) The partition of $p(C|A, B)$ based on $A = a_1$ and $A = a_2$. (iii) Algorithm 1 correctly identifies the CSI $p(C|A = a_1, B) = p(C|A = a_1)$.

A B C	$p(C\|A, B)$
a_1 b_1 c_1	p_1
a_1 b_1 c_2	p_2
a_1 b_2 c_1	p_1
a_1 b_2 c_2	p_2
a_2 b_1 c_1	p_3
a_2 b_1 c_2	p_4
a_2 b_2 c_1	p_5
a_2 b_2 c_2	p_6

(i)

A B C	$p(C\|A, B)$
a_1 b_1 c_1	p_1
a_1 b_1 c_2	p_2
a_1 b_2 c_1	p_1
a_1 b_2 c_2	p_2
a_2 b_1 c_1	p_3
a_2 b_1 c_2	p_4
a_2 b_2 c_1	p_5
a_2 b_2 c_2	p_6

(ii)

$\rightarrow$

A C	$p(C\|A = a_1)$
a_1 c_1	p_1
a_1 c_2	p_2

(iii)

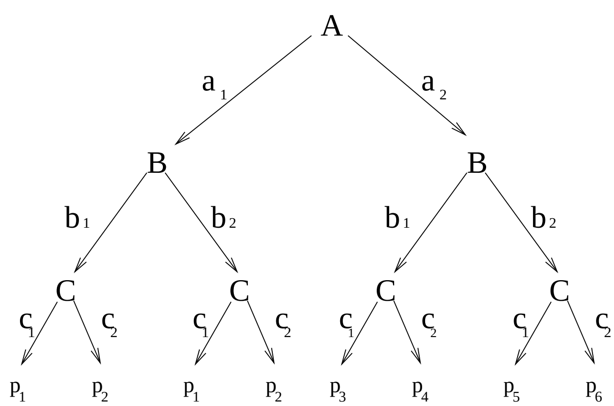

Figure 2.5: The *initial* CPD-tree for $p(C|A, B)$.

Although the algorithm is sound, Algorithm 1 may fail to detect some CSIs that are present in the distribution. The problem is due to the ordering of the nodes in the tree. Depending on the node ordering, the number and nature of CSIs detected may vary. As an example, it can be observed that if the ordering of the nodes in the initial tree in Figure 5.5 is changed from $A - C - D - E$ to $C - D - A - E$, Algorithm 1 does not detect any CSIs. Note that the variable being conditioned on must remain at the lowest level in the tree (leaf nodes).

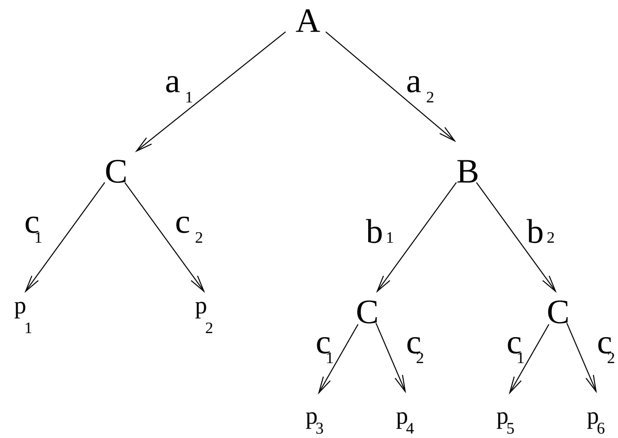

Figure 2.6: The *resulting* CPD-tree for $p(C|A, B)$.

Since node ordering is important in the algorithm, a heuristic can help determine a better, if not optimal node ordering. In the BN, the node with the most incoming arrows should be used as a level 1 node (assuming leaf nodes with probability values are labeled level 0) in the CPD-tree. The subsequent levels should contain remaining nodes in descending order of number of incoming arrows in the DAG. Variables with no incoming arrows should be at the top of the tree.

Finally, the main set-back with the algorithm is its exponential computational complexity. However, as we discuss later in the document, the complexity issue does not pose a major problem when dealing with an analysis of the semantics of discovered independencies, which is the focus of the thesis.

2.5 Direction of Generalizations

2.5.1 Contextual Weak Independence (CWI)

Throughout the development of CSI, we have not paid any attention to the natural "grouping" of the information, or analyzing the semantics of the available information. The main focus was on complexity reduction, fewer variables per distribution, fewer operations required for inference. So again, once the questions of *representation*, *inference* and *discovery* were answered, the natural, logical transition was once more one of generalization, how can we generalize even more, and make it even more compact? One generalization is contextual weak independence (CWI) [4]. The result from utilizing CWI is an even more generalized version of CSI, where context is found within a CSI. It is a further generalization of context for more specific subgroups of data. It is slightly less intuitive in practice but the computational saving can be superior, although the independencies are harder to find. Essentially, the results were that once again, an adequate representation is possible, inference can be carried out (once again, with a new operator, this time called the weak-join) [6], and discovery methods have also been suggested [7].

2.5.2 Further Generalizations of Independencies

CI is a way to decompose JPDs. As a result, more efficient probabilistic inference may be possible. CSI is a general case of CI, which can yield better inference. CWI is a generalization of CSI, which in turn, has shown to yield even better inference [6]. From that pattern, a natural question arises. It seems that better probabilistic inference is possible the further we decompose, *so why not break down n-tuple distributions into n individual distributions?* That would be the extreme case where every distribution only has one unique record.

Decomposing an n-tuple distribution into n distributions is not necessarily efficient since the decomposition is not based on any independence between variables. It is an arbitrary decomposition. When a query comes in, the distribution must be reconstructed before the query can be processed since no independencies held initially. Therefore, no savings are obtained and marginalizing out a variable from the distribution will require the same number of additions as it would have, had the distribution not been decomposed into singleton decompositions.

2.5.3 Refinement with CSI

By taking advantage of context-specific independence (CSI), the indirect compact representation of the joint probability distribution (JPD) was improved in many cases with an even more compact representation. Subsequently, CSI was further generalized with CWI. However, consideration of the potential cognitive interpretations of this type of independence was given little attention. This is the problem we address in the next chapter.

Chapter 3

Attribution Theory

In this chapter, we discuss existing theories of attribution, and discuss how human adults attribute causes to effects. Current methods for categorizing particularities about people's unexpected behaviours, namely biases in attribution, are also discussed in this chapter. As stated in the introduction, one of the main contributions of this thesis is the use of *context-specific independencies* (CSI) to distinguish between situational factors and dispositional causes when examining a situation regarding people's personalities.

Recall that in social psychology, attribution theory deals with explanations of behaviour. The theory assumes that people try to understand why others do the things they do by attributing multiple causes to that particular behaviour. For example, consider the issue of illegal drugs. You have a collegue who you consider is an intelligent, well-organized, respectful, kind person, whose lifestyle and ideas you admire. This person has confided in you in the past and you value his/her friendship very much. The individual in question is also a regular consumer of illegal drugs. As a friend, you may be disappointed about this type of behaviour but more than anything you are also worried about health and legal risks your friend is taking. Naturally, you wish to understand why the behaviour persists. Upon reflection, you realize many factors may be contributing to, or causing your friend's behaviour. Possible factors include a simple character flaw, an addiction, a misperception of the effects of drugs as a stress reliever, peer pressure, etc. The true explanation matters to you, and in general, the conclusions we make about an individual's behaviour determine our reactions to the particular individual. Even more generally, our interactions with other people have roots in the attributions or explanations we make of what they say and do. Attribution theory explores how people associate causes to events and how their subsequent actions will be affected by this cognitive perception, namely the event they choose as causal.

3.1 Determination of Attribution

In theorizing about attribution, we are interested not only in the true cause of behaviour but also in how we, as human adults, assign a cause to another person's behaviour, whether the inferred cause is true or not. For example, suppose the true cause of your friend's regular consumption of illegal drugs is a false belief, acquired in a magazine article, about the benefits it may have on acne, in combination with her quasi-obsession with clear skin, and a family history of oily skin. However, attribution is equally interested in understanding how other individuals would attribute the behaviour. For instance, you might attribute the cause of her behaviour to a completely different factor, such as her peers. Two fundamental principles [37] play a key role in determining the attributions people make, namely the *discounting principle*, and the *covariation principle*.

3.1.1 Discounting Principle

The discounting principle can be thought of as the principle of a likely cause. Kelley [37] states that in general, people tend to accept the more frequently observed scenario as being causal, and disregard all other possibilities. For example, if a floor shop supervisor monitors a worker closely, then this boss' superior considers this a plausible explanation for the worker's hard work (i.e. supervision), and will therefore likely disregard other plausible reasons for the behaviour, such as the worker's motivation to do a good job [65]. According to the discounting principle, once a plausible cause for a particular behaviour is identified, the search for an explanation ends. People are more interested in *a* cause than *the* cause.

3.1.2 Covariation Principle

The covariation principle is applicable when two events are associated over a series of instances. If event B always happens when event A occurs, and does not take place when A is absent, people often infer that one causes the other. This principle also applies to the negation. If event B never happens when event A occurs, but always occurs in the absence of A, people will tend to infer that event A causes B not to occur. An example of the covariation principle is as follows: if a person becomes visibly nervous when they are around water, but seems calm otherwise, we are likely to attribute the nervousness to the person's feelings towards water, and not to a personality trait. We will not categorize the individual as simply being a nervous person in general.

Before moving on to theories of attribution, based on the principles of discounting and co-variation, we distinguish two general categories of causes people use to understand other people's behaviour: *situational* and *dispositional*.

3.2 Causes for Attributions

Heider [66], who first argued that attributions are fundamental to social relations, suggests that as a general rule, we make *internal* attributions to explain other people's mistakes, such as "the event was due to something particular about the person", while we tend to make *external* attributions to explain our own errors, such as "the unfortunate event occurred due to some uncontrollable factor in the environment". For example, if our favourite NHL team loses a game, we, as supporters of the team, tend to attribute the loss to the team, "the team didn't play a very good game", while if they win, we are more likely to say "we won(!)". In practice, we tend to go through a two-step process of attribution [66]. We start by making an automatic internal attribution, followed by a slower consideration of external factors, to see whether an external attribution is more sensible. That is, we revisit and modify our mental maps in accordance with our data. However, with the time constraints and distractions we deal with constantly, we often do not get to that second step of assessing whether an external attribution is more appropriate. This problem is known as *automatic believing* [67, 68] in the social psychology literature. Automatic believing explains why internal attribution is more likely than external attribution in general.

Since Heider's [66] original definition of internal and external attributions, theories have been developed that try to explain how people form *situational* or *dispositional* attributions. The terms situational and dispositional are essentially a refinement of Heider's original terminology. A cause that explains actions in terms of a social setting or environment is defined as *situational*. For example, if we observe a person purchasing a vanilla ice cream cone at the ice cream parlour, when flavour availability on the given day is limited to vanilla, we will attribute the cause to the situation. We will likely not conclude that the individual prefers vanilla, or really wanted vanilla ice cream. On the other hand, *dispositional* causes are based on characteristics of the person in question. For example, if we witness an individual at the ice cream parlour selecting vanilla from 24 ice cream flavours, we are more likely to conclude that the person prefers this particular kind, or wanted that flavour for some personal reason. Dispositional causes are also referred to as *personal* causes.

Keeping in mind that a true disposition is an internal factor about a person and a situational cause is an attribution to the environment, therefore external, we turn to attribution theories that aim to explain how people form situational and dispositional attributions. The theories build on the principles of discounting and covariation (see Section 3.1). We present here the three theories that have proved most influential [38].

3.3 Theories of Attribution

In this section, we discuss three theories of attribution aimed at explaining the choice people make between situational and dispositional attributions. The first is Jones and Davis's model of correspondent inferences [39], which concerns a single social interaction. The second is Kelley's covariation model [40], which consists of a relationship over time, and finally the third is Weiner's model of achievement attributions [41], which deals with situations involving success or failure.

3.3.1 Theory of Correspondent Inferences

Jones and Davis [39] extend Heider's theory; they introduce cues that are used to infer the cause of an action. They claim we use information about the behaviour of a person as well as effects of the particular behaviour to make a correspondence inference, in which the behaviour is either attributed to a disposition (personality trait) or a situation, and is based on a sole observation. The theory of correspondent inferences studies how three main categories of "logical" cues, namely *free will*, *non-common effects*, and *non-conforming* are used to infer the cause of an action. Two "non-logical" cues are also used, namely *hedonic relevance* and *personalism*. Based on the cues, if the act corresponds to a true characteristic of the actor, we say a correspondent inference was made. The cues help us decide if a correspondent inference is to be made, or if the behaviour should simply be attributed to situational factors.

Logical Cues

Free Will: When looking for an explanation of someone's behaviour, we focus upon *freely chosen* actions and mostly ignore ones that are clearly coerced. For example, when a salesperson acts in a seemingly exagerated friendly manner, we are not likely to make a correspondent inference, since it is quite plausible that the actor's behaviour is coerced by simply following a manager's orders. However, if a perfect stranger with no intentions to sell us anything acts very kindly towards us, we are more likely to believe that the act was freely chosen, and that the stranger is truly a genuinely kind person.

Non-Common Effects: When attempting to make a decision about the possibility that someone's behaviour corresponds to an underlying disposition, we also make use of the *non-common effects principle*. A non-common effect is a behaviour that can easily be attributed to a person, due to its large deviation from the norm. Internal attributions are more easily asserted when fewer rare or *non-common effects* between the choices of the person under investigation are present. For example, consider a person choosing between two jobs that are very different in many ways, including salary. If the lower salary job is chosen, we cannot conclude that the person is not money-driven, since many other aspects of the different jobs may have led to the decision, or many *common* effects are present. When alternatives have many aspects in common, there are fewer things that differentiate them to help us make inferences about the person. When the behaviour is not what we would have forecast, we assume that it is due to their internal preferences or character traits. Thus, if more non-common effects are present, we infer underlying disposition. For example, if a person is making a decision between two jobs that differ only by location and salary, it is natural for us to attribute the person's final decision to their individual preferences. If the lower salary job is chosen, it is easy to infer that the person is likely not money-driven [69]. In this example, very few aspects differ between the two options, therefore the candidate's choice can easily be attributed to his/her personal preferences.

Non-Conforming: Also known as social desirability, another consideration when attempting to decide if there is a correspondence is *non-conformism.* We are more likely to conclude that there is correspondence if the effects of the behaviour are socially undesirable. Politeness illustrates this idea clearly. If someone is noticeably rude in a social situation, we are likely to conclude that this person is simply an unpleasant person. On the other hand, if someone demonstrates conventional politeness, we will not feel as though the event in question has taught us anything new about the individual. In other words, we have no new evidence to add to our knowledge base. Once again, we deal with uncertainty due to incomplete information. We don't know enough about the person's behaviour in different situations to confirm whether his/her politeness is genuine.

Non-Logical Cues

The non-logical cues are biases that arise from our own personal reactions to an event. If the act had a direct impact on ourselves as an individual, we tend to make more confident correspondent inference, which is defined as *hedonic relevance.* Also, if we feel the actor *intended* to benefit or harm us, we will, once again, be more confident in a correspondent inference. This second non-logical cue is defined as *personalism.*

In summary, to make correspondent inferences, information is required about five factors: whether the behaviour under investigation is voluntary and freely chosen (*free will*), what is unexpected about the particular behaviour (*non-common effects*), whether the behaviour is socially desirable (*non-conforming*), whether the behaviour impacts the person making the inference (*hedonic relevance*), and finally, whether the behaviour is of personal interest to the person doing the inferring (*personalism*).

Classic Experiment in Correspondent Inferences

A series of controlled experiments has proved consistent with the model of correspondent inferences [70, 71]. In one experiment [71], participants listened to an interview with a job applicant. Half the participants were led to believe that the applicant was interviewing for a job as a member of a submarine crew, while the other half were led to think the interviewee was applying for an astronaut job.

The submarine crewmember job was described as one requiring a friendly, outgoing, and cooperative personality, while the astronaut job required a more reserved, inner-directed, quiet individual. The other essential element for the attribution activity was that half the participants were exposed to an interview with an outer-directed, friendly applicant, while the other half heard an interview with a more reserved, inner-directed person.

The interesting finding in this experiment is its consistency with the theory of correspondence attribution. When participants were prompted to make judgments of what the applicant's personality was really like, those hearing role-consistent behaviour rated the person near the neutral point on the relevant personality dimensions, which indicates much uncertainty. On the other hand, when the behaviour of the applicant did not match the type of personality consistent with the job description, participants attributed the behaviour to the applicant's personality. So the inner-directed person applying for the submariner job was confidently rated as quiet, while the outer-directed astronaut applicant was confidently rated as friendly and extroverted. In such cases, the participants didn't have any alternative explanations for the behaviour of the applicants. Since the behaviour was not a reflection of job requirements, correspondent inferences were confidently made.

Although several experiments have supported the model of correspondent inferences, there are several limitations to the theory. First, there is a possibility that the observers making the inferences (us, on a daily basis) decide on the commonality of effects by making comparisons between the person's actual behaviour and non-chosen actions rather than intentional ones. Nisbett and Ross [72] have studied this problem and they have concluded that people actually rarely consider that non-chosen actions are in fact "non-chosen". The second limitation ties in with the first. We sometimes make correspondent inferences even when we judge the actor's actions to be unintentional [73]. Finally, the process involved in drawing inferences about other people's behaviour is far more complex than correspondence inference theory suggests.

3.3.2 Theory of Covariation Model

Kelley proposes that people are intuitive scientists. He argues that people make causal attributions based on the information that is available to them at the time of decision. The type of causal inference Kelley suggests relates to the principle of covariation and has roots in the work of the philosopher Mill's "joint methods of agreement and difference" [74]. This method is inductive and proposed as being rational (normative). Note that once again, we are in a situation of incomplete information. When relevant information is available to us from different sources, we can detect the *covariation* of observed behaviour and the possible causes for the behaviour in question. Thus, Kelley's model [40, 75] addressed what information is used by the perceiver to arrive at a causal attribution. It is a logical model for judging whether a particular action should be attributed to some characteristic of the person making the attribution (internal) or rather to the environment (external).

Based on his model, called Kelley's cube, Kelley proposes that when making causal attributions, the perceiver is faced with one of two types of information, namely *covariation information* or *configuration information*, both described below.

Covariation Information The perceiver possesses information from *multiple observations*, at different times and from different situations, and can perceive the covariation of an observed effect and its causes. For example, if a particular individual is consistently unpleasant to you, it is possible that the individual is simply an unpleasant person, or it could be that you are not a likeable person. If you hold information about this person's attitude towards others, and you are aware of how others treat you, you are then in a position to make a decision based on covariation.

Configuration Information The perceiver is faced with a *single observation* and must take account of the configuration, or the current available information. For example, if you see someone running over a dog with their car, you will use information about the configuration or arrangement of factors to form an opinion of the driver in question. If the roads were slippery or if the visibility was reduced, this increases the likelihood that you will make a situational attribution of the driver's behaviour. On the other hand, if driving conditions were optimal, you are much more likely to make a dispositional attribution of the driver's behaviour. You may conclude that he is either a poor driver or perhaps an inconsiderate being, or even a very distracted person.

Kelley's principle states that *an effect is attributed to a condition that is present when the effect is present and absent when the effect is absent.* The mathematical basis of his theory is based on the statistical technique ANOVA. We examine the changes in a predictive variable (the effect) by varying explanatory variables (the conditions).

