Toward a formal representation of radical interpretation

MSc Thesis *(Afstudeerscriptie)*

written by

Eric Flaten
(born November 17th, 1965 in Elbow Lake, United States)

under the supervision of **Prof. dr. ing. R.A.M. van Rooij** and **Prof. dr. M.J.B. Stokhof**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
April 30, 2020	Dr. Benno van den Berg (Chair)
	Prof. dr. M. Aloni
	Prof. dr. ing. R.A.M. van Rooij (Supervisor)
	Prof. dr. M.J.B. Stokhof (Supervisor)
	Prof. dr. F.J.M.M. Veltman

Institute for Logic, Language and Computation

Acknowledgement

The models described in this paper were created using the GeNIe Modeler, available free of charge for academic research and teaching use from BayesFusion, LLC, http://www.bayesfusion.com/.

Contents

List of Figures

Chapter 1

Introduction

What is meaning? This question is important, because the answers it elicits provide insights into philosophy, logic, linguistics, and other fields of research with which these three interact. Yet, Donald Davidson was frustrated with the inadequate answers that were offered at the time, so he asked a different "less intractable" question: "What would it suffice an interpreter to know in order to understand the speaker of an alien language, and how could he come to know it?" (Davidson 1994, p. 126) In the process of pursuing answers to his question, he created the thought experiment of radical interpretation which was first published in Davidson (Davidson 1973, reprinted in Davidson 1984a) and (Davidson 1974). Later, Davidson restated his question as a "doubly hypothetical" one:

- Given a theory that would make interpretation possible, what evidence plausibly available to a potential interpreter would support the theory to a reasonable degree? (Davidson 1973, reprinted in Davidson 1984a, p. 125)

For the sake of clarity and ease of exposition the above "doubly hypothetical" question will be separated into the following two questions:

- What kind of theory would make interpretation possible?
- What plausibly available evidence would allow potential interpreters to tell that the theory was correct?

Why do these questions matter to Davidson? One is that we will have better answers for the question "What is meaning?" and better insight on the intentional for he says, "we will gain an important insight into the nature of the intentional (including, of course, meaning), in particular into how the intentional supervenes on the observable and non-intentional." (Davidson 1994, p. 124) One way this will happen is that we will be able to go from patterns among observable facts (such as a person's linguistic behavior) to

"facts of a more sophisticated kind (degree of belief, comparison of differences in value)". (Davidson 1980, reprinted in Davidson 1984a, p. 154)

Why does Davidson call his theory *radical* interpretation? Because his radical interpreter begins the process of interpretation without any knowledge of her speaker or his language. In other words she has to interpret "from scratch". Also, the interconnection between belief and meaning makes her situation even more challenging as highlighted below.

> The interdependence of belief and meaning is evident in this way: a speaker holds a sentence to be true because of what the sentence (in his language) means, and because of what he believes. Knowing that he holds the sentence to be true, and knowing the meaning, we can infer his belief; given enough information about his beliefs, we could perhaps infer the meaning. But radical interpretation should rest on evidence that does not assume knowledge of meanings or detailed knowledge of beliefs. (Davidson 1973, reprinted in Davidson 1984a, p. 134)

Since the radical interpreter begins without any interpretations or beliefs and needs each to help with the other, Davidson says "we must somehow deliver simultaneously a theory of belief and a theory of meaning".[1] (Davidson 1974, p. 312) Therefore, in searching for a single theory for interpreting, Davidson has to find or create two. The first is a theory of interpretation designed along a Tarski-style theory of truth. (Davidson 1973, Davidson 1974) The second is a theory of belief based on the decision theories of (Ramsey 1931) and (Jeffrey 1965). (Davidson 1974)

1.1 Davidson's theory of interpretation

Davidson's answer to "What kind of theory would make interpretation possible?" is a theory of truth in Tarski's style that is modified to apply to natural language. (Davidson 1973, reprinted in Davidson 1984a, p. 130) A Tarski-style theory of truth entails for every sentence s in the object language a sentence of the following form which Tarski called Convention T:

s is true (in the object language) if and only if p.