Kelley proposes three dimensions as independent variables in his model: *Person (P)*, *Stimuli (S)*, and *Time/Modalities (T)*. Those dimensions are expressed in his Person $\times$ Stimuli $\times$ Time/Modalities cube. Kelley has also introduced to his model three information variables, namely *consensus*, *distinctiveness*, and *consistency*, to measure covariation along the three different dimensions. According to the model, which has later been revised by Kelley and Michela [76], the cause of a given P's response to a certain S on a particular occasion T is inferred based on the individual's perception of the degree of:

- *consensus* between P's response to S and other people's response to S (on occasion T);
- *distinctiveness* of P's response to S from P's responses to other stimuli (on occasion T);
- *consistency* of P's response to S on occasion T with P's response to S on other occasions.

In the first case, we search for particularities about a specific person, in the second case, we search for a particular stimulus, and in the third case, we attempt isolate a particular occasion.

Table 3.1 illustrates this principle. For instance, on the dimension of consensus, we are analyzing responses to a particular constant stimulus over a number of people, while keeping the time constant.

Table 3.1: Interactions between the dimensions and information variables in Kelley's Covariational Principle

	person (P)	**stimuli (S)**	**time (T)**
Consensus	Variable	Comparison Constant	Invariable
Distinctiveness	Comparison Constant	Variable	Invariable
Consistency	Comparison Constant	Invariable	Variable

The comparison constant is an instance of the factor for which we verify every instance of the variable. The invariable factor is simply kept constant to avoid noise in the interactions, but is not of interest to the result. The table helps to mentally conceptualize the analysis of the three dimensions over the three information variables.

We note here that the information variables of consensus and consistency are inversely proportional to covariation. A high level of consensus indicates a low covariation between the particular person and the effect under investigation. On the other hand, distinctiveness is directly proportional to covariation. A high level of distinctiveness indicates a high covariation.

From his proposed model, Kelley made three specific attributional predictions for configurations of the three information variables (*consensus*, *distinctiveness*, *consistency*). Of the three configurations, the first should lead to *person attribution*, the second to *stimulus attribution*, and finally the third to *circumstance/time/occasion attribution*.

Low consensus, low distinctiveness, and high consistency (LLH) should lead to a person attribution, high consensus, high distinctiveness, and high consistency (HHH) should lead to a stimulus attribution, and finally low consensus, high distinctiveness, and low consistency (LHL) should lead to a circumstance attribution. In terms of covariation, which is what we are truly interested in, the LLH configuration for example, indicates high covariation along the person dimension of the cube and low covariation along the two others. In the case of HHH, we have high covariation along the stimulus dimension and low covariation along the two others. Similar reasoning can be used for the remaining dimension.

Recall from statistics that ANOVA is a general technique that can be used to test the hypothesis that the means among two or more groups are equal, under the assumption that the sampled populations are normally distributed. An example of the analysis of variance model of covariation from McArthur's [77] work is presented in Figure 3.1. In this example, the goal is to attribute the statement "John laughs at comedian" to either the person (John himself), the stimulus (the comedian), or the circumstances (a comedian is on stage). We have mentioned that an LLH configuration, namely low consensus, low distinctiveness, and high consistency, should lead to a person attribution. In Figure 3.1, a low consensus between people implies that *few* people are laughing. A low distinctiveness between entities implies that there is nothing particular about the stimulus (the comedian) that distinguishes him/her from all other stimuli (other comedians). Finally, high consistency implies that John laughs at all comedians, and not only the particular one on stage in this instance. With these three values, we can safely attribute the laughter to a particularity about John, i.e. John finds comedians funny.

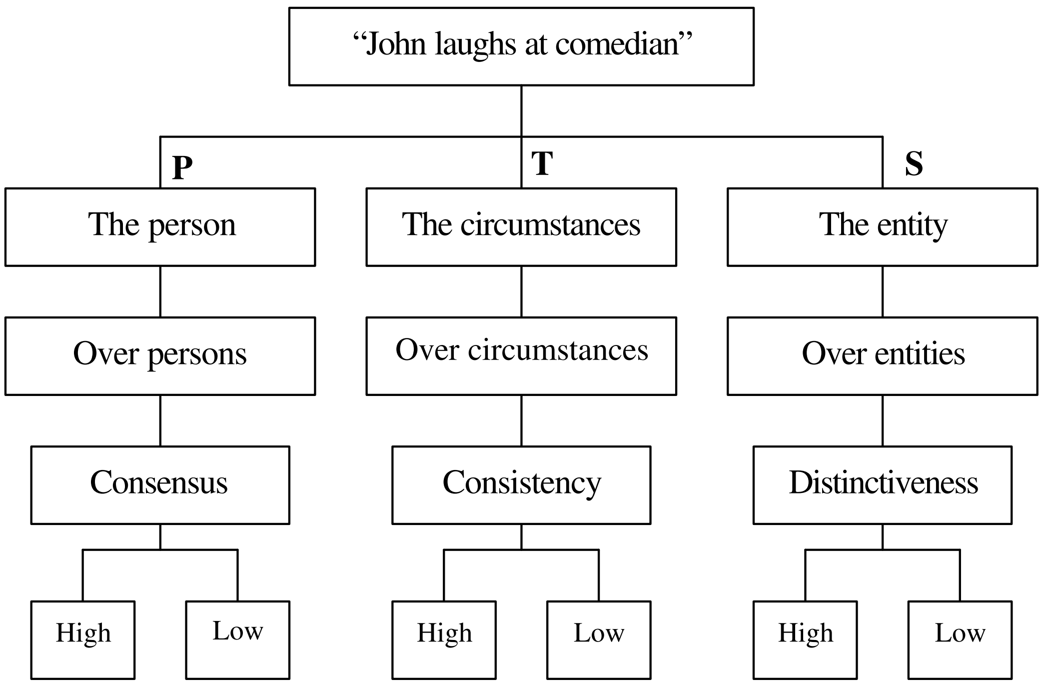

Figure 3.1: McArthur's Comedian example of Kelley's Covariational Principle

3.3.3 Theory of Achievement Attributions

People's social experiences are often evaluated in terms of successes and failures. Those evaluations can be defined in very concrete ways: being on the school honour roll, passing the talent show audition, winning a hockey game, getting a new job, etc. However, we also experience success and failure in more subtle ways: being respected by our peers, having our presence requested at social events, etc. Many having experienced divorce must deal with feelings of failure [78]. Also in dealing with loneliness, attributions regarding success and failure are crucial [79]. Weiner's *theory of achievement attributions* uses the covariation and discounting principles (see Section 3.1) in assessing situations involving success and failure.

Weiner's [41, 80] theory suggests that achievement attributions are constructed in two steps. First, we must decide if the success or failure was caused internally (something about the actor) or externally (something about the situation) (see Section 3.2). Second, we need to decide whether

Table 3.2: Attributions about achievement: A final grade in an undergraduate class

	Stable		Unstable	
	Internal	External	Internal	External
Controllable	typical effort	professor dislikes student	unusual effort	unusual disruption by other student
Not controllable	lack of ability	task difficulty	mood	luck

the cause was *stable* or *unstable* in nature, where a stable cause is one that is consistent over time. An example of a stable cause could be, for instance, consistent effort by the actor throughout the duration of a controlled period, such as a university semester.

In the later version of Weiner's theory [80], a third dimension was added: whether the actor had the power to control the occurrence of the event. For example, the mood of your instructor when he/she is marking your term paper is something you cannot likely control. In total, the theory allows for eight different explanations for success or failure. An example, taken from Weiner [81], is presented in Table 3.2.

The problem then becomes one of predicting the attributional choice. We need a way to decide what caused the particular success or failure. Freize and Weiner [82] conducted an experiment in an attempt to find such an explanation. The participants were given information about a situation in which a person had experienced either success or failure at a task. The following information was provided to the participants:

- whether the person had succeeded or failed at a similar task in the past (consistency data - *time*);
- whether most people had succeeded or failed at a similar task in the past (consensus data - *person*);
- whether the person had succeeded or failed at their next attempt at the same task (distinctiveness data - *stimuli*).

Given the above information, participants were asked to explain why they believed the actor had succeeded or failed at the task. The majority of participants attributed the outcomes to:

- *internal* causes when performance was *consistent* with past performances by the *same* actor and *different* from outcomes of *other* actors (low consensus);

- *external* causes when *consistency* for the actor was high and *consensus* was high as well.

3.4 Tendencies in Making Attributions

Attribution theory is based on the assumption that humans are rational thinkers, and therefore always use available information in a rational way to make decisions. Based on that belief, the human subject is a "naive scientist", since we believe in the systematic search for relevant information, followed by a *rational* logical explanation of behaviour. Although the theory seems to do fairly well in practice, research shows that we often make attributions that are *not* based on rational conclusions, thus the "naive scientist" is fallible. These irrational attributions create *biases* in the attribution process, which can have a significant impact on our conclusions. We discuss some biases in attribution here.

3.4.1 Correspondence Bias

Recall the discussion of Correspondent Inferences in Section 3.3.1 according to which we infer that people's actions refer to their true disposition (internal factors). According to the theory, people make the inference between other people's actions and their true disposition based on logical evidence, a rational reason. However, even when logic and evidence suggest otherwise, people appear to have biases that lead them to conclude that the person who performed the act was predisposed to do so. We ignore the role of the environment in the person's decision to engage in the behaviour. This is referred to as *Correspondence Bias*. For example, in an argument, if a person speaks hurtful words to another, the victim is more likely to attribute the instigator's rude, inconsiderate behaviour to their personality rather than to something they may have said to hurt them first.

Fundamental Attribution Error

In the literature, this tendency to exaggerate or overestimate the importance of dispositional factors, and to undermine the importance of other people and other angles of a situation has been referred to as the *Fundamental Attribution Error* [42]. Even when people are instructed to argue a certain position, and when this fact is known to the observers, observers still tend to attribute the behaviour to the actor's true disposition [70]. As an illustration, consider a lawyer in a courtroom. Even though observers are aware that the lawyer's job is to defend his client regardless of his personal opinions, he will still be perceived by the public as sharing the guilt.

Actor/Observer Bias

Another aspect of the correspondence bias is that people are more likely to attribute other people's actions to character traits (disposition), and their own actions to the specific situations they face. This is referred to as the *actor/observer* bias. Note that this bias doesn't contradict the correspondence bias; actors and observers tend to attribute behaviour to dispositions, only *more* so for *other* people's behaviour [83].

A well-known experiment [84] illustrates this idea. A group of male students were asked to write about why they had chosen the university major they did, as well as why they chose to be with the girlfriend they were with. The students were also asked to repeat the same activity, except the second time, making attributions about their friends' choices of University majors and girlfriends.

When making attributions about themselves, the students mainly attributed their own decisions to external reasons, e.g., "I decided to major in psychology because it is interesting". However, when making attributions about their friends, they associated the decisions to dispositional factors, e.g., "He's going out with her because he's insecure".

Reasons for Actor/Observer Differences in Attribution

According to Heider [66], people make different attributions for themselves and for others mainly because they have different perspectives on the same event. The "actor's behaviour captures the attention of observers by engulfing their field of perception" [66], whereas actors are generally unable to observe themselves, and they are more aware of the situation they are in than they are of themselves. Interestingly, however, when a situation is videotaped and showed to the actor, the actor tends to make attributions based more on disposition than without the video replay [85].

Another reason explaining the actor/observer differences is in the access to information. Actors and observers have access to different information and actors know how they acted in a given situation. In general terms, actors have access to *consistency* and *distinctiveness* information (see Section 3.3.2) and are better able to judge how different situations influence their behaviour. Without this additional information, observers can only make correspondent inferences (i.e. friendly people do friendly things - see Section 3.3.1).

3.4.2 Self-Serving Bias

A second bias in rational attribution is one that helps us protect our ego and self-esteem, namely *self-serving* bias. Human adults tend to attribute their successes to internal factors and detach themselves from their failures by attributing them to external factors. A straightforward example is one of students receiving grades as a measure of performance assessment. Students receiving high grades are likely to attribute the success to internal factors such as ability or effort, while students receiving poor marks tend to attribute the perceived failure to external factors such as task difficulty or bad luck [43].

We are more likely to take credit for our success and deny responsibility for our failure when we have explicitly chosen to engage in the particular situation for which we are making attributions, and are highly involved with. This also applies to activities where the results of our performance is public rather than private [86].

3.4.3 Defensive Attributions

A third category of biases deals with our need to feel secure. An experiment by Walster [44] supports this bias referred to as *defensive attribution*. In Walster's experiment, a driver (Lennie) parks his car at the top of a hill, the hand break comes loose, the car rolls down the hill, and causes damage. When asked to indicate to what extent they attributed responsibility to Lennie, participants held Lennie responsible for the incident when the damage was severe or when someone was hurt. There is no logic in such conclusions since Lennie's negligence (to get his brake checked) was no different when the gravity of repercussions varied. Similar findings have been reported in other experiments as well [87].

The interesting finding in the above-described experiment is that the severity of consequences seemed to affect the attribution of responsibility. It has been suggested [44] that we act defensively to disassociate ourselves from the possibility of a threatening event. We often blame the victims of violent crimes or accidents, since blaming the incident on bad luck or "God's Will" enforces the idea that such a situation could possibly happen to us. Shaver [88] highlights two types of scenarios that lead us to adapt this defensive behaviour:

- the situation we are assessing is similar to our own;
- the victim is similar enough to ourselves that we could imagine ourselves in their position.

3.4.4 Illusion of Control

Although much of what happens to us in life is beyond our control, our response to this intimidating thought is to cling to an *illusion of control*, which can be defined as an exaggerated belief in our own capacity to determine what happens to us in life [45]. For example, people often like to select their own lottery ticket at the store due to the false illusion that maybe they know how to select the winning ticket. Wortman [89] conducted an experiment in which he provided each participant with a can containing two different coloured marbles, each marble representing a different prize. Some participants were informed about which one of the two marbles was associated with the desirable prize, while other participants were not. Also, participants either got to select their own marble (without looking) or were simply given one. In either case, the participants had no control over the marble they were getting, but those who got to select their own marble (although without seeing which one they were selecting), attributed more responsibility to themselves.

Throughout this chapter, we have discussed how people make attributions, and what causes people to make certain attributions. Causes of attribution are very important as we rely heavily on them to *predict* future attributions. Furthermore, to predict future attributions or to justify current attributions, theories of attribution are available [39, 40, 41], as we discussed in earlier sections of this chapter. Although those theories can help in predicting attributions, many deviations from the theories' predictions exist. As we discussed in this chapter, these are called biases in attribution, and they arise when a particular behaviour can not be explained by existing theories.

In this thesis, we show how causes of attribution can be determined from contextual consideration on probability distributions. We present those results in Section 5.3 of Chapter 5. The next chapter focusses on the preliminaries to the third theme of the thesis, namely an existing model for causal attribution by human adults.

Chapter 4

Cheng's Probabilistic Contrast Model

In this chapter, we discuss Cheng and Novick's probabilistic contrast model dealing with causal attribution of human adults. More precisely, Cheng and Novick aim to understand what information humans actually use when making attributions. In their model, they address the importance of the role of context in determining attribution, although their model is unable to discover context from a data distribution. They use a notion of *focal sets* to distinguish between contexts, where focal sets are predetermined by a human expert. We describe focal sets in further detail later in this chapter. In Chapter 5, we propose a method for identifying focal sets from probability distributions. This focal set discovery technique may be very useful, as contexts are often unknown before attributions are made.

For the moment, we discuss the different aspects of causal attribution that need to be taken into account in building Cheng and Novick's probabilistic contrast model. Then, we present their initial purely covariational contrast model, followed by a discussion of focal sets. Finally, for completion, we discuss Cheng's improved contrast model, Power PC.

4.1 High Level Attributional Considerations

This section addresses Cheng and Novick's preliminary question about attribution: *is the process of attribution inherently biased?* [19]. To answer this question, they claim it is essential to establish a clear distinction between *process* and *data*. This is the first topic we address in this section. Then we discuss the biases Cheng and Novick considered in building their model. The biases described are simply a specialized set of the more general categories of biases discussed previously in Chapter 3. Finally, we close this section with a discussion of Cheng and Novick's initial response to bias.

4.1.1 Process versus Data

When asked if the process of causal attribution is inherently biased, Cheng and Novick argue that to answer this question, it is essential to make the distinction between the data on which the causal inference process operates and the process of inference computation itself, to address the issue that people's causal inferences are rational (normative).

The two cognitive psychologists also claimed that previous research on covariation-based causal inference had either failed to make this distinction between process and data or had not accurately identified what information people use to make causal assessments [19]. Consequently, it becomes almost impossible to establish whether "observed biases occur in the inference process per se or in the data on which the inference process operates" [19].

Cheng and Novick compare observed bias to the highly influential work of Henle [90] on logic. "Just as false conclusions in deductive reasoning can be reached by the use of valid deductive rules when the premises on which the rules operate are false [90], so observed bias in inductive reasoning may be due to the nature of the input (i.e., the set of information on which inference is computed and the pattern of that information) rather than to biases in the process of inference computation itself." [19].

4.1.2 Deviations from Normative Covariational Statements in Causal Attribution

In the 1980s, research on causal attribution found deviations from the predictions of the covariational models [91, 51, 92, 72]. Three types of deviations noted were:

- bias against using consensus information;
- bias toward attributing effects to a person;
- tendency to make multiple other attributions unpredicted by the model, conjunctive ones in particular.

Once again, note that the three biases above are consistent with the ones presented in Chapter 3. Their format is slightly different to allow experimentation with Kelley's cube (see Section 3.3.2). The three biases we discuss below are explained from the point of view of Cheng and Novick [19] and refer to issues with the solely covariational model of Kelley's Analysis of Variance [37] as well as the one of Jaspars, Hewstone and Fincham's Inductive Logic Model [92]. Cheng and Novick initially thought the three problems (biases in attribution) we describe below would be alleviated by their purely covariational probabilistic contrast model.

Bias Against Using Consensus Information

Researchers in causal attribution agree that there is an underuse of consensus information [47, 91]. McArthur [77] conducted an experiment where he varied two levels (high, low) of consensus, distinctiveness, and consistency information. He found that consistency information accounted for 41% of the variance in circumstance attribution and distinctiveness accounted for 12% of the variance in stimulus attribution. However, consensus information only accounted for 6% of the variance in person attribution. Recall from Table 3.1 in Section 3.3.2 that the variable dimension for the *person* dimension is *consensus*, the one for *stimuli* is *distinctiveness*, and finally, the variable dimension for the *Time/Modalities* dimension is consistency. Similar underuse was observed in predicting total variance in causal attribution. Consistency accounted for 20% of the variance, distinctiveness accounted for 10%, and finally, consensus information accounted for a mere 3% of the variance. On the other hand, to blur the situation, it has been noted that this consensus bias doesn't occur under all conditions [93, 94, 95, 96, 97, 98, 99, 100].

Bias Toward Attributing Effects to a Person

The bias toward attributing effects to a person has also received a lot of support [77, 49, 96]. For instance, in a study conducted by Jaspars et al. [92], they noticed that 82% of the subjects in McArthur's [77] study attributed an effect to the person when the presence of the person was necessary and sufficient to produce the effect. However, only 62% of subjects made a stimulus attribution when the presence of the stimulus was necessary and sufficient to produce the effect. Finally, in the case of circumstance, only 33% made the circumstance attribution, when its presence was necessary and sufficient to produce the effect. Also, like consensus bias, person bias doesn't occur in all situations.

The bias toward person attribution is consistent with Ross's [42] "fundamental attribution error" (see Section 3.4.1). Briefly described, the fundamental attribution error refers to the human tendency to attribute behaviour to enduring dispositions, such as attitudes or personality traits. We underestimate influence of situational factors on others' behaviour, and we overestimate influence of dispositional factors on others' behaviour.

Tendency to Make Multiple Other Unpredicted Attributions

The third problem deals with a tendency for the subject to make other unpredicted attributions, conjunctive ones in particular. When the subject must make a causal attribution that deals with more than one dimension of Kelley's cube, the predictions made by covariational methods are not reliable. Such deviations are summarized in [19].

4.1.3 Cheng and Novick's Initial Response to Bias: Incomplete Information

In social psychology, more specifically in attribution theory, when we talk about "incomplete information," we refer to configurational information, which we discussed previously in Section 3.3.2.

There have been many suggestions as to why biases occur in attribution and why they appear so capriciously under some conditions but not others [47, 48, 49, 50, 51, 92]. Cheng and Novick [19] suggested that the infiltration of bias was due to incomplete information.

Recall from Section 4.1.1 the distinction between process and data. In testing covariation-based models, most experiments have assumed that the input they were getting from subjects were consensus data, distinctiveness data, as well as consistency data. However, one problem that was overlooked, which has since been noted by several researchers [19, 101, 92, 96], is that information on the three variables of consensus, distinctiveness, and consistency were assumed to cover all information required to make causal attributions. In reality however, the information variables only contain a subset of the information required for accurate attribution.

Consider the following illustration. The variable of consensus represents the degree of agreement between the person under investigation, and other people in their responses to a particular stimulus on a particular occasion (see Table 3.1). Unfortunately, that has often been wrongly interpreted to mean the amount of response agreement between the person under investigation and other people with respect to *all* stimuli on *all* occasions. Therefore, when considering the eight regions of information in Kelley's [40] cube as in Figure 4.1, keeping in mind that Region 0 is the target event to be explained, consensus information covers only Region 1. With similar reasoning, we allocate distinctiveness to Region 2 and consistency to Region 3. In Figure 4.1, there are L persons, represented by the vertical axis, M stimuli, represented by the horizontal axis, and finally, there are N occasions, each represented by an individual cube.

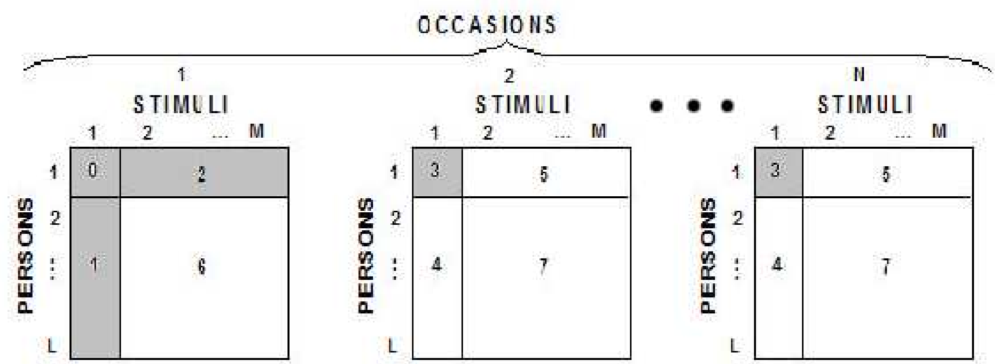

Figure 4.1: The eight information regions in Kelley's cube. Shaded regions indicate configurational information

Given the above reasoning, when conducting experiments with configurational information (see Section 3.3.2), it is not possible to know what assumptions subjects are making spontaneously about the occurrence of an effect in the nonconfigurational parts of the cube (Regions 4-7) in Figure 4.1.

From the above analysis of the attended regions on Kelley's cube, Cheng and Novick [19] suggest that since the attribution of causality is a function of the data on which the rules of inference operate, as well as a function of the inference rules themselves, the presumed biases found in earlier experiments may be a reflection of the assumptions made by the subjects in the unattended regions of the cube (other contexts), rather than of the inference process. Subjects have no choice but to make assumptions regarding the pattern of information for the unspecified cells in the cube.

In the cognitive literature, it has been found that often, people use other information than the data they are provided when analyzing a situation [52, 53, 54].

4.2 Description of the Probabilistic Model of Causal Attribution

In this section, we describe Cheng and Novick's initial purely covariational contrast model. We discuss how contrasts are computed, we make a distinction between a main-effect contrast, and an interaction contrast, and we distinguish between facilatory causes and inhibitory causes. Finally, we discuss Cheng and Novick's focal sets for deciding what factors should be taken into account in different situations where attribution takes place. This component of the model is vital as erroneous conclusions are almost inevitable if the wrong factors are assumed to be used by the person making the attribution. In their model, Cheng and Novick decide on the factors chosen to form a focal set before any attribution is made. They use their human expertise to determine relevant focal sets. We revisit focal sets in Chapter 5 and propose a method to discover focal sets from probability distributions.

4.2.1 A Contrast Model

Cheng and Novick [19] describe their model as being a probabilistic analogy of statistical *contrasts*. The word contrast is used in this context to signify the contrasts between a value on a particular dimension of the cube (Figure 4.1) and other values on that same dimension. It also refers to contrasts involving specific combinations of values as opposed to other combinations of values. The interpretations we can infer from these contrasts are particularities about a specific person, or particularities about the person in conjunction with a particular stimulus. Those observations are the results of *specific* contrasts rather than *overall* main-effects or interactions in an instance of an ANOVA.

Due to Cheng and Novick's shared belief that human adults do not mentally perform a process analogous to complex quantitative computations underlying statistical contrasts, their model does not require such complex mental computations. With the probabilistic contrast model, the subjects are simply required to be capable of making estimations and comparisons of proportions. Research shows that naive subjects perform adequately at this type of task [102, 103, 104].

4.2.2 Computations of Contrasts

To compute contrasts in a causal scenario, we must assume we are dealing with attended dimensions that are present in the event we are trying to explain. We need to make an assumption regarding the information we suspect is available to the subject making the attribution (human must decide appropriate context). Cheng and Novick present two types of contrast, namely a *main-effect contrast* and an *interaction contrast*.

4.2.3 Main-Effect Contrast

A cause C of an effect E is defined to be a factor for which the probability of the observed effect, $p(e|i)$, when factor i is present is significantly greater than $p(e|\neg i)$, the probability of the effect when factor i is absent. The proportions computed for a main-effect contrast are approximations of the probabilistic conditional independencies of the effect given the presence or absence of event i. In other words, a factor is considered to be a cause if its presence statistically significantly increases the likelihood of the effects, where the presence of the effect is assumed. The main-effect contrast can be expressed as follows:

$$\Delta p_i \;=\; p(e|i) - p(e|\neg i). \tag{4.1}$$

We illustrate the idea of main-effect contrast with Kelley's cube, as depicted in Figure 4.2, where we assume we are searching for Person attribution. If the proportion of cells in which the target effect occurs for Person 1 (top of cube in Figure 4.2) is larger than the corresponding proportion for other persons (remainder of cube in Figure 4.2), then we attribute to Person 1 a target effect involving Person 1. In other words, we attribute the effect to Person 1, since the effect occurs consistently in the presence of Person 1, but not other Persons.

So far, we have limited the discussion to causes that consist of a single factor. For example, consider a situation where the effect for which we seek the cause in "the door slams shut" with the constantly present factor of "window is open". In this situation, we consider "wind factor" to be the variable factor suspected to be the cause. In the absence of wind, the door does not slam shut, whereas in the presence of wind, the door slams shut. We can safely conclude that the presence of wind is in fact causing for the door to slam shut.

Not all effects are the product of a single cause. For example, consider an allergy. If a person is allergic to a particular substance, he/she will react to the exposure to the substance but not to the exposure to the substances to which the person is not allergic. However other people who are not allergic to the particular substance will not react to it. The person's allergy to a substance is then attributable to the conjunction of the particular person and the specific substance.

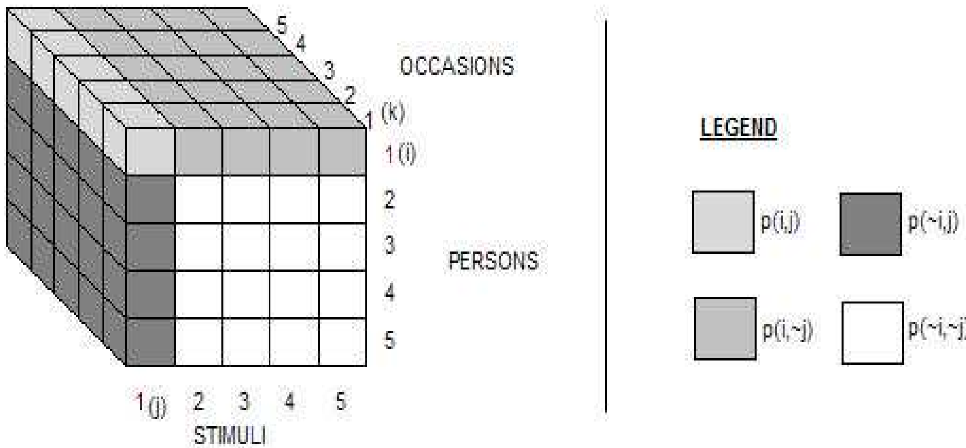

Figure 4.2: Relevant information for specifying a main-effect contrast and an interaction contrast according to Cheng and Novick's Probabilistic Contrast Model

4.2.4 Interaction Contrast

Whereas a main effect contrast requires a difference to be noticeably large between the proportion of cases in which the effect is observed in the presence of the presumably causal factor and the absence of that factor, an interaction contrast requires a notably large difference between such differences for levels of an orthogonal factor. For example, we may have the interaction between the simultaneous presence of positively charged clouds and negatively charged clouds as the cause of thunder. It consists of a difference of differences, or a second order difference. A two-way interaction contrast involving potential causal factors i and j is defined as:

$$\Delta p(i,j) \;=\; (p(i,j) - p(\neg i,j)) - (p(i,\neg j) - p(\neg i \neg j)), \tag{4.2}$$

where p denotes the proportion of cases in which the effect i occurs when a potential contributing factor j is either present or absent.

As an example, consider again the cube in Figure 4.2. Let Person i be a particular value along the person dimension and Stimulus j be a particular value along the stimulus dimension. In Figure 4.2, we illustrate this situation for $i = 1$ and $j = 1$. Now, suppose we want to determine if the interaction between Person i and Stimulus j can be considered a potential cause for a particular event involving Person i, Stimulus j, and Occasion k. To make such a conclusion regarding a two-way interaction with the probabilistic contrast model, four proportions need to be considered.

For events in which Person i is present:

- let $p(i,j)$ be the proportion of occasions on which the effect occurs when Stimulus j is also present;

- let $p(i, \neg j)$ be the associated proportion in the absence of Stimulus j.

For events in which Person i is absent:

- let $p(\neg i, j)$ be the proportion of occasions on which the effect occurs when Stimulus j is also present;
- let $p(\neg i, \neg j)$ be the associated proportion in the absence of Stimulus j.

If the difference between $p(i, j)$ and $p(i, \neg j)$ is significantly greater than the difference between $p(\neg i, j)$ and $p(\neg i, \neg j)$, or symmetrically if the difference between $p(i, j)$ and $p(\neg i, j)$ is significantly greater than the difference between $p(i, \neg j)$ and $p(\neg i, \neg j)$, then we conclude that the conjunction of Person i and Stimulus j is a cause of the target effect.

Note that, in general, interaction contrasts involving n factors are defined as n^{th} order differences, where n is any positive integer. That idea is true in theory, however, Cheng and Novick [19] do mention the computation will presumably be intractable but they have not studied n-way interactions in any depth.

4.2.5 Facilitatory (Generative) versus Inhibitory (Preventive) Causes

Cheng and Novick's initial probabilistic contrast model has two types of causes: generative, and preventive. A generative cause is a factor that increases the likelihood of the effect under investigation. In an analogous fashion, we can define a preventive cause as a factor whose presence decreases the likelihood of occurrence of the effect for which we seek a cause. For example, the presence of bug spray on someone's body reduces the likelihood that the person wearing the bug spray will get bit by mosquitoes. Therefore, when "mosquito bites" is the effect, the "presence of bug spray" is a preventive (inhibitory) cause. Hilton and Slugoski [51] and Kelley [75] also discuss generative versus preventive cause.

An interesting aspect of noting whether a cause is generative or preventive is the preservation of directionality. In the standard ANOVA model, all differences are squared and therefore all information about directionality is lost. In the probabilistic contrast model, main-effect contrasts see a positive difference for generative causes and a negative difference for preventive causes. For interactions, the difference of differences is positive for generative combinations of factors and negative for preventive combinations.

4.2.6 Note on Alternative Causes

In the probabilistic contrast model, multiple alternative causes are allowed and are distinguished from conjunctive causes (interactions). Recall that conjunctive causes involve multiple necessary factors for the difference in presence versus absence of factors to be significant enough to be deemed causal. Any factor whose presence increases the probability of occurrence of the effect is a possible cause of the effect in question. If many factors alter the probability, then each one of those is considered to be a cause of the effect, and none is necessary. That is, the effect will occur if any (not necessarily all) of the multiple alternative causes are present. The multiple alternative causes can be a combination of main-effect contrasts and/or conjunctive causes (interactions).

4.2.7 Focal Sets and Computation of Contrasts

In order to compute contrasts between compatible events only, Cheng and Novick introduced to their model the idea of *focal sets*: contextually selected sets of events over which covariation is computed. To arrive at the idea of a focal set, a distinction must be made between a *novel candidate cause*, an *enabling condition*, and a *causally irrelevant factor* in the context of the probabilistic contrast model.

Generally speaking, the *novel candidate cause* of an effect is an event that occurs, and because of that occurrence, the effect of the event happens as well. If the event we call the cause had not taken place, the effect would not have subsequently occurred. In the philosophy literature, the idea of *enabling conditions* and how they differ from actual causes has been studied extensively [105, 106, 107, 108, 109, 110, 74, 111]. An *enabling condition* is a factor that enables the cause to occur. For example, if investigators are asked to report the cause for a plane crash, they will likely attribute the crash to factors such as pilot error, wind shear, or malfunctioning of a critical component, rather than say, the gravitational pull of the earth, although without the gravitational pull of the earth, the crash would likely not have occurred. However, when a crash does not occur, the gravitational pull of the earth is still present. Finally, a *causally irrelevant* factor changes nothing in the occurrence of the effect. If we consider again the plane crash example, we can say that the colour of the airplane seats, or the number of TV screens in the airplane are causally irrelevant factors with respect to the crash.

Given the distinction between causes, enabling conditions, and causally irrelevant factors, we can address the necessity for a more rigid definition of context, when making causal attribution. The same way we would not blame the gravitational pull of the earth for a plane crash, we would not attribute the presence of oxygen to the onset of a forest fire. However, we cannot state that the variable “oxygen” is causally irrelevant the same way that the presence of rocks in the forest is causally irrelevant with respect to starting a forest fire. Ideally, we want to disregard any causally irrelevant factors and abstract our model to only relevant factors. However, if we remove the variable “oxygen” from the model “causing fire”, the model is correct as long as the context is limited to “forest fire”. However, if we change the context to “laboratory”, where precautions are taken to exclude oxygen during parts of an experiment, the presence of oxygen may in fact cause a fire. The idea of focal sets resolves this ambiguity. In order for an event or factor to be considered causally irrelevant of a particular effect, it must be unable to produce the effect in question in every context or domain, which are referred to as *focal sets*. However, in the case where an event is a cause in a particular context, or focal set, but not in others, we call that event an *enabling condition* in all other focal sets in which the event is not a cause. For example, for the effect "fire", "oxygen" is a potential cause in the context "laboratory", thus it is an enabling condition in the context "forest". On the other hand, "presence of rocks" is causally irrelevant.

4.3 Power PC

After studying the problem of incompleteness of information responsible for bias in causal induction, Cheng and Novick arrived at the popular conclusion that in general, covariation does not always imply causation.

Novick wrote about the evolution of their approach to the problem of causal induction:

> Since our 1992 paper, Patricia Cheng and I have radically revised our view of causal induction. Although our previous model, which is purely covariational, explains a wide range of findings regarding causal inference, it cannot explain why even untutored reasoners do not equate covariation with causation. An opposing approach – the power or mechanism approach – has attempted to address reasoners' intuitive understanding of this fundamental inequality, but it has been unable to specify the process that transforms information from the available noncausal input to a causal judgment. Cheng [21] formulated a revised version of our probabilistic contrast model – the power PC theory – that demonstrates that an integration of the covariation and power approaches can overcome the problems confronting each approach.

Work in psychology on the issue of causal induction has been dominated by the approach of covariation, which we discussed in Section 3.3.2, and also the *power approach*, which we discuss further in this section. Traditionally, the two approaches have been regarded as opposing views, or competitors for true explanation of causality. However, after working with their probabilistic contrast model, Cheng and Novick came to the conclusion that neither covariation alone nor causal power alone can explain the inferences humans make about causal relations. In the remainder of this section, we describe the clues that lead to believe in the necessity for a collaboration between the two, and we briefly describe Cheng's [21] improved probabilistic contrast model, which she refers to as *Power PC*. Power PC uses both accounts of causality, namely covariation and causal power. In the next section, we give an overview of the roots in philosophy of both views, as well as the problems encountered when considered in isolation.

4.3.1 Roots of Covariation and Causal Power Studied in Isolation

When speculating about causality, its existence, and its true nature, philosophy studies the question: "How does a reasoner come to know that one thing causes another?" The work that has been done in the study of behaviour on this question of causal induction has been dominated by two views that traditionally have been seen as competing for the real truth. The first view, namely covariation, traces its roots to the philosopher David Hume [55], while the second view, namely causal power, stems from the philosophy of Kant [56]. We discuss the general beliefs entailed by both approaches.

Roots of Covariation - Hume

Hume's covariational approach of human causation is motivated by the belief that the ultimate source of all information available to the human being lies in the sensory inputs, which is believed to contain *no* explicit causal relations. From that assumption, all acquired causal relations have to somehow be computed from sensory input. From our sensory input, we can infer information such as the presence versus absence of candidate causes and of effects. We also can assess temporal and spatial information about events. The belief in having access solely to these "observable" facts about the world to use as input to our causal inferences leaves covariation as the only logical choice for any attempts at making inferences from which to assess causation. We simply decide on the strength of the causal relationship based on the degree of covariation between the potential cause and the target effect.

Problems with Covariation Approach Used in Isolation

Covariation models of causality share a common problem: covariation does not always imply causation. Although event B may always follow event A under all circumstances, it is impossible to make an absolute conclusion that A causes B. For example, sunrise may occur everyday after a rooster on a farm crows, but sunrise doesn't happen any other time during the day when the rooster doesn't crow, yet it would be incorrect to state the conclusion that the rooster's crowing causes the sun to rise. The question left unanswered here is: What takes one from covariation to causation?

Roots of Causal Power - Kant

The causal power approach, which traces its roots to Kant [56], suggests that there exists an amount of a priori knowledge that "serves as a framework for interpreting input to the causal induction process" [21]. Kant proposed that humans have some innate knowledge about the existence of causality. We automatically believe that all events are caused: "All alterations take place in conformity with the law of the connection of causes and effect" [56] (p. 218). This suggests that humans speculate on a strong assertion that there is something abstract between the cause and the effect, besides the simple temporal factor, that associates the cause and the effect, and without which the effect would not be produced. This abstract concept of "something more" has been referred to in the literature as generative source, causal mechanism, causal propensity, and causal power [112, 113, 114, 115, 116, 117, 118].

Cheng [21] uses the term *causal power* to cover all the above variants and defines it as: "the intuitive notion that one thing causes another by virtue of the power or energy that it exerts over the other." For example, when the sun warms one's back, one may explain the rise in temperature of their back by virtue of the sun emitting energy, which in turn travels to reach the skin.

In summary, the causal power view suggests that effects do not simply follow their causes, instead, effects are *generated* by their causes. The observable statistical characteristics of covariational (but not causal) sequences are very similar to the ones of actual causal sequences, except they are "missing the critical connection provided by the understanding of a causal power" [21]. The sunrise following crowing is an example of this type of sequence.

Problems with Causal Approach Used in Isolation

The main problem with causal power alone is that it is not computational. That is, it does not define an explicit mapping between the ultimate input and the output in the process of causal induction. Also, the approach suffers from a problem of circularity. Specific causal powers are, by definition causal. Finally, it leaves us with the question: Unless knowledge of these powers is innate, how do reasoners come to know them?

4.3.2 Combination of Covariation and Causal Power

Cheng argues that both views, covariation and causal power, contain an element of truth in explaining human causal inference. She claims that covariation is a component of the process of causal induction, and reasoners *do* have an a priori framework for interpreting input to that process. Based on this belief, Cheng extends the original purely covariational probabilistic contrast model [19] to her improved Power PC model, in which she assumes that the issue of causal inference can be divided into two parts: how an acquired causal relation is first induced, and how prior domain-specific causal knowledge (whether innate or learned) influences subsequent causal knowledge. Many experiments support this belief that causal induction and the influence of domain-specific prior causal knowledge are separable processes [119, 19, 120].

Cheng's Power PC theory is motivated not only by the problems discussed above, but also by phenomena of natural causal induction that are inexplicable by any known psychological approach. Those phenomena address situations in which we cannot assume non-causality based on given evidence, although we cannot conclude the existence of a causal relation either [21].

4.3.3 Power PC Model

Cheng's Power PC model assumes that "there are such things in the world as causes that have the power to produce an effect and causes that have the power to prevent an effect and that only such things influence the occurrence of an effect (c.f. [114, 56])" [21]. Although Cheng's Power PC is intended as a purely psychological account of human causation, its computational flavour seems to adapt to the realm of AI techniques for Decision Making in Uncertainty. Cheng's Power PC model has later been explained by Clark Glymour [121], who claims that Cheng's models of his models of causation "turn out to be Bayes nets under a particular parameterization, which means that we can use what is known about search and estimation for Bayes nets to extend Cheng's theoretical results".

Keeping in mind that Cheng's Power PC model considers not only the covariational view of causality, but also the causal power view, which implies we are working with unobservable factors, and therefore not directly computable, the goal is to use the information we do have to isolate the value that represents the power of a certain cause C to produce a target effect E.

Cheng's Power PC model considers only causal factors with binary values: the factor can either be *present* or *absent*. In addition, only factors that are *present* can play a causal role in the relationship. A second categorization deals with the type of causal relationship. The model supports two types of causal relationships, namely a *generative* causal factor (increases the probability of the effect) and a *preventive* causal factor (decreases the probability of the effect). A third categorization distinguishes between *simple* and *interactive* (or compound) causal powers.

If two or more simple causal powers are present and have the power to produce the same type of effect, they will produce an instance of that effect independently of each other. One simple causal power of producing an effect doesn't change the probability of another simple causal power to produce the same type of effect. For example, if factors A and B both have simple, non-interactive, generative causal powers to produce an effect E, then when A and B are present, A may cause E, B may cause E, or they may both independently cause E. The probability that A, if A is present, causes E is independent of the probability that B, if present, causes E. Note that the above is *not* the probability of E given A and E given B, $p(E|A)$, $p(E|B)$.

The *interactive* causal power of A and B to produce E can be the result of A acting alone, B acting alone, A and B acting separately, or finally, A and B acting conjointly to produce (or prevent in the case of preventive causes) the effect E. Finally, the Power PC model operates on the *joint frequency of candidate causes of effect E*. Given the joint frequency data as well as the target effect E, taking into account that *unobserved* causes of E may also be acting, the model looks to answer the following question: Given the above described data, how do people judge the efficacy or causal power of any particular cause?

According to Cheng's Power PC, there is no direct link between a cause and an effect without the presence of a causal power between the cause and the effect. Humans do not infer that some event causes another event to occur unless they have some knowledge or intuition of a specific generative source (causal power) linking the cause to the effect. People have an intuitive notion that one thing causes another by virtue of the power or energy that it exerts over the other. It is this additional component of the human causal inference process that leads to the question: How are causal relations constructed from input available to one's information-processing system and distinguished from noncausal ones, including noncausal covariations?

Cheng's results about causal powers are interpreted from the point of view of a reasoner who looks to infer the magnitude of the unobservable causal power from observable events based on his/her theoretical explanation. Given that, the power of a candidate cause X to produce an effect E is the theoretical entity that can only be estimated. We denote the causal power of X to produce E as q_{xe}. This parameter, q_{xe} represents the proposition that X causes E given that X occurs. It may take on two possible values, 1, and 0. The parameter q_{xe} is an indicator function, and takes on value 1 in the situation where, if the causal factor occurs, it acts to bring about E, while 0 represents that the causal factor, even if it occurs, does not bring about E. It is important to understand the distinction between the conditional probability of E given X and the power of X to cause E. The former can be described as the probability of E occurring in the presence of X, $p(E|X)$, which can be estimated directly from observable events, while the latter it is the probability with which X produces E when X is present.

4.3.4 Mathematical Derivation of Causal Power

In this section, we give an overview of the mathematical formulation of the problem, which allows for the unobservable causal power to be computed indirectly using observed data.

Obtaining Power for Generative Causes

In Cheng's Power PC model, the only way to obtain the causal power for a candidate cause is indirectly, through other observable data. In order to compute the causal power of a generative candidate cause, Cheng makes some assumptions about the available information. In this section, we present the derivation of generative and preventive causes of Cheng's model [21] as formalized by Glymour [121]. First, suppose that people know and believe that all unobserved causes of a target effect E are generative, and that at least one generative candidate cause of E is observed. Also, we assume for simplicity of the demonstration, without loss of generality, that there is one observed generative causal factor C and one unobserved causal factor U. The target effect E occurs if and only if C occurs *and* C can generate the effect E, or U occurs *and* U can generate the target effect E. We also let the parameter q_{ce} represent the proposition that C causes E given that C occurs, and similarly for q_{ue}.

The parameters q can take on two different values or states: 1 if the causal factor's occurrence acts to bring about the target effect E and 0 if even the presence of the causal factor doesn't act to bring about the target effect E. Finally, C, U, and E are all binary variables as well. They take on value 1 if they occur and 0 otherwise. With that terminology in mind, a target effect E is assigned a value 1 if and only if $q_{ce}C = 1$ or $q_{ue}U = 1$, since the effect only occurs if it has the power to do so (q_{ce} or q_{ue}) and the corresponding event occurs (C or U). Therefore, the probability of an effect E is as follows:

$$p(E = 1) \;=\; p(q_{ce}C = 1 \;\vee\; q_{ue}U = 1) \tag{4.3}$$

For brevity, we express values of 1 as the variable alone, and values of 0 as the variable preceded by the negation $\neg$, so that the expression in Equation 4.3 becomes Equation 4.4.

$$p(E) \;=\; p(q_{ce}C \;\vee\; q_{ue}U) \tag{4.4}$$

For any propositions A and B, the probability of A or B is $p(A) + p(B) - p(AB)$, hence:

$$p(E) \;=\; p(q_{ce}C) + p(q_{ue}U) - p(q_{ce}q_{ue}CU) \tag{4.5}$$

Assuming independence between q_{ce}, q_{ue}, C, and U, Equation 4.5 can be rewritten as:

$$p(E) \;=\; p(q_{ce}) \cdot p(C) + p(q_{ue}) \cdot p(U) - p(q_{ce}) \cdot p(q_{ue}) \cdot p(CU). \tag{4.6}$$

From Equation 4.6, the probability of target effect E given that candidate cause C occurs and U does not occur, $p(E|C, \neg U)$, can be reduced to:

$$p(E|C\neg U) \;=\; p(q_{ce}), \tag{4.7}$$

which justifies describing $p(q_{ce})$ as the causal power of C to produce E. However, this still doesn't answer the initial question; it doesn't explain how we can know, or even estimate the causal power of C to produce E. We still need a way to estimate the probability value of $p(q_{ce})$, the causal power of candidate cause C to generate the target effect E. Keeping in mind the assumption that C and U are independent, we can separate Equation 4.6 into Equations 4.8 and 4.9 as follows:

$$p(E|C) \;=\; p(q_{ce}) + p(q_{ue}) \cdot p(U) - p(q_{ce}) \cdot p(q_{ue}) \cdot p(U), \tag{4.8}$$

$$p(E|\neg C) \;=\; p(q_{ue}) \cdot p(U). \tag{4.9}$$

Computing the difference, or the probabilistic contrast between Equation 4.8 and 4.9,

$$\Delta P_{CE} \;=\; p(E|C) - p(E|\neg C), \text{we have} \tag{4.10}$$

$$\begin{aligned}
\Delta P_{CE} &= p(q_{ce}) + p(q_{ue}) \cdot p(U) - p(q_{ce}) \cdot p(q_{ue}) \cdot p(U) - [p(q_{ue}) \cdot p(U)] \\
&= p(q_{ce}) + p(q_{ue}) \cdot p(U) \cdot [1 - p(q_{ce}) - 1] \\
&= p(q_{ce}) - p(q_{ue}) \cdot p(U) \cdot p(q_{ce}) \\
&= p(q_{ce}) \cdot [1 - p(q_{ue}) \cdot p(U)]. \qquad (4.11)
\end{aligned}$$

Hence, from Equation 4.11, we obtain:

$$p(q_{ce}) \;=\; \frac{\Delta P_{CE}}{1 - p(q_{ue}) \cdot p(U)}. \tag{4.12}$$

By substitution from Equation 4.9, we obtain:

$$p(q_{ce}) \;=\; \frac{\Delta P_{CE}}{1 - p(E|\neg C)} \tag{4.13}$$

From Equation 4.13, we can conclude that the causal power of candidate cause C to generate the target effect E can be estimated indirectly from ΔP_{CE} along with the conditional probability of E given $\neg C$. This result is computed entirely from observations of C and E, and some simple assumptions about the nature of the alternate causal factor U. In other words, although powers, or unobservables are not directly observable on their own, their values can be inferred from the power PC model. Similar assumptions yield the same results no matter the number of unobserved causes, as long as they are all generative and independent of C. Also, a similar derivation is possible if we have more observed causal factors, (i.e. D) independent of C, and we condition on the absence of causal factor D.

Cheng [21] refers to this derivation as the transformation of metaphysics into testable mathematics, and states that three additional predictions can be made when the context is appropriate:

- There should be pairs of cases where people judge the causal powers to be different but the ΔPs to be identical;

- When an effect *always* occurs in the absence of the particular causal factor, people should question the power of the factor to produce the effect rather than to decide the factor has no influence. In other words, we should conclude that the data is inconclusive rather than that the model is bad;

- When the effect never occurs when the particular causal factor is absent, people should judge how efficient the factor is in predicting the effect by the ΔP, as in the purely covariational probabilistic contrast model.

Obtaining Power for Preventive Causes

Obtaining power for preventive causes can be done in a similar manner as the one used for computing power for generative causes. We suppose all unobserved causes U of target effect E are generative, and there is one observed candidate preventing cause F of E. The effect E will occur if U occurs, and U acts to bring about E, and F does not prevent E from occurring, $E = q_{ue}U \cdot (1 - q_{fe}F)$. We have:

$$p(E) = p(q_{ue}U) \cdot (1 - q_{fe}F), \tag{4.14}$$

and we wish to obtain the value of $p(q_{fe})$, the power of candidate preventive cause F to prevent the occurrence of target effect E, or the probability with which F will prevent E from occurring. By separating 4.14 into conditionals, we obtain:

$$p(E|\neg F) = p(q_{ue}U) \tag{4.15}$$

$$\begin{aligned} p(E|F) &= p(q_{ue}U) \cdot p(\neg q_{fe}) \\ &= p(E|\neg F) \cdot (1 - p(q_{fe})) \\ &= p(E|\neg F) - [p(E|\neg F) \cdot p(q_{fe})]. \end{aligned} \tag{4.16}$$

The value of $\Delta P_{FE} = p(E|F) - p(E|\neg F)$ can then be expressed as:

$$\Delta P_{FE} = -[p(E|\neg F) \cdot p(q_{fe})], \tag{4.17}$$

which yields the causal power of F to prevent the occurrence of E:

$$p(q_{fe}) = \frac{-\Delta P_{FE}}{p(E|\neg F)}. \tag{4.18}$$

From Equation 4.18, once again, the preventive causal power can be estimated directly, as it was for generative candidate causes. From the equation, we can predict that in an appropriate context, if an effect never occurs when the potential preventive cause is present, people cannot make conclusions about the preventive causal power, because it is undefined (division by zero).

Cheng conducted an experiment confirming that prediction. If one wants to test a new antibiotic and therefore applies it to a culture but not to a control culture, and if all cells in both cultures dies, it is not logical to conclude that the antibiotic has no effect. Rather, one should conclude that the experiment is not good because there is most likely some other factor independently killing the cultures.

Generative Interactive Causes

In the previous two sections, we noticed that although causal powers themselves are not directly observable in nature, which is the main disadvantage of this view of causality, Cheng's Power PC model can deduce the causal power from data about observable potential causes (generative and preventive) and target effects. Cheng's Power PC model is also capable to compute, in a similar way, the causal power of generative interactions. When two factors A and B generatively interact, the explicit mathematical model is:

$$E = q_{ue}U \oplus q_{ae}A \oplus q_{be}B \oplus q_{ab}AB, \tag{4.19}$$

where $\oplus$ represents boolean addition and q_{ab} represents the proposition that if both A and B occur, they interact to cause E. The new problem we face here is that we must somehow isolate the value of the causal power of the interaction, that is $p(q_{ab})$ in addition to the causal power of the individual causes to produce the effect on their own, as before. If we condition on the absence of B the interaction term vanishes and the equation is reduced to $E = q_{ue}U \oplus q_{ae}A$, and similarly for conditioning on the absence of A. Therefore, the simple causal powers of A and B to produce E can be estimated as before, without interaction. We can also condition on the absence of both A and B to obtain $E = q_{ue}U$, which allows us to compute the probability that E is produced by unobserved causes, $p(q_{ue}U)$. From the combination of the information here and the techniques presented in the previous two subsections, it is possible to solve for $p(q_{ab})$ from the probability of E when A and B are both present. The derivations are presented in [21] and [121].

In this chapter, we discussed Cheng and Novick's covariational probabilistic contrast model and its implications. We then discussed the justification for a model that accounts for causal power in addition to covariation. Based on that justification, we described Cheng's Power PC model.

One important prerequisite for either version of the model is contextual consideration with focal sets. Without a consideration of context, novel candidate causes, enabling conditions, and causally irrelvevant factors, may be erroneously identified. In Cheng and Novick's models, such focal sets are identified by a human expert before contrasts are computed.

In Chapter 5, Section 5.4 of this thesis, we propose a method to identify focal sets from probability distributions by means of CIs, CSIs, and model decomposition. We extract causally irrelveant factors from distributions with the notion of CI. Then, we distinguish between novel candidate causes and enabling conditions with CSI.

Chapter 5

Context as a Tool for the Refinement of Causal Models

Since CSI can be discovered inside distributions, it can provide more accurate inference since the contextual information would be ignored otherwise. Also, because CSI discovery produces a lossless decomposition of a CPD, then intuition supports that there must be a particular semantic meaning to the relationships where CSIs hold as opposed to where they don't. Our investigation deals with the type of useful relationships that can be uncovered and used to increase our knowledge and understanding of a particular situation under study. The hypothesis motivating this idea is that if a different set of independencies holds within the same distribution, but for different subsets, the subsets in question must be important in some logical way, and may have to be treated differently. The research results presented in this chapter demonstrate how such consideration of context can, in a variety of situations, provide a new way to separate an initial seemingly correct but possibly erroneous causal model into two or more new models that take into account independencies too specific for the initial model. This decomposition of the causal model allows for a more accurate representation of the situation being modeled.

We demonstrate how the notions of context unify existing ideas about human reasoning in both the cognitive science and AI literature. We also show how the algorithmic aspects suggest that context may be what helps human reasoners be efficient in the realm of vast knowledge bases. We provide a treatment of context in the setting of BNs, that gives a useful account of AI and attribution problems by using context to refine cognitive causal models. Finally, we provide a method for the identification of Cheng's focal sets from probabilistic distributions, without the assistance of a human expert.

Our main investigation tool is the first extension to CI, namely CSI (see Chapter 2). We take a closer look at its useful semantics, and the many problems it can help solve. Situations to apply CSI have been overlooked in the AI literature, as the goal has predominently been to accelerate the process of inference and to obtain more compact representations. The discovery of a CSI has a deep semantic meaning, which is highlighted in the present research.

5.1 Importance of Data Preprocessing to Build a Mental Map

When manipulating information, Artificial Intelligence is mainly preoccupied with the algorithmic process that turns the input into output. Of importance are mathematical soundness, accuracy, speed, compactness, memory requirements, etc. When modeling behaviours for which a correct algorithm is not yet known, if the tentative algorithm for predicting behaviour produces seemingly erroneous results, the attempt to rectify the problem is likely to be correcting the algorithm. However, a careful consideration of the input provided to the algorithm could change that assumption. Perhaps the algorithm is functioning adequately, whereas the input data may not be as expected.

It is clear that if the algorithm is not provided the correct data as input, it is impossible to obtain correct output. Thus, on the input data side of the question, the errors that lead to incorrect output are measurement errors. There are two scenarios where data is unmeasured and therefore incomplete. One scenario is when the relevant information is simply not in the model. We term this scenario *unmeasured-out*. Alternately, it could be hidden inside a variable, typically by means of an independence that holds in a particular context. We call this second scenario *unmeasured-in*. We show how, in certain situations, this problem can be alleviated. For the remainder of this chapter, we address the problem of unmeasured-in, where independencies hide in the data and make potentially erroneous models appear to be correct.

5.2 Problem Solving and Seemingly Paradoxical Scenarios

In this section, we show how a formal consideration of context can help correct erroneous assumptions used in problem solving. When relevant independencies hide inside variables, erroneous inference is almost inevitable. In the extreme case, that type of error may lead to seemingly paradoxical scenarios. To emphasize this problem and show how CSI may help solve it, we discuss a well-known paradox, namely *Simpson's Paradox*, and describe how it can be resolved by means of CSI [36].

5.2.1 Simpson's Paradox

Simpson [122] makes a point about a particularity of a subset of combinations of fractions that makes intuitively implausible relationships seem mathematically correct. Simpson's paradox occurs when arithmetic inequalities are reversed when individual proportions are aggregated. The result is called *Simpson's reversal of inequalities*. A generalization of the type of expression that results in such reversal is illustrated in Table 5.1. Note the use of the dashline in Table 5.1. Simpson's reversal of inequalities is not a rule; it does not yield the statements mutually inconsistent. Rather, it is

Table 5.1: Simpson's Reversal of Inequalities

(i)	a_1 / b_1	$<$	a_2 / b_2
(ii)	c_1 / d_1	$<$	c_2 / d_2
(iii)	$(a_1 + c_1)/(b_1 + d_1)$	$>$	$(a_2 + c_2)/(b_2 + d_2)$

an instance that, when true, creates an instance of Simpson's paradox. More formally, based on Table 5.1, an instance of Simpson's paradox occurs when $\exists\ a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2 \in \mathbb{Z}\ s.t.\ (i) \wedge (ii) \wedge (iii)$.

Cohen and Nagel [123] introduce a classic example of Simpson's paradox. They gathered data about death rates from tuberculosis in Richmond, Virginia and New York, New York and found the following propositions held true:

> For African Americans, the death rate was *lower* in Richmond than in New York. For Caucasians, the death rate was also *lower* in Richmond than in New York. However, for the total combined population of both African Americans and Caucasians, the death rate was *higher* in Richmond than in New York.

Scrutiny of the data reveals that Caucasians are naturally less likely to get tuberculosis, whether they live in Richmond or in New York. At the time of the survey, there were more Caucasians than African Americans living in New York, therefore a higher proportion of the New York population was less at risk. The reverse held true for Richmond, which caused the seemingly paradoxical scenario. The answer to this perplexing scenario is simply that since there were more Caucasians than African Americans living in New York at the time of the survey, and Caucasians are less likely to get tuberculosis, their dominance in New York skews the data significantly. Only with contextual consideration of ethnic origins separately can we derive the more acurate and realistic scenario.

In the next subsection, we present a causal scenario where context is not considered, to show how faulty conclusions and counterintuitive associations can be obtained from mathematically sound equations. We then show how Simpson's paradox can be understood in terms of independencies hidden in specific contexts in the data.

5.2.2 A Seemingly Correct Causal Model

The causal model in Figure 5.1 describes a possible causal relationship between the variables *(M)elanoma* , *(S)unscreen*, and *Skin-(T)ype*. According to the DAG, wearing sunscreen has a direct causal influence on the incidence of melanoma, and skin-type has a direct causal influence on wearing sunscreen, and on the incidence of melanoma. The corresponding JPD factorization for variables M, S, and T is the following:

$$p(M, S, T) \; = \; p(M|S, T) \cdot p(S|T) \cdot p(T). \tag{5.1}$$

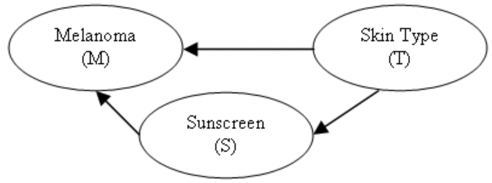

Figure 5.1: Causal model describing the causal relationship between use of sunscreen, skin-type, and incidence of melanoma.

Although the causal model in Figure 5.1 seems reasonable and intuitive, a recent study showed that sunscreen users might be at risk of melanoma [124]. Although the available data seems to yield that result, it remains counterintuitive. We show how such possibly erroneous conclusions could faultily penetrate into the system. Although the notion of causation is frequently associated with concepts of necessity and functional dependence, "causal expressions often tolerate exceptions, primarily due to missing variables and coarse descriptions" [61].

Note that in this section, we show a contrived example to emphasize the extreme case where the hidden independencies actually result in an instance of Simpsons paradox and makes it appear as though wearing sunscreen actually *causes* melanoma. However, even if we didn't take it to the extreme, we would remain in a situation where it may simply look as though sunscreen has *no* effect on the disease melanoma, which is not as strong as making such a faulty conclusion as saying sunscreen *causes* melanoma, but nonetheless still incorrect, as evidence point to many benefits of wearing sunscreen. With contextual consideration we can we obtain a more realistic scenario.

Assume a situation where the department of health is considering a promotion of the use of sunscreen as a measure to prevent being exposed to the disease melanoma. The promotion encourages both dark-skinned people and light-skinned people to wear sunscreen. However, statistics gathered from a typical sample of the population, shows some puzzling and questionable results.

For the remainder of this example, we assume the domains of variables *(M)elanoma*, *Skin-(T)ype*, and use of *(S)unscreen* to be binary. The variables may take on the following sets of values respectively: $\{(y)\text{es}, (n)\text{o}\}$, $\{(l)\text{ight}, (d)\text{ark}\}$, and $\{(y)\text{es}, (n)\text{o}\}$. The numbers here are contrived to illustrate the example.

In the sample set, 50 people with dark skin wore sunscreen and only 10 got melanoma. On the other hand, out of 80 dark-skinned people *not* wearing sunscreen, 20 got melanoma. Of all dark-skinned people in the sample set, 20% of those who wore sunscreen got melanoma, while 25% of those who didn't wear sunscreen were victims of the disease.

In the light-skinned portion of the sample set, out of 80 who wore sunscreen, 60 got melanoma, while 40 out of 50 people who didn't wear sunscreen got sick. In total, 75% of light-skinned people who wore sunscreen got melanoma, while 80% of those who didn't protect their skin were affected.

Yet, altogether 130 people wore sunscreen and 130 people didn't wear sunscreen. Of the 130

Table 5.2: Simpson's reversal of inequalities in the *Sunscreen*, *Skin-Type*, and *Melanoma* problem, where proportions are a function of the occurence of *Melanoma*.

	Sunscreen		**No Sunscreen**
Dark Skin	10/50 (20%)	$<$	20/80 (25%)
Light Skin	60/80 (75%)	$<$	40/50 (80%)
All Subjects	70/130 ($\approx$ 53.8%)	$>$	60/130 ($\approx$ 46.2%)

people who did in fact wear sunscreen, 70 got melanoma and of the 130 people who didn't wear sunscreen, 60 people got the disease. The percentage of people who did wear sunscreen and still got melanoma is greater than the percentage of people who didn't wear sunscreen and got melanoma. Table 5.2 shows Simpson's reversal of inequalities (see Table 5.1) in the above example.

This problem is perplexing. How can it be that both dark skin and light skin favor the use of sunscreen and yet overall, *not* wearing sunscreen is better than wearing sunscreen? The sample sizes are equal for both groups, sunscreen (130) and no sunscreen (130), and also for light skin (130) and dark skin (130). In addition, the problem doesn't arise due to small sample size, as the problem remains for any multiple of the numbers. In fact, as we increase the sample size, we only solidify confidence in the reversal of inequalities. For a factor of 1 million for example, we can add or remove a fair number from each of the millions without altering Simpson's reversal of inequalities.

The answer to this bewildering example is simply that it is less likely for the dark-skinned person to get melanoma, independent of their use of sunscreen. In the example, of the people wearing no sunscreen and getting melanoma, more have dark skin than light skin, and the reverse is true for those who wear sunscreen. Of those with dark skin, only 30 out of 130 got melanoma, whereas 100 out of 130 light-skinned people got melanoma, where there were more people wearing sunscreen.

More formally, in the context where the skin-type is dark, wearing sunscreen and getting melanoma are independent. We can formalize Simpson's paradox using *context-specific independence* (CSI) [3].

5.2.3 Correcting the Model

As discussed in Chapter 2, there are situations where CI cannot capture independencies that hold only in certain contexts. Although those independencies are not visible when all contexts of the variables are considered, the presence of independencies that are only true in certain contexts will affect the causal model, and perhaps yield causal links that either do not exist in reality, or are much stronger than what the model shows, if context was considered. Also, consideration of CSI may improve causal inference even in cases where the relationships do not result in paradoxical statements.

Based on the indirect specification of the JPD (see Section 2.3), the JPD $p(T, S, M)$ can be

Table 5.3: CPD for $p(M|T,S)$, the probability of *Melanoma* given *Skin-Type* and *Sunscreen*.

T	S	M	$p(M\|T,S)$
L	Y	Y	N_1
L	Y	N	N_2
L	N	Y	N_3
L	N	N	N_4
D	Y	Y	N_5
D	Y	N	N_6
D	N	Y	N_5
D	N	N	N_6

Table 5.4: CSI decomposition of CPD $p(M|T,S)$ in Figure 5.3.

T	S	M	$p(M\|T,S)$
L	Y	Y	N_1
L	Y	N	N_2
L	N	Y	N_3
L	N	N	N_4
D	Y	Y	N_5
D	Y	N	N_6
D	N	Y	N_5
D	N	N	N_6

(i)

T	S	M	$p(M\|T=l,S)$
L	Y	Y	N_1
L	Y	N	N_2
L	N	Y	N_3
L	N	N	N_4

T	S	M	$p(M\|T=d,S)$
D	Y	Y	N_5
D	Y	N	N_6
D	N	Y	N_5
D	N	N	N_6

(ii)

T	M	$p(M\|T=d)$
D	Y	N_5
D	N	N_6

(iii)

expressed as the product of the CPD's in the model depicted in Figure 5.1.

$$p(T,S,M) \quad = \quad p(T) \cdot p(S|T) \cdot p(M|S,T). \tag{5.2}$$

From the indirect specification of the causal model in Figure 5.1, in Equation 5.1, it is fair to state that the multiplication of CPDs $p(T), p(S|T)$, and $p(M|S,T)$ define the complete causal model in terms of the available information.

The CPD $p(M|S,T)$ can be decomposed as follows. Since variable T is binary, we can separate the CPD $p(M|S,T)$ into two contexts (l and d), without making any reductions. We obtain one partial CPD for context $T{=}l$, and one partial CPD for $T{=}d$, as in Equation 5.3. This separation is also illustrated in Figure 5.4 (ii).

$$p(M|S,T) \quad = \quad p(M|S,T=l) \odot p(M|S,T=d) \tag{5.3}$$

This decomposition is useful if a CSI doesn't hold in one partial function or the other. Since variable S is independent of M in context $T=d$, a reduction is possible in the partial CPD where $T=d$. In context $T=d$, $p(M|S,T=d) = p(M|T=d)$. The decomposition of the CPD is presented in Equation 5.4 and illustrated in Figure 5.4 (iii).

$$p(M|S,T) = p(M|S,T=l) \odot p(M|S,T=d)$$
$$= p(M|S,T=l) \odot p(M|T=d) \quad (5.4)$$

By substitution, we obtain the following final decomposition of the available causal model.

$$p(T,S,M) = p(T) \cdot p(S|T) \cdot p(M|S,T=l) \odot p(M|T=d)$$

Note that S is no longer included in the CPD for M when $T = d$. Using CSI, we established that given *Skin-Type* = *dark*, variables *Melanoma* and *Sunscreen* are conditionally independent. The associated CPD is shown in Table 5.3, and the CSI decomposition for that CPD is presented in Table 5.4.

To eliminate the problem, we can detect CSI in the input data and therefore build a set of representative causal models for relevant subsets of the data. The CPD-Tree algorithm [7], allows for decomposition of the CPDs based on CSI, where the detection is entirely performed from data.

The detection method is straightforward. Initially, we express the CPD as a tree, as in Figure 5.2 (left), which is taken from the CPD $p(M|S,T)$. To build the initial tree, the leaf nodes are the ones with the most incoming arrows in the corresponding causal model, and the root node, the least. The detection algorithm is described in Section 2.3.

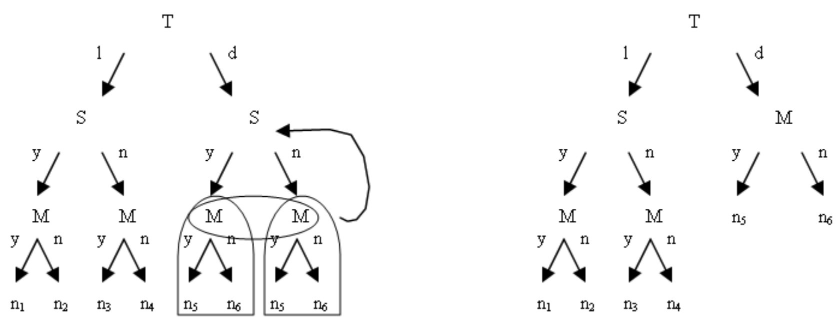

Figure 5.2: CPD-Trees for CSI detection from data.

Figure 5.2 (right) shows the tree after CSI detection. The resulting figure, where *Skin-Type* = *dark*, does not mention sunscreen. Given variable *Skin-Type* = *dark*, variables *Melanoma* and *Sunscreen* are conditionally independent. From the now known independencies, the resulting CPDs for $p(M|S,T)$ are the two CPDs in Table 5.4, and therefore the resulting causal models for the contexts *Skin-Type* = *light* and *Skin-Type* = *dark* respectively are shown in Figure 5.3.

In summary, the detection of CSI results into two causal models, each expressing different independencies based on contexts of the variables, therefore capturing the problems with the paradoxical data and repairing it with the detection method.

Figure 5.3: Resulting causal models after CSI detection with CPD-Trees.

5.3 Correspondent Inferences in Attribution Theory

In this section, we show how a consideration of contexual independencies can help discover hidden dispositions and situational factors (see Section 3.2) in causal relations [46]. Once again, we emphasize here that without consideration of CSI, the model we present would seem intuitively correct, and false conclusions would be drawn from it.

In the determination of causal attribution, we are interested in not only the true cause of behaviour, but also in how we, as human adults, assign a cause to another person's behaviour, whether the inferred cause is true or not. When seeking to understand another individual's behaviour, people generally use two types of information, namely *situational* factors, and *dispositional* causes (see Section 3.2). Situational factors explain actions in terms of a social setting or environment, while dispositions are causes based on characteristics of the person whose behaviour we seek to understand. When attributing a cause to a person, it is very important that the inference comes from dispositional factors, and not situational ones. However, the distinction between the two is often blurred in data. For example, a job applicant who fails to attend a recruitment meeting may be perceived as anti-social or uninterested (disposition), when in reality, the individual may live out of province, and will only relocate if hired (situation). A more thorough examination of the context of the situation painted by the available information may reveal hidden clues about the nature of the factors being considered (situational or dispositional). Discovery of such clues (context-specific independencies) may yield more accurate causal models to describe the situation at hand.

In this section, we address how consideration of context can help uncover hidden factors about individuals and how the discovered independencies will change and improve our believed causal model, by isolating situational factors and true dispositions, to distinguish between the causal repercussions in both cases. We show that if the contextually hidden information is considered, it can help us learn whether the attribution was based on a person's true disposition or on situational factors. In addition, we may discover that two different causal models should be used for the same scenario based on the type of attribution that was made (dispositional or situational). We discuss a method for correcting such erroneous models by finding the hidden contextual variables. For the remainder of the section, the terms factor and variable will be used interchangeably.

5.3.1 Correspondent Inferences in Attribution

As discussed in Chapter 3, situational factors explain actions in terms of a social setting or environment, while dispositions are causes based on characteristics of the person in question. Jones and Davis' Correspondent Inference theory [39] suggests that we use information about the behaviour of a person as well as effects of the particular behaviour to make a correspondent inference, in which the behaviour is either attributed to a disposition or a situation, and is based on a sole observation. This theory is interesting for hidden variable discoveries, as we have a single observation about each individual, and discover independencies between variables when we look at a group of individuals performing a similar task.

5.3.2 Causal Model where Dispositional and Situational Factors May Lead to Erroneous Conclusions

Company ABC is interested in better understanding what type of applicant is likely to be a successful employee within the company. ABC is a large corporation and receives applications from across the country. The CEO likes to interview as many qualified applicants as possible. However, although a large percentage of applicants meet all the requirements, to reduce recruitment cost the CEO would like to interview only a subset of the qualified applicants. The CEO would like to learn more about the current employees of his company to understand what type of applicant would likely be successful in interview.

The causal model in Figure 5.4 describes the causal relationship between five variables directly related to the potential success in interview of a typical applicant, including the success variable itself. For simplicity, we assume each variable is binary. The five variables are the following: *(A)pplicant's experience with dealing with the public*, {0 = no experience, 1 = experience}, *(W)eekend outings organized by company regularly to promote dynamics within personnel*, {0 = uninterested, 1 = interested}, *(P)reparation for interview*, {0 = little preparation, 1 = extensive preparation}, *(R)esearch about company done by applicant prior to interview*, {0 = no, 1 = yes}, and finally *(S)uccess in job interview*, {0 = no, 1 = yes}.

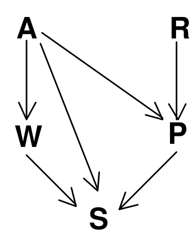

Figure 5.4: Causal model for job interview.

According to the DAG, there is a direct causal relationship between the applicant's experience with the public (A) and their interest in making their involvement in the company a part of their social life (W). There is also a direct causal influence from A to P, the time and effort spent on job interview preparation. Finally, the last causal relationship involving A is clear, namely that there is a relationship between A and a successful job interview (S). Researching the company prior to the job interview (R) is causally related to preparation for the interview (P), which in turn is directly causally related to S, a successful interview. Finally, an interest in socializing outside work hours (W) is directly related to a successful job interview (S). The JPD factorization for variables A, R, P, W, and S contains the following CPDs:

$$p(A, R, P, W, S) \; = \; p(A) \cdot p(R) \cdot p(P|A, R) \cdot p(W|A) \cdot p(S|A, W, P). \tag{5.5}$$

Although the causal model in Figure 5.4 seems reasonable and intuitive, we will see later that discovery of hidden independencies paints a different picture that can lead to bad hiring decisions if left unattended.

5.3.3 Discovery of Hidden Independencies

Since BNs operate on the general notion of CIs between variables, it is difficult to consider hidden independencies in the data or even to be aware of their presence. In his attempts to understand applicants and their potential fit within the company, while not interviewing all qualified applicants, the CEO of ABC gathers factors about the applicants that he feels are relevant indicators of success. For every hiring session, he organizes an informal social recruiting session specifically for the applicants, and although not mandatory, he expects most candidates to attend. Since this session is an indicator of motivation and interest, the CEO compiles the applications of the applicants who did not attend the session in past hiring rounds, to look for indicators of a lower applicant success rate, which is exactly what one would expect. Based on the arrows in the causal model in Figure 5.4, the variables having a direct relationship with successful interview S are A, P, and W. The associated CPD for $p(S|A, W, P)$ is presented in Table 5.5.

Based on the information in the distribution, we see that some applicants who did not attend the session were very successful in interview while others were not. For example, when candidates had previous experience with the public, they were more likely to be hired. Also, when they prepared extensively for the interview, they were more likely to be hired as well, although they did not attend the session. More examples can be derived from Table 5.5. There is no clear indication that not attending the recruitment session had a direct impact on overall success. If that were the case, all probability values in the distribution would be quite low since none, or few of the applicants from this group would have had successful interviews. Below, we see how a discovery method for hidden variables reveals strong influences hidden in this seemingly inconclusive CPD, and revealing situational factors about the individuals that are not to be attributed to true dispositions about

Table 5.5: The CPD $p(S|A, W, P)$.

A	W	P	S	$p(S\|A, W, P)$
0	0	0	0	0.80
0	0	0	1	0.20
0	0	1	0	0.10
0	0	1	1	0.90
0	1	0	0	0.80
0	1	0	1	0.20
0	1	1	0	0.10
0	1	1	1	0.90
1	0	0	0	0.15
1	0	0	1	0.85
1	0	1	0	0.15
1	0	1	1	0.85
1	1	0	0	0.05
1	1	0	1	0.95
1	1	1	0	0.05
1	1	1	1	0.95

the person, but rather to the situation.

5.3.4 CSI Discovery

The CEO of ABC did not consider context. In this subsection, we see that a consideration of context changes the original model in Figure 5.4. Once again, we use the CSI detection method Refine-CPD-Tree [7] discussed in Section 2.3. Using the algorithm, we verify if a tree reduction is possible. If such a reduced tree exists, the data contains a CSI, which is an indication of a hidden variable that could perhaps correct a faulty model that may otherwise appear correct.

In our example, we have a CPD that contains all available information relevant to making a decision about the potential success of an interview by a job applicant, as depicted in Table 5.5. Recall that no variables can be removed from that distribution based on CI, since the independence would have to hold for all values in the distribution. The Refine-CPD algorithm can determine if context-specific independencies reside in the data. The CPD in Table 5.5 can be represented as the CPD-tree in Figure 5.5.

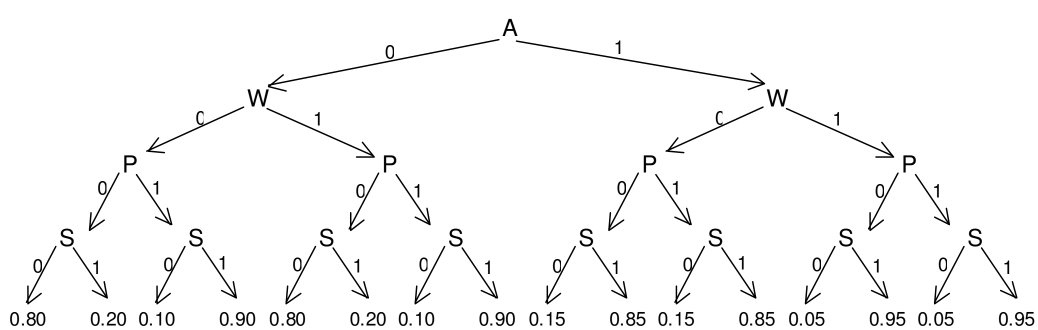

Figure 5.5: Initial CPD-tree for $p(S|A, W, P)$.

Running the Refine-CPD algorithm yields the refined CPD-tree in Figure 5.6. The variable W no longer appears on the left side of the tree, in the context $A = 0$. In addition, on the right side of the tree, in context $A = 1$, the variable P no longer appears. This suggests a hidden relationship in variable A in context $A = 0$ and in context $A = 1$.

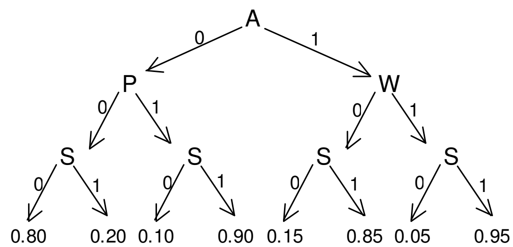

Figure 5.6: Refined CPD-tree for $p(S|A, W, P)$.

5.3.5 Uncovering Hidden Dispositions and Situational Factors

The previous subsection showed that a CSI discovery algorithm can uncover hidden relationships in a CPD when no causal independencies can be inferred by considering the entire dataset. The example showed that some contexts of A may help explain the relevance of the applicants' absence to the recruitment session. If we look again at Table 5.5, and consider $A = 0$ and $A = 1$ separately, we observe that removing W from the distribution in configurations where $A = 0$ doesn't change the likelihood of occurrence of S, whereas such a removal would be impossible in the context $A = 1$. In $A = 0$, $p(S|A = 0, P, W) = 0.80$ when $P = 0$ and $S = 0$, 0.20 when $P = 0$ and $S = 1$, 0.10 when $P = 1$ and $S = 0$, and finally, 0.90 when $P = 1$ and $S = 1$. In context $A = 1$, saying $p(S|A = 1, P, W) = 0.15$ when $P = 0$ and $S = 0$, is not completely correct since it is also true that in context $A = 1$, $p(S|A = 1, P, W) = 0.05$ when $P = 0$ and $S = 0$. This inconsistency persists because of the values of W, $p(S|A = 1, P, W) = 0.15$ with $P = 0$ and $S = 0$ only when $W = 0$. Also, $p(S|A = 1, P, W) = 0.05$ with $P = 0$ and $S = 0$ only when $W = 1$. Therefore, the value of W *does* change the probability of successful interview in context $A = 1$, so no removal is possible. We conclude that in context $A = 0$, variables S and W are independent given variable P. Such a separation is legal since no information is lost, because the union-product operator (see Section 2.4.2) can reconstruct the original CPD. From the resulting CPDs, we may now make more adequate judgments about the individuals. The CPD after refinement is presented in Table 5.6 (iii).

Isolation of contexts suggests different causal models depending upon the value of A. An examination of the semantics of the reduction reveals that in context $A = 0$ (no experience with the public), variable W plays no role in estimating the success of the candidate's interview. Recall that variable W dealt with the candidate's interest in participating in company weekend outings. Since

Table 5.6: Variables S and W are conditionally independent given P in context $A = 0$, while S and P are conditionally independent given W in context $A = 1$.

(i)

$AWPS$	$p(E \vert A,W,P)$
0000	0.80
0001	0.20
0010	0.10
0011	0.90
0100	0.80
0101	0.20
0110	0.10
0111	0.90
1000	0.15
1001	0.85
1010	0.15
1011	0.85
1100	0.05
1101	0.95
1110	0.05
1111	0.95

(ii)

$AWPS$	$p(S \vert A{=}0,W,P)$
0000	0.80
0001	0.20
0010	0.10
0011	0.90
0100	0.80
0101	0.20
0110	0.10
0111	0.90

$AWPS$	$p(S \vert A{=}1,W,P)$
1000	0.15
1001	0.85
1010	0.15
1011	0.85
1100	0.05
1101	0.95
1110	0.05
1111	0.95

(iii)

APS	$p(S \vert A{=}0,P)$
000	0.80
001	0.20
010	0.90
011	0.10

AWS	$p(S \vert A{=}1,W)$
100	0.15
101	0.85
110	0.05
111	0.95

this subset of candidates have no experience with the public and do not seem eager to participate in weekend outings, we are led to believe that their absence from the recruitment session was due to a true disposition of the person. Therefore, in context $A = 0$, perhaps a different set of variables may better explain what would cause these candidates' interviews to be successful. However, without the discovery of this CSI between S and W in context $A = 0$, we cannot make that conclusion.

On the other hand, in context $A = 1$ (experience with the public), we notice that those who didn't attend the recruiting session were influenced by the weekend outings W. Their probability of success was higher when the value of W was equal to 1. Therefore, it is important to keep W in the model for that second subset of candidates since knowing W *does* change our belief in S. However, still in context $A = 1$, after running the discovery algorithm, variable P disappears. Recall that variable P dealt with preparation for the interview. Since P doesn't affect our belief in S in context $A = 1$, we can conclude that these individuals' performance is not affected by whether they prepare for the interview or not. Given that and the fact that they are eager to participate in weekend outings, it is difficult to attribute their non-attendance to the recruiting session to a true disposition. With this new knowledge acquired from the discovery of an unmeasured-in variable, we have enough information to believe that there is something particular about candidates who didn't attend the session, but yet have experience with the public and are eager to socialize with co-workers. With this information, we can look at the applications of those particular applicants to see if our unmeasured-in discovery leads us to identify that perhaps some important information has been left out of the model (unmeasured-out), but for which we could not see the importance unless we discovered the unmeasured-in variable. In this case, we may discover that such candidates all live outside the city, and therefore could not attend the session despite their desire to socialize. This new information would also coincide with their desire for socializing with co-workers on weekends (moving to a new city for a job), and their experience with the public would be a much better indicator of their success in interview than their amount of preparation (unlike their $A = 0$ counterpart). In context $A = 1$, behaviour should clearly be attributed to the situation rather than a true disposition. From this analysis, it is clear that different causal models should be used for the two groups, as the factors that would lead to a successful interview differ greatly between the two. We now see how we can refine the causal models based on the discovered independencies.

5.3.6 Refining the Model

Since there is no longer mention of variable W in context $A = 0$, we can refine our causal model by removing the direct causal link between W and S, and similarly in context $A = 1$ for variable P. With the uncovered hidden contexts of variable A, when considering the probability of a successful job interview S, given all factors that have a direct causal link with S, the initial causal model in Figure 5.4 can be represented by two more specific causal models that account for differences between the two groups. Those refined causal models are illustrated in Figure 5.7, where the left side represents the refined model for context $A = 0$ (disposition), and the right side represents the refined model for context $A = 1$ (situation).

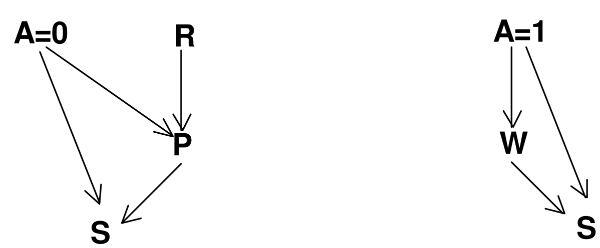

Figure 5.7: Causal Models After Discovery

Based on the more specific representations of the original causal model, it is now possible to categorize groups of individuals. Candidates in the context $A = 0$, where S and W are independent given P, are likely to be indifferent about the company's weekend activities, as they are disinclined to attend. Candidates in $A = 1$ are likely to be motivated by the idea of a social work culture, since they would be moving to a new city if they were hired. As for interview preparation, candidates in the context $A = 0$ are likely to spend more time and effort on preparation so that they feel more comfortable during the interview by contemplating as many interview scenarios as possible, due to their lack of interpersonal experience. Meanwhile, candidates in the context $A = 1$ are likely to spend less time preparing than those in context $A = 0$.

This example clearly indicates that the reason for attributing a cause to a particular individual differs greatly when we use clues about the situation surrounding the individual at the time of decision, rather than clues about a true disposition of the individual. In addition, the discovery of unmeasured-in hidden variables can help identify elements surrounding a situation (based on what variables remain in a context, and which ones disappear) to establish different causal models for dispositions and situations.

5.4 Use of Context in Discovering Focal Sets

Cheng and Novick have studied context implicitly through their *probabilistic contrast model* [19]. The main purpose of their long standing research efforts are the determination of causal judgements in human adults: *how do we decide on the causes of effects?* A portion of their work deals with the consideration of context in making attributions. They describe how context defines focal sets that may change the important causal factors at play in a setting. However, in their model, these focal sets are not discovered in probability distributions but dictated by the expert's intuition. We further this investigation by mapping focal sets and context-specific independencies (CSIs) so that CSIs can be used to discover focal sets for Cheng and Novick's probabilistic contrast model.

In this work, we discover focal sets by discovering CIs and CSIs in the probability distributions. With the discovered independencies, we can isolate causally irrelevant factors and novel candidate causes. Enabling conditions are also identified but vary slightly in definition from Cheng's enabling conditions. What we define as an enabling condition is equivalent to an *alternative cause*. To have a model that discovers a focal set following Cheng's definition exactly, we would need to further divide our discovered alternative causes into *actual enabling causes* and *genuine alternative causes*, which we do not do in this work. For the remainder of this section, we treat the term *enabling condition* as encompassing the more general concept of *alternative cause*.

If contexts are established before the attribution takes place, any potential cause must fit within one of the predetermined contexts, or it will be deemed causally irrelevant. However, if contexts could be discovered as they appear in the distributions, surprising independencies forming focal sets that the human experts did not think about ahead of time could be discovered in the information. This type of context is more interesting than predefined contexts since it may teach us something new about a situation and may provide more insight on how a particular situation should be handled from that point forward.

In the remainder of this section, we first discuss the necessity for focal sets. We then identify the factors we need to consider to adapt Cheng and Novick's models to offer a discovery method based on existing AI tools, namely context-specific independencies (CSI) and the Refine-CPD discovery algorithm. We then present the method for discovering focal sets, and end the section with an example.

5.4.1 Necessity of Focal Sets

As discussed in Chapter 4, focal sets are contextually selected sets of events over which covariation is computed. If a focal set is not determined ahead of time, Cheng's model may erroneously compute probabilistic contrasts between incompatible events. For example, when seeking a cause for the effect *Darkness*, we may attempt to compute a contrast in the variable *position of light switch*. Recall that Cheng defines a cause as a factor for which the probability of the observed effect when factor i is present, $p(i)$, is significantly greater than $p(\neg i)$, the probability of the effect when factor i is absent. In addition, the proportions computed for a main-effect contrast are approximations of the probabilistic conditional independencies of the effect given the presence or absence of event i. In our example, the binary variable *position of light switch* is considered to be a cause if the value *light switch "off"* significantly increases the likelihood of the effect *Darkness*. This main-effect contrast can be expressed as follows:

$$\Delta p(Darkness) \;=\; p(light\ switch\ \text{``}off\text{''}) - p(light\ switch\ \text{``}on\text{''}).$$

If the contrast is high, the factor *light switch "off"* is attributed as a cause for *Darkness*, which seems intuitively correct. However, if the contrast between *light switch "off"* and *light switch "on"* is nil, one must disqualify the factor *light switch "off"* as ever being a cause of the effect *Darkness*, which is clearly wrong for all instances, but very plausible if there is a short circuit, for example. If the same situation is addressed with contextual consideration, in the context "short circuit", we can safely say that whether the light switch is on or off will not impact the effect *Darkness*. However, in a different context, say "power functioning normally," the position of the light switch is a genuine potential cause of the effect *Darkness*.

Without a consideration of context, we may have to conclude that the position of the light switch is never correlated with the effect *Darkness*. With context in mind, we can easily make the distinction between the two situations without generalizing an observation to all instances. Although the position of the light switch is irrelevant in the context "short circuit," *Position of light switch* is not causally irrelevant, since it may be causal in other contexts, such as when the power is functioning normally.

Although consideration of context allows for effects to be attributed to different causes in different contexts, in Cheng and Novick's model, the contexts must be known ahead of time. This constraint leaves little room for unpredicted interesting contexts to provide insight to an otherwise perplexing situation, such as in Section 5.2, and Section 5.3 of this chapter. In the present research, we suggest a remedy for this limitation by discovering focal sets from a probabilistic distribution.

Before we elaborate on the discovery, we revisit the three types of factors making up focal sets, namely novel candidate causes, enabling conditions, and causally irrelevant factors. For the discovery method, we need ways of identifying all three types of factors in a probabilistic distribution.

5.4.2 Factors Identified in a Focal Set

Recall from Chapter 4 that a novel candidate cause of an effect is an event that occurs, and *because* it occurs, the effect of that event happens as well. For example, when the power is functionning normally, *light switch "off"* is a novel candidate cause for *Darkness*, since when *light switch "off"* occurs, the effect *Darkness* occurs as well. However, the definition of novel candidate cause is not sufficient for a functional focal set. There are situations where even when the effect *Darkness* occurs, the factor *light switch "off"* is not present, such as in the context "short circuit". In this context, *Darkness* will be present whether the light switch is "on" or "off". Nonetheless, it would be wrong to deem *light switch "off"* causally irrelevant for the effect *Darkness* altogether. For that reason, Cheng and Novick define an enabling condition.

Recall that an enabling condition is a factor that enables the cause to occur. In our more general definition of enabling condition, if a factor is a novel candidate cause in a particular context, we call it an enabling condition in all other contexts where it is not causal for the same effect. Therefore, since *light switch "off"* is a cause for *Darkness* in the context "power functioning normally," we call it an enabling condition in the context "short circuit." Recall that if we label *light switch "off"* causally irrelevant in a context where it is not causal, the label causally irrelevant holds for the entire distribution, and *Position of light switch* is deemed non-causal for the entire distribution, even where it is causal.

Finally, the last factor to consider is Cheng and Novick's causally irrelevant factor. An event is said to be causally irrelevant if, in its presence, the effect does not occur more frequently, or less frequently than in its absence. For example, *Number of chairs in room* is causally irrelevant for the effect *Darkness*.

In the following section, we suggest an approach to identify the three above mentioned components of focal sets in probability distributions.

5.4.3 Discovering Focal Set Components

In this section, we discuss how the AI tools for uncertain reasoning presented in Chapter 2 can be adapted to discovering focal sets from probabilistic distributions. We use the notion of conditional independence (CI) for discovering causally irrelevant factors, context-specific independence (CSI) to make the distinction between enabling conditions and novel candidate causes. We extend the Refine-CPD algorithm to discover CIs and eliminate causally irrelevant factors from the distribution in question. Then, CSI discovery from the remaining distribution allows us to distinguish between novel candidate causes and enabling conditions.

Identifying Causally Irrelevant Factors

Since a causally irrelevant factor I never contributes to producing an effect E regardless of the context, factor I can safely be removed from the CPD $p(E|X, I)$, where X is the set of all other factors being considered as potential candidate causes for effect E. For that reason, we can say that E and I are conditionally independent (CI) of each other given X, or $p(E|X, I) = p(E|X)$. Therefore, in discovering focal sets pertaining to the effect E, we must first identify all CIs in the CPD $p(E|X, I)$, remove them from the distribution, and label them as causally irrelevant factors in all focal sets for effect E.

For the CI discovery, we present a new algorithm, which operates on subtrees, like the Refine-CPD algorithm for discovering CSIs.

Algorithm 2 *FIND CI*

Input: a CPD $p(E|F_1, F_2, \ldots, F_n)$*, Number of Factors* n

Output: Causally Irrelevant Factors for Effect E

begin

1. $x = 1$

2. **repeat until** $x = n$

3. *build CPD-tree from distribution with* F_x *as root node*

4. **if** *left subtree of* F_x = *right subtree of* F_x

5. *put* F_x *in set of causally irrelevant factors, remove* F_x *from the distribution*

6. $x = x + 1$

7. **end repeat**

8. *return set of irrelevant factors*

end

The algorithm follows from the definition of CI. For effect E, if the factor F_x at the root yields two identical subtrees (i.e. regardless of the value of F_x, probability values are the same when all else remains), then the effect E is independent of the factor F_x, given all other factors in the CPD. An example of discovery of causally irrelevant factors is presented in Section 5.4.4.

Distinguishing Between Novel Candidate Causes and Enabling Conditions

Once the causally irrelevant factors have been identified, all factors remaining in the CPD conditioning on the target effect E should either be enabling conditions or novel candidate causes. Recall that one factor F_x may be a novel candidate cause in one context and an enabling condition in another.

Since a novel candidate cause N is dependent upon the effect E, no independencies should be found between N and E. However, an enabling condition B is a factor that is non-causal in the context where CSI is found, but a novel condidate cause in another context. Therefore, if we find a CSI to hold between the effect E and one of the remaining factors F_x, we know F_x must be a novel candidate cause in another context, otherwise it would have been identified as a causally irrelevant factor in the previous step, and removed from the CPD.

Therefore, in discovering enabling conditions, we must detect CSI in the remaining CPD, after the removal of causally irrelevant factors, for effect E. If we find a CSI to hold between E and F_x, F_x can be identified as an enabling condition in the context where it is discovered, and F_x can also be identified as a novel candidate cause in another context where the particular CSI doesn't hold. This step can be achieved with the Refine-CPD tree algorithm presented in Section 2.4.4. An example of this type of discovery is presented in Section 5.4.4 as well.

5.4.4 Example of Complete Focal Set Discovery

A violent storm has hit the town of Faux Col VII and left the entire town without power. The habitants of Faux Col VII have always believed that without power, there would never be any light (the town was not known for its knowledge in astronomy). However, with this power outage, they have been proved wrong because they still experience light, and are perplexed by the cause of darkness, if not the power outage. They are especially concerned with whether medical facitilies will be left in the dark or will be able to benefit from this mysterious light, which is *not* a result of functioning electricity.

The goal of this example is to identify the cause of darkness in the town's medical facilities to determine whether or not they will function as usual during this prolonged power outage. We simplify the model to illustrate the discovery of focal sets from probabilistic distributions. First, let *(S)etting* be a binary variable taking on two possible values, *local medical clinic* and *operating room*, The value *local medical clinic* represents the location where family physicians meet their patients about everyday ailments, and is identified by value 0, while *operating room* is the town hospital's generator controlled operating room, and is identified by value 1. The second variable is *Position of (L)ight Switch*, a binary variable which takes on the value 0 if the light switch indicates "on", and 1 if the light switch indicates "off". The third binary variable, *(C)arpet Colour*, takes on value 0 if the carpet is black, and value 1 if the carpet is white. Finally, the binary variable *Darkness* takes on value 0 if it is not dark, and value 1 if it is dark. From this information, we build the CPD $p(D|S, L, C)$ illustrated in Table 5.7.

For the CPD in Table 5.7, we first identify CIs to account for causally irrelevant factors. Following Algorithm 2 presented in the previous section, we iterate through all factors as the root node and compare the left and right subtress. Figure 5.8 illustrates the iteration of the algorithm where the root node is *(C)arpet Colour*.

In the CPD-tree in Figure 5.8, we notice that the left subtree is identical to the right subtree,

Table 5.7: Conditional Probability Distribution for Effect *Darkness*, $p(D|S, L, C)$.

S	L	C	D	$p(D|S, L, C)$
0	0	0	0	0.40
0	0	0	1	0.60
0	0	1	0	0.40
0	0	1	1	0.60
0	1	0	0	0.40
0	1	0	1	0.60
0	1	1	0	0.40
0	1	1	1	0.60
1	0	0	0	0.99
1	0	0	1	0.01
1	0	1	0	0.99
1	0	1	1	0.01
1	1	0	0	0.25
1	1	0	1	0.75
1	1	1	0	0.25
1	1	1	1	0.75

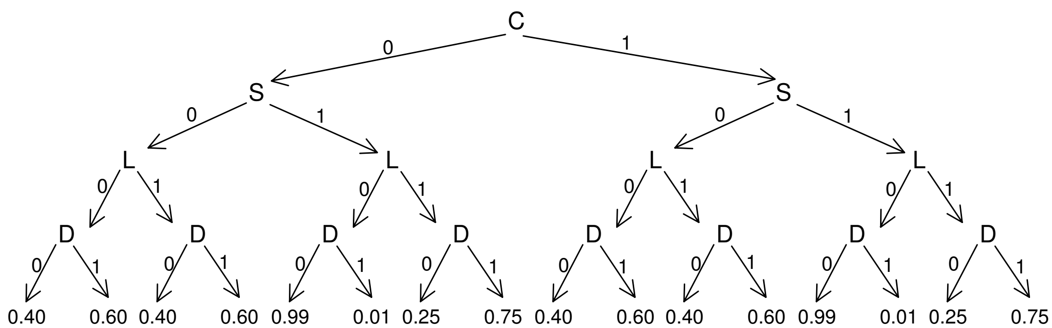

Figure 5.8: CPD-Tree for $p(D|S, L, C)$ for CI Identification with Algorithm 2.

thus making C a causally irrelevant factor. Therefore, C can be labeled as causally irrelevant in every focal set for effect D and removed from the CPD $p(D|S, L, C)$. No other factors are causally irrelevant so we move to distinguishing between enabling conditions and novel candidate causes. The resulting CPD after the removal of causally irrelevant factor C is presented in Table 5.8.

As mentioned previously, enabling conditions are discovered by means of CSI. They represent an independence that holds in one or more, but not all contexts. Since all causally irrelevant factors have been removed from the distribution, we know that if a CSI is found in a context K between and effect E and a factor F_x, F_x is an enabling condition in context K, and a novel candidate cause in other contexts where effect E and factor F_x are dependent upon one another.

In our example, from the remaining CPD $p(D|S, L)$ in Table 5.8, we build the initial CPD-tree, as depicted in Figure 5.9, and apply to it the Refine-CPD algorithm from Section 2.4.4.

We obtain the resulting, refined CPD-tree in Figure 5.10.

The refined CPD-tree in Figure 5.10 entails the decomposed CPD in Table 5.9.

From this decomposed CPD, we see that in the context *(S)etting = local medical clinic*, the effect

Table 5.8: Conditional Probability Distribution for Effect *Darkness*, After Removal of Variable C, $p(D|S, L)$.

S	L	D	$p(D\|S, L)$
0	0	0	0.40
0	0	1	0.60
0	1	0	0.40
0	1	1	0.60
1	0	0	0.99
1	0	1	0.01
1	1	0	0.25
1	1	1	0.75

Table 5.9: Decomposed Conditional Probability Distribution for Effect *Darkness*, After Removal of Variable C, and after CSI Detection.

(i)

SLD	$p(D\|S,L)$
000	0.40
001	0.60
010	0.40
011	0.60
100	0.99
101	0.01
110	0.25
111	0.75

(ii)

SLD	$p(D\|S=0,L)$
000	0.40
001	0.60
010	0.40
011	0.60

SLD	$p(D\|S=1,L)$
100	0.99
101	0.01
110	0.25
111	0.75

(iii)

S D	$p(D\|S=0)$
0 0	0.40
0 1	0.60

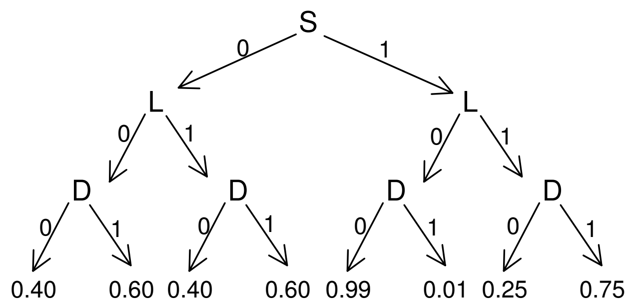

Figure 5.9: Initial CPD-Tree for $p(D|S, L)$ for CSI Identification with Algorithm 1.

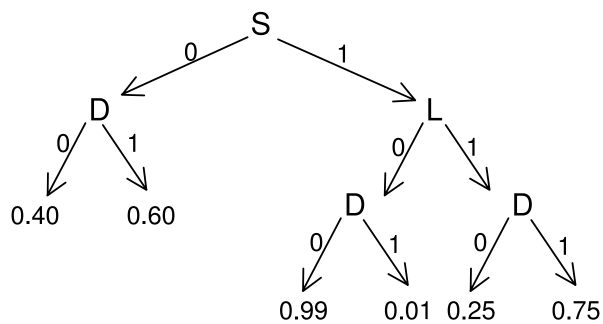

Figure 5.10: Refined CPD-Tree for $p(D|S, L)$ after CSI Identification with Algorithm 1.

(D)arkness is independent of variable *Position of (L)ight Switch*, while in the context *(S)etting = operating room*, the effect *Darkness* and the variable *Position of (L)ight Switch* are dependent upon eachother. Therefore, *Position of (L)ight Switch* is an enabling condition in context *(S)etting = local medical clinic*, and *Position of (L)ight Switch* is a novel candidate cause in the context *(S)etting = operating room*. The discovered focal sets for the effect *Darkness* are illustrated in Figure 5.11.

Once the focal sets are established, other possible causes for darkness or light in the local medical clinic and in the operating room can be investigated. For instance, it could be that the clinic has large windows and lets in natural light throughout a large portion of the day. On the other hand, to avoid infiltration of bacteria, the operating room may have no windows at all, thus only benefiting from light when the light switch is positioned to "on".

In this chapter, we have revisited the three themes mentioned in the introduction, and suggested how contextual consideration can benefit each situation. In problem solving and seemingly paradoxical scenarios, CSI allows us to decompose existing models to let surface the hidden information that may make a situation seem paradoxical. In cases where the hidden information doesn't go as far as reversing inequalities, the hidden information may still make variables seem relevant to the situation when they are not, and vice versa. In attribution theory, CSI discovery has allowed us to better distinguish between situational factors and dispositional causes without reverting back to attributional biases. Finally, we have provided a method for discovering focal sets from probability

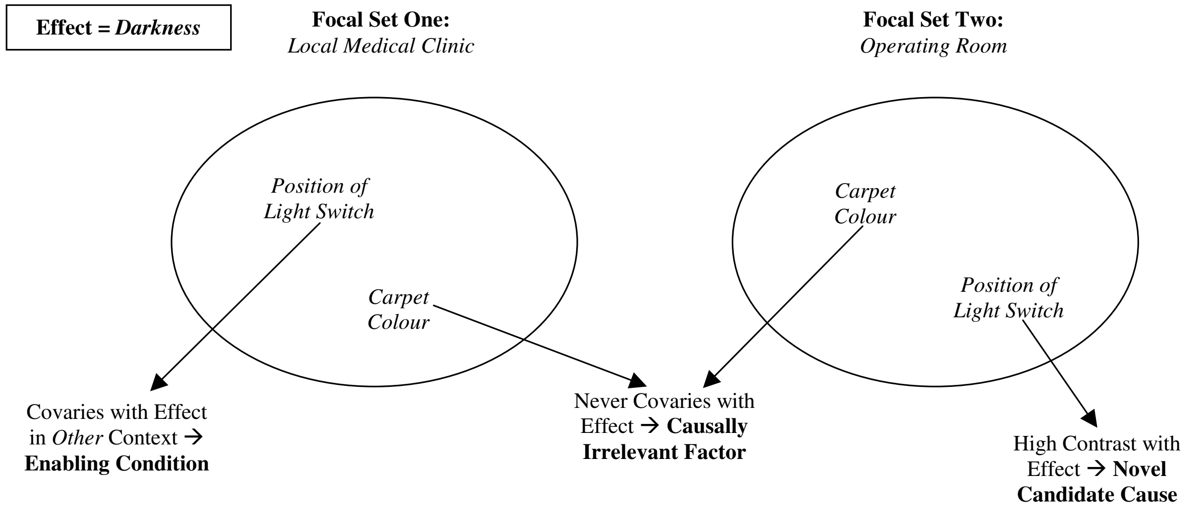

Figure 5.11: Resulting Focal Sets for Effect *Darkness*.

distributions for Cheng and Novick's probabilistic contrast model, without requiring the assistance of the human expert.

Chapter 6

Conclusions and Future Work

Here we summarize the major findings in this thesis and suggest some possible directions for further study.

6.1 Conclusions

This thesis has demonstrated the usefulness of building and also infering contextual models from independencies found in probabilistic distributions for decision making, problem solving, and attribution. The contextual models have proved to be able to clarify erroneous situations by considering information about the situation, that lies inside the distributions. These hidden independencies are also capable of determining surprising contexts where the human expert may fail to see relevance in considering a particular subset of the available information as a unique context. In this thesis, we have presented several results promoting the discovery of contexts from probability distributions and the construction of refined causal models after contextual consideration.

From an algorithmic viewpoint, we have shown that context can play a major role in providing human reasoners with the insight necessary to efficiently decipher large knowledge bases. Once a context is established by the CSI discovery algorithm, the human can then decide whether it is necessary to look elsewhere in the environment for an omitted variable, or whether the information is sufficient and suggests something new about the situation at hand.

Also, we contributed towards a cognitive account of context, primarily by mapping the idea of CSI with Cheng and Novick's concept of focal sets for contextual consideration in their probabilistic contrast model. We showed how contextual discoveries may be used to offer a computational mechanism able to discover focal sets in probabilistic distributions.

The contributions in this thesis have clearly demonstrated a potential for tools in Artificial Intelligence to provide insight in reasoning about everyday matters, the type studied in the cognitive science literature. We focussed primarily on Bayesian networks as a representational and inferencial formalism, and showed that with an understanding of context, not only do BNs provide an efficient reasoning formalism, but also provide an account of several cognitive phenomena.

In terms of problem solving, we showed how contextual consideration can help correct erroneous assumptions used in problem solving. When relevant independencies lie within the distribution of variables, erroneous inference is almost inevitable. In the extreme case, that type of error may lead to seemingly paradoxical scenarios. To emphasize this problem and show how CSI may help solve it, we showed how an instance of Simpson's Paradox could appear in a situation where context has not been considered and described how the situation was resolved by means of CSI. In attribution theory, we showed how a consideration of contexual independencies can help discover hidden dispositions and situational factors in causal relations. Once again, we emphasized that without consideration of CSI, an attribution model could seem intuitively correct, and yet false conclusions would be drawn from it. Finally, in terms of adult judgment of causal induction, we presented a method for discovering focal sets in probabilistic distributions in Cheng and Novick's probabilistic contrast model, where focal sets were previously determined by a human expert.

6.2 Future Work

Consideration of context for acquiring knowledge about a situation and for refining current representational models have not been given a great deal of attention in the literature to date. For that reason, many areas have not yet been explored. In this section, we discuss an idea that could be explored in order to extend the work presented in this thesis. The extension is an explicit consideration of measurement of the data and its impact on the types of deductions we can make from the model.

There are many ways to measure and quantify information to model a situation. Based on the type of measurement that is used, different conditional independencies could be present. As future work, we propose an investigation of the impact of variation of measurement, as well as the potential for contextual consideration to improve the discrepancies between different measurements for the same situation. We illustrate these two issues by means of an example.

6.2.1 Issue 1: Dependence of Two Measurements and How They Yield Different CIs

Imagine a factory where every time the factory produces a chemical C, it also produces a pollutant B.

We assume two measurements, M_1 and M_2, are used to assess whether or not a factory is operating, and both measurements are variables in the causal graph describing the causal relationship between the chemical (C) produced by the factory as well as the byproduct (B). The graph is illustrated in Figure 6.1.

The causal model in Figure 6.1 consists of four binary variables C, B, M_1, and M_2, where variable C represents the chemical intended to be created by the factory, and variable B represents

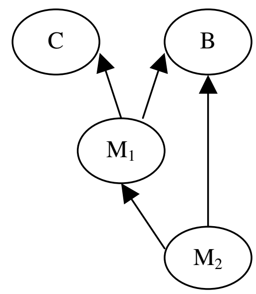

Figure 6.1: Causal model for factory operation.

the bad chemical, byproduct created through the process of fabrication of the good chemical (C). Variable M_1 represents whether or not the factory is operating based on the temperature of the boiler. When it has reached a temperature of 350F, the factory is said to be operating. Variable M_2 is a second indicator of the operating of the factory, where the factory is said to be operating when the boiler switch is set to "on".

When the switch is set to "on", it causes for the temperature of the boiler to rise, but also, it creates a byproduct due to the other processes that are activated upon setting the switch.

When the temperature reaches 350F, the factory starts to produce the chemical it is intended to produce, but also, from that production, an undesirable byproduct is created as well.

If M_1 is used as a measurement to assess whether or not the factory is operating, we can say that C and B are conditionally independent given M_1. Any amount of C produced tells us nothing about the amount of B produced or vice versa, given M.

However, if M_2 is used as a measurement, given M_2, it is impossible to conclude that C and B are independent because of the existence of M_1, which although not being used as a measurement of the factory's activity or inactivity, is a factor in yielding C and B independent of one another.

Given this simple example, we see that it may be worth studying in more depth the impact the choice in measurement may have on the distribution. Perhaps we could answer questions such as: what variables best measure the model?

6.2.2 Issue 2: Measurement Differences Explained with CSI

If we modify the above example slightly, we can show that context has a role to play in selecting a type of measurement.

Assume the factory owner, Mr. O. Zone, for his protection, claims that the pollutant B is produced when the chemical C is used in the sewage treatment plant. He defends his claim with the following argument: if the pollutant were a byproduct of the factory itself, then the occurence of the pollutant would be independent of the occurence of the chemical given that the factory that produced the chemical.

Table 6.1: Initial CPD for factory operation before consideration of measurement.

M	C	B	P(B|M,C)
M1	Y	Y	0.8
M1	Y	N	0.2
M1	N	Y	0.8
M1	N	N	0.2
M2	Y	Y	0.9
M2	Y	N	0.1
M2	N	Y	0.6
M2	N	N	0.4

In fact, what is happening here is that Mr. O. Zone **states** his claim based one measurement of the information but **defends** his claim with information relevant to another scale of measurement of the information.

If we separate the problem into two possible measurements, we obtain a much more adequate explanation of Mr. O. Zone's claim. Although Zone's statements are true, they are inconsistent with the type of measurement he uses to state and defend.

The two measurements we will use to explain the owner's reasoning are M_1 and M_2, where M_1 = *factory is operating when factory production activities begin*, while M_2 = *factory is operating when chemical transformation in sewage begins*. Note that M_1 and M_2 have a different meaning here than in the short example in the previous section.

Since the factory produces the chemical, there is a sure causal relationship between M_x and C regardless of whether $x = 1$ or $x = 2$. Also, based on the distribution in Table 6.1, there is a causal link between C and B as well as M_x and B. With this information, we obtain the causal graph in Figure 6.2.

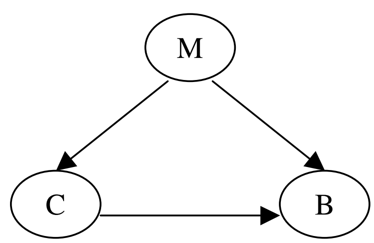

Figure 6.2: Initial Causal model for factory operation before consideration of measurement.

According to the above causal graph, Mr. O. Zone is absolutely correct in making and reasoning his argument. However, the subtleties of the measurement corrupt the validity of Mr. O. Zone's claim. Mr. O. Zone is layering both measurements to make his argument valid and claiming no independence between C and B. A close inspection of the distribution shows that, in accordance with Mr. O. Zone's claim, in the context $M = M_1$, C and B are in fact independent, while they are dependent in the context $M = M_2$. When M_1 is the preferred measurement, the fabrication

Table 6.2: CSI decomposition of CPD for factory operation after consideration of measurement.

M	C	B	P(B|M,C)
M1	Y	Y	0.8
M1	Y	N	0.2
M1	N	Y	0.8
M1	N	N	0.2
M2	Y	Y	0.9
M2	Y	N	0.1
M2	N	Y	0.6
M2	N	N	0.4

M	B	P(B|M=M1)
M1	Y	0.8
M1	N	0.2

M	C	B	P(B|M=M2,C)
M2	Y	Y	0.9
M2	Y	N	0.1
M2	N	Y	0.6
M2	N	N	0.4

of the chemical is independent of the creation of a byproduct. Knowing the amount of chemical fabricated tells us nothing about the state of the byproduct. On the other hand, when M_2 is the preferred measurement, C has a direct causal influence on the presence of B in the air. Table 6.2 shows the CSI decomposition of the distribution based on the CSI $I_{M=M_1}(\{B\},\{M\},\{C\})$.

The type of decomposition illustrated in Table 6.2 is the type of decomposition we have been using throughout the thesis for building contextual models. Based on the results in the thesis, the logical result in this situation would be that, in the context $M = M_1$, the causal link from C to B can be deleted, creating two more accurate causal models for the two contexts of M, namely M_1 and M_2, as depicted in Figure 6.3.

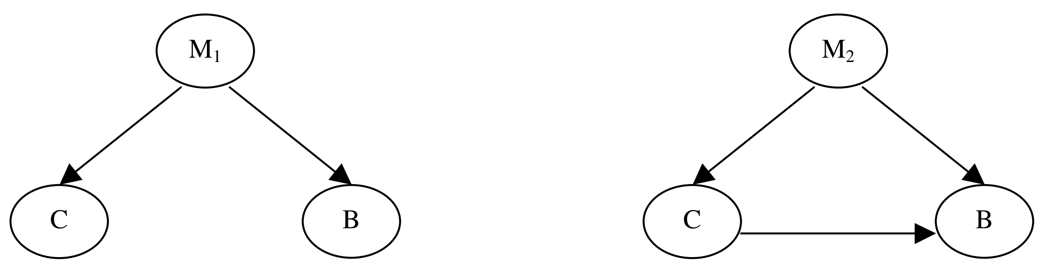

Figure 6.3: Refined causal model for factory operation after consideration of measurement.

With the two contexts of measurement considered exclusively as in the causal graphs above, Mr. O. Zone can no longer escape the responsibility of creating a byproduct through the operations of his factory.

This example, although very preliminary, shows the potential for considering measurement as an issue that can be addressed by means of contextual models.

References

[1] G.B. Saxe. Selling candy: A study of cognition in context. In M. Cole, Y. Engestroem, and et al, editors, *Mind, culture, and activity: Seminal papers from the Laboratory of Comparative Human Cognition*, pages 330–337, New York, NY, US, 1997. Cambridge University Press.

[2] D. Kirshner and J.A. Whitson. *Situated Cognition: Social, semiotic, and psychological perspectives*. Erlbaum, Mahwah, NJ, 1997.

[3] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in bayesian networks. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 115–123, 1996.

[4] S.K.M. Wong and C.J. Butz. Contextual weak independence in bayesian networks. In *Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 670–679, 1999.

[5] Nevin Lianwen Zhang and David Poole. On the role of context-specific independence in probabilistic inference. In *IJCAI '99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1288–1293, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[6] C.J. Butz and M.J. Sanscartier. On the role of contextual weak independence in probabilistic inference. In *Fifteenth Canadian Conference on Artificial Intelligence (AI)*, pages 185–194, 2002.

[7] C.J. Butz and M.J. Sanscartier. A method for detecting context-specific independence in conditional probability tables. In *3rd International Conference on Rough Sets and Current Trends in Computing (RSCTC02)*, pages 344–348, 2002.

[8] C.J. Butz and M.J. Sanscartier. Acquisition methods for contextual weak independence. In *3rd International Conference on Rough Sets and Current Trends in Computing (RSCTC02)*, pages 339–343, 2002.

[9] J.F. Wagner. Transfer in pieces. *Cognition and Instruction*, 24(1):1 – 71, 2006.

[10] D. Grodner, E. Gibson, and D. Watson. The influence of contextual contrast on syntactic processing: evidence for strong-interaction in sentence comprehension. *Cognition*, 95(3):275 – 296, 2005.

[11] E. Kaiser and J.C. Trueswell. The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94(2):113 – 147, 2004.

[12] L. Zhu and G. Gigerenzer. Children can solve bayesian problems: The role of representation in mental computation. *Cognition*, 98(3):287 – 305, 2006.

[13] T. Regier and S. Gahl. Learning the unlearnable: the role of missing evidence. *Cognition*, 93:147 – 155, 2004.

[14] R. Treiman, B. Kessler, and S. Bick. Influence of consonantal context on the pronunciation of vowels: a comparison of human readers and computational models. *Cognition*, 88(1):49 – 78, 2003.

[15] M. Buehner and P.W. Cheng. Causal learning. In K.J. Holyoak and R.G. Morrison, editors, *Handbook of Thinking and Reasoning*, pages 143 – 168, New York, NY, 2005. Cambridge University Press.

[16] A. Majid, A.J. Sanford, and M.J. Pickering. Covariation and quantifier polarity: What determines causal attribution in vignettes? *Cognition*, 99(1):35 – 51, 2005.

[17] K.H. Teigen and G. Keren. Surprises: low probabilities or high contrasts? *Cognition*, 87(2):55 – 71, 2003.

[18] M.R. Waldmann and Y. Hagmayer. Estimating causal strength: the role of structural knowledge and procesing effort. *Cognition*, 82(1):27 – 58, 2001.

[19] P.W. Cheng and L.R. Novick. A probabilistic contrast of causal induction. *Journal of Personality and Social Psychology*, 58:545–567, 1990.

[20] P.W. Cheng and L.R. Novick. Covariation in natural causal induction. *Psychological Review*, 99:365–382, 1992.

[21] P.W. Cheng. From covariation to causation: A causal power theory. *Psychology Review*, 104:367–405, 1997.

[22] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, 77(2):257 – 286, 1989.

[23] M. Drummond. Situated control rules. In *First International Conference on Principles of Knowledge Representation and Reasoning*, pages 339–343, Toronto, 1989. Morgan Kaufmann.

[24] M.J. Schoppers. Universal plans for reactive robots in unpredictable environments. In *Tenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1039–1046, Milan, 1987.

[25] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[26] D. McDermott and J. Doyle. Nonmonotonic logic 1. *Artificial Intelligence*, 13:41–72, 1980.

[27] J. McCarthy. Circumscription: A form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[28] G. R. Simari and R. P. Loui. A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence*, 53(2-3), 1992.

[29] D.L. Poole. On the comparison of theories: Preferring the most specific explanation. In *Ninth International Joint Conference on Artificial Intelligence*, pages 144–147, 1985.

[30] P. Gardenfors. *Belief Revision: an introduction*. Cambridge University Press, 1992.

[31] H. Reichenbach. *The theory of probability*. University of California Press, Berkeley, CA, 1949.

[32] H. E. Kyburg. The reference class. *Philosophy of Science*, 50, 1983.

[33] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Fransisco USA, 1988.

[34] D. Poole. Probabilistic partial evaluation: Exploiting rule structure in probabilistic inference. In *IJCAI '97: Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 1284–1291, 1997.

[35] D. Poole and N. Zhang. Exploiting contextual independence in probabilistic inference. *JAIR*, 18:263–313, 2003.

[36] M.J. Sanscartier and E. Neufeld. Causality, simpson's paradox, and context-specific independence. In *Eighth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 233–243, 2005.

[37] H.H. Kelley. Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, and B. Weiner, editors, *Attribution: Perceiving the causes of behavior*, pages 1–26, Morristown, N.J., 1972. General Learning Press.

[38] J.E. Alcock, D.W. Carment, and S.W. Sadava. *A Textbook of Social Psychology.* Prentice-Hall Canada, Toronto, 5th edition edition, 2005.

[39] E.E. Jones and K.E. Davis. From acts to dispositions: The attribution process in person perception. In L. Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 2, Orlando, FL, 1965. Academic Press.

[40] H.H. Kelley. Attribution theory in social psychology. In D. Levine, editor, *Nebraska Symposium on Motivation*, pages 192–238. Lincoln: University of Nebraska Press, 1967.

[41] B. Weiner. *Achievement motivation and attribution theory.* General Learning Press, Morristown, N.J., 1974.

[42] L. Ross. The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in experimental social psychology*, 10:173–240, 1977.

[43] W.M. Bernstein, W.G. Stephan, and M.H. Davis. Explaining attributions for achievement. a path analytic approach. *Journal of Personality and Social Psychology*, 37:1810–1821, 1979.

[44] E. Walster. Assignment of responsibility for an accident. *Journal of Personality and Social Psychology*, 3:73–79, 1966.

[45] E.J. Langer. The illusion of control. *Journal of Personality and Social Psychology*, 32:311–328, 1975.

[46] M.J. Sanscartier and E. Neufeld. Discovering hidden dispositions and situational factors in causal relations by means of contextual independencies. In *to appear in Twenty-Eighth Annual Conference of the Cognitive Science Society*, 2006.

[47] L.B. Alloy and N. Tabachnick. Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, 91:112–149, 1984.

[48] S.M. Kassin. Consensus information, prediction, and causal attribution: A review of the literature and issues. *Journal of Personality and Social Psychology*, 37:1966–1981, 1979.

[49] B.R. Orvis, J.D. Cunningham, and H.H. Kelley. A closer examination of causal inference : The role of consensus, distinction and consistency information. *Journal of Personality and Social Psychology*, 32:605–616, 1975.

[50] Amos Tversky and Daniel Kahneman. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge University Press, New York, 1982.

[51] D.J. Hilton and B.R. Slugoski. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1):75–88, 1986.

[52] J.S. Bruner. On perceptual readiness. *Psychological Review*, 64:123–152, 1957.

[53] R.J. Harris. Comprehension of pragmatic implications in advertising. *Journal of Applied Psychology*, 62:603–608, 1977.

[54] M.K. Johnson, J.D. Bransford, and S.K. Solomon. Memory for tacit implications of sentences. *Journal of Experimental Psychology*, 98:203–205, 1973.

[55] D. Hume. *A Treatise of Human Nature*. Millar, London, 1739/1987.

[56] I. Kant. *Critique of pure reason*. St-Martin's, New York, 1781/1965.

[57] R. von Mises. *Probability, Statistics and Truth, 2nd rev. English ed.* Dover, New York, 1981.

[58] R. Carnap. *Formalization of Logic*. Harvard University Press, Boston, 1943.

[59] J. McCarthy and P.J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4, 1969.

[60] E. Castillo, J. M. Gutirrez, and A. S. Hadi. *Expert Systems and Probabilistic Network Models*. Springer-Verlag, New York, 1997.

[61] J. Pearl and T.S. Verma. A theory of infered causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452. Morgan Kaufmann, 1991.

[62] David Poole. Context-specific approximation in probabilistic inference. In *UAI*, pages 447–454, 1998.

[63] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, USA, 2000.

[64] M.J. Sanscartier. *Probabilistic Reasoning Using General Forms of Conditional Independence*. M.Sc. Thesis, University of Regina, 2002.

[65] L.H. Strickland. Surveillance and trust. *Journal of Personality*, 26:200–215, 1958.

[66] F. Heider. *The Psychology of Interpersonal Relations*. Wiley, New York, 1958.

[67] D.T. Gilbert. How metal systems believe. *American Psychologist*, 46:107–119, 1991.

[68] D.T. Gilbert. Ordinary personology. *The handbook of social psychology*, 2:89–150, 1998.

[69] ChangingMinds.org. Correspondent inference theory. Maintained by Syque 2002 – 2006 `http://changingminds.org/explanations/theories/correspondent_inference.htm` , 2006.

[70] E.E. Jones and V.A. Harris. The attribution of attitudes. *Journal of Experimental Social Psychology*, 3:1–24, 1967.

[71] E.E. Jones, K.E. Davis, and K. Gergen. Role playing variations and their informational value for person perception. *Journal of Abnormal and Social Psychology*, 63:302–310, 1961.

[72] R.E. Nisbett and L. Ross. *Human inference: Strategies and shortcomings of social judgement*. Prentice-Hall, Inc, Englewood Cliffs, N.J., 1980.

[73] M.A. Hogg and G.M. Vaughan. *Social Psychology. Third Edition*. Prentice Hall, 2002.

[74] J.S. Mill. Collected works of john stuart mill. *System of logic*, 7&8, 1872/1973.

[75] H.H. Kelley and J.L. Michela. The processes of causal attribution. *American Psychologist*, 28:107–128, 1973.

[76] H.H. Kelley and J.L. Michela. Attribution theory and research. *Annual Review of Psychology*, 31:457–501, 1980.

[77] L.A. McArthur. The how and what of why: Some determinants and consequences of causal attributions. *Journal of Personality and Social Psychology*, 22:171–193, 1972.

[78] R. Weiss. *Marital Separation*. Basic Books, Inc., New York, 1975.

[79] L.A. Peplau and D. Perlman. *Loneliness: A sourcebook of current theory, research and therapy*. John Wiley & Sons, New York, USA, 1982.

[80] B. Weiner. *Human Motivation*. Holt, Rinehart & Winston, New York, USA, 1980.

[81] B. Weiner. A theory of motivation for some classroom experiences. *Journal of Educational Psychology*, 71:3–25, 1979.

[82] I.H. Frieze and B. Weiner. Cue utilization and attribution judgments for success and failure. *Journal of Personality*, 39(4):591–606, 1971.

[83] D. Watson. The actor and the observer: How are their perceptions of causality divergent? *Psychological Bulletin*, 92:682–700, 1982.

[84] R.E. Nisbett, C. Caputo, P. Legant, and J. Marecek. Behavior as seen by the actor and as seen by the observer. *Journal of Personality and Social Psychology*, 27:154–164, 1973.

[85] M.D. Storms. Videotape and the attribution process: Reversing actor's and observer's points of view. *Journal of Personality and Social Psychology*, 27:165–175, 1973.

[86] G.W. Bradley. Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology*, 36:56–71, 1978.

[87] J.M. Burger. Motivational biases in the attribution of responsibility for an accident: A meta-analysis of the defensive-attribution hypothesis. *Psychological Bulletin*, 90:496–512, 1981.

[88] K.G. Shaver. Defensive attribution; effects of severity and relevance on the responsibility assigned for an accident. *Journal of Personality and Social Psychology*, 14:101–113, 1970.

[89] C.B. Wortman. Some determinants of perceived control. *Journal of Personality and Social Psychology*, 31:282–294, 1975.

[90] M. Henle. On the relation between logic and thinking. *Psychological Review*, 69:366–378, 1962.

[91] E. Bordiga and N. Brekke. The base rate fallacy in attribution and prediction. In J.H. Harvey, W. Ickes, and R.F. Kidd, editors, *New directions in attribution research*, volume 3, pages 63–95, Hillsdale, NJ, 1981. Erlbaum.

[92] J. Jaspars, M. Hewstone, and F.D. Fincham. Attribution theory and research: The state of the art. In J. Jaspars, F. D. Fincham, and M. Hewstone, editors, *Attribution Theory and Research: Conceptual Development and Social Dimensions*, pages 3–36, London, UK, 1983. Academic Press.

[93] R.D. Hansen and J.M. Donoghue. The power of consensus: Information derived from one's own and other's behavior. *Journal of Personality and Social Psychology*, 35:294–302, 1977.

[94] M. Hewstone and J. Jaspars. A re-examination of the roles of consensus, consistency and distinctiveness: Kelley's cube revisited. *British Journal of Social Psychology*, 22(1):41–50, 1983.

[95] J.A. Kulik and S.E. Taylor. Premature consensus on consensus? effects of sample-based versus self-based consensus information. *Journal of Personality and Social Pyschology*, 38:871–878, 1980.

[96] D.J. Pruitt and C.A. Insko. Extension of the kelley attribution model: The role of comparison-object consensus, target-object consensus, distinctiveness, and consistency. *Journal of Personality and Social Psychology*, 39:39–58, 1980.

[97] D.N. Ruble and N.S. Feldman. Order of consensus, distinctiveness, and consistency information and causal attribution. *Journal of Personality and Social Psychology*, 34:930–937, 1981.

[98] B. Weiner. *Perceiving the causes of success and failure.* General Learning Press, Morristown, N.J., 1974.

[99] G.L. Wells and J.H. Harvey. Do people use consensus information in making causal attributions? *Journal of Personality and Social Psychology*, 35:279–293, 1977.

[100] M. Zuckerman. Actions and occurrences in kelley's cube. *Journal of Personality and Social Psychology*, 36:647–656, 1978.

[101] F. Försterling. Models of covariation and attribution : How do they relate to the analogy of analysis of variance? *Journal of Personality and Social Psychology*, 57:615–625, 1989.

[102] L.B. Alloy and L.Y. Abramson. Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology*, 108:441–485, 1979.

[103] J. Robinson. *Essays in the Theory of Economic Growth.* Martin's Press, New York, USA, 1964.

[104] E.H. Shuford. Percentage estimation of proportion as a function of element type, exposure time, and task. *Journal of Personality and Social Psychology*, 61:430–436, 1970.

[105] H.J. Einhorn and R.M. Hogarth. Judging probable cause. *Psychology Bulletin*, 99:3–19, 1986.

[106] M.L.A. Hart and A.M. Honore. *Causation in The Law.* Clarendon Press, Oxford, 1985.

[107] D.J. Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107:65–81, 1990.

[108] D. Kahneman and D.T. Miller. Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93:136–153, 1986.

[109] J.L. Mackie. Causes and conditions. *American Philosophical Quarterly*, 2:245–255, 1965.

[110] J.L. Mackie. *The Cement of the Universe.* Clarendon Press, Oxford, 1974.

[111] S.E. Taylor. Adjustment to threatening events: a theory of cognitive adaption. *American Psychologist*, 38:1161–1173, 1983.

[112] W. Ahn and J. Bailenson. Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, 31:82–123, 1996.

[113] W. Ahn, C.W. Kalish, D.L. Medin, and S.A. Gelman. The role of covariation versus mechanisminformation in causal attribution. *Cognition*, 54:299–352, 1995.

[114] M. Bullock, R. Gelman, and R. Baillargeon. The development of causal reasoning. *The developmental psychology of time, W. J. Friedman (Ed.)*, pages 209–254, 1982.

[115] R. Harre and E.H. Madden. *Causal powers: A theory of natural necessity.* Basil Blackwell, Oxford, 1975.

[116] A. Michotte. *The perception of causality (trans. T. R. Miles & E. Miles).* Basic Books, New York, 1963.

[117] P.A. White. A theory of causal processing. *British Journal of Pychology*, 80:431–454, 1989.

[118] P.A. White. Use of prior beliefs in the assignment of causal roles: Causal powers versus covariation-based accounts. *Memory & Cognition*, 23:243–254, 1995.

[119] A.G. Baker, P. Miercier, F. Vallee-Tourangeau, R. Frank, and M.F. Pan. Selective associations and causality judgments: The presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19:414–432, 1993.

[120] D.R. Shanks. Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17:433–443, 1991.

[121] C. Glymour. *The Mind's Arrows: Bayes nets and graphical causal models in psychology.* MIT Press, Cambridge, Massachusetts, 2001.

[122] E.H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, 13(B):238–241, 1951.

[123] M.R. Cohen and E. Nagel. *An Introduction to Logic and Scientific Method*. Brace and Co., New York: Harcourt, 1934.

[124] L.K. Dennis, L.F. Beane Freeman, and M.J. Vanbeek. Sunscreen use and the risk for melanoma: a quantitative review. *Annals of Internal Medicine*, 139(12):966–978, 2003.

Appendix A

Verify Validity of Conditional Independencies (CIs) in Probability Distributions

In this appendix, we illustrate how to detect a CI from an MPD.

In the joint distribution $p(A, B, C, D)$ in Figure A.1, we want to verify if variables A and C are conditionally independent given variable B. In other words, does the CI $I(A, B, C)$ hold in $p(A, B, C, D)$? This question can be answered with the alternative definition of CI, namely Equation (2.2). If the CI $I(A, B, C)$ holds, then according to the definition, the following is also true

$$p(A, B, C) \ = \ \frac{p(A, B) \cdot p(B, C)}{p(B)}. \tag{A.1}$$

In order to verify the validity of the above equation, we must compute the marginal of $p(A, B, C, D)$ on $\{A, B, C\}$, as shown in Figure A.2. We compare the result to the product of the marginal on $\{A, B\}$ and on $\{B, C\}$ and divide it by the marginal on $\{B\}$. This result is shown in Figure A.3. Since the resulting distribution in Figure A.3 (right) is the same as the one in Figure A.2, we can conclude that $I(A, B, C)$ holds in the given distribution.

A	B	C	D	$p(A, B, C, D)$
0	0	0	0	0.1
0	0	0	1	0.1
0	0	1	1	0.2
1	0	0	0	0.1
1	0	1	0	0.1
1	1	1	1	0.4

Figure A.1: An example joint distribution for reading CI.

A	B	C	$p(A, B, C)$
0	0	0	0.2
0	0	1	0.2
1	0	0	0.1
1	0	1	0.1
1	1	1	0.4

Figure A.2: The marginal $p(A, B, C)$ of $p(A, B, C, D)$ in Figure A.1.

A	B	$p(A,B)$
0	0	0.4
1	0	0.2
1	1	0.4

$\cdot$

B	C	$p(B,C)$
0	0	0.3
0	1	0.3
1	1	0.4

$/$

B	$p(B)$
0	0.6
1	0.4

$=$

A	B	C	$p(A,B,C)$
0	0	0	0.2
0	0	1	0.2
1	0	0	0.1
1	0	1	0.1
1	1	1	0.4

Figure A.3: The marginals $p(A,B)$, $p(B,C)$, and $p(B)$ of $p(A,B,C,D)$ in Figure A.1, and the resulting marginal $p(A,B,C) = p(A,B) \cdot p(B,C)/p(B)$.