[1] At least since Davidson 1967, Davidson's use of the phrase "theory of meaning" is non-standard. Instead of this phrase referring to an abstract concept of meaning apart from any specific language, Davidson's use of this phrase is language-specific. A more accurate phrase is "a semantic theory for language-L". In light of this when directly quoting Davidson the phrase "theory of meaning" will be replaced with "[a semantic theory of a language]" using the square brackets $[\cdot]$ to signal this switch. Thanks to Martin Stokhof for pointing out this non-standard use.

Specific instances of this form are created by replacing s with a canonical description of it and replacing p by a translation of s. Convention T uses an undefined semantic notion called *satisfaction* which relates sentences to infinite sequences of objects from the variables of the object language. Tarski provided a finite number of axioms: "some give the conditions under which a sequence satisfies a complex sentence on the basis of the conditions of satisfaction of simpler sentences, others give the conditions under which the simplest (open) sentences are satisfied. Truth is defined for closed sentences in terms of the notion of satisfaction." (Davidson 1973, reprinted in Davidson 1984a, p. 131)

Tarski designed his theory of truth for formal languages which do not have indexical objects such as "I", "here", or "now". But Davidson says natural languages are "replete with indexical features, like tense, and so their sentences may vary in truth according to time and speaker". (Davidson 1973, reprinted in Davidson 1984a, p. 131) In light of this Davidson says, "The remedy is to characterize truth for a language relative to a time and a speaker. The extension to utterances is again straightforward." (Davidson 1973, reprinted in Davidson 1984a, p. 131)

Davidson (Davidson 1973, reprinted in Davidson 1984a, p. 131) claims that a Tarski-like theory of truth that has been modified to fit natural language can be used for a theory of interpretation and defends this claim by asking and answering the three questions below:

1. Can a theory of truth be given for a natural language?
2. Can a theory of truth be verified by appeal to evidence available before interpretation has begun?
3. If the theory were known to be true, would it be possible to interpret utterances of speakers of the language?

Below is a summary of his answers to these questions.

1. *Can a theory of truth be given for a natural language?* (Davidson 1973, reprinted in Davidson 1984a, p. 132)

Davidson believes this is possible and proposes two stages for applying a theory of truth in detail to a natural language. (Davidson 1973, p.132) Stage One involves characterizing truth for "a carefully gerrymandered part of the language", which will "no doubt [be] clumsy grammatically", will involve "an infinity of sentences which exhaust the expressive power of the whole language" and these sentences will give "the logical form, or deep structure, of all sentences." (Davidson 1973, reprinted in Davidson 1984a, p. 133)

Stage Two involves matching each of the remaining sentences to sentences in Stage One.

2. *Can a theory of truth be verified by appeal to evidence available before interpretation has begun?* (Davidson 1973, reprinted in Davidson 1984a, p. 133)

In answering Question 2, Davidson proposes one change and makes some observations about T-sentences. One, he proposes reversing how Convention T is used: "[By] assuming translation, Tarski was able to define truth; the present idea is to take truth as basic and to extract an account of translation or interpretation." (Davidson 1973, reprinted in Davidson 1984a, p. 134) Two, "T-sentences mention only the closed sentences of the language, so the relevant evidence can consist entirely of facts about the behaviour and attitudes of speakers in relation to sentences (no doubt by way of utterances)." (Davidson 1973, reprinted in Davidson 1984a, p. 131) Three, "truth is a single property which attaches, or fails to attach, to utterances, while each utterance has its own interpretation; and truth is more apt to connect with fairly simple attitudes of speakers." Four, Davidson suggests using the attitude of a speaker holding true a sentence, because of the principle of charity an interpreter assumes that when a speaker makes an utterance he holds that utterance true. In this sense she can tell that a speaker holds a sentence true even if she has no idea what that sentence means. With these observations and changes he says:

> There is no difficulty in rephrasing Convention T without appeal to the concept of translation: an acceptable theory of truth must entail, for every sentence s of the object language, a sentence of the form:
>
> s is true if and only if p, where 'p' is replaced by any sentence that is true if and only if s is.
>
> Given this formulation, the theory is tested by evidence that T-sentences are simply true; we have given up the idea that we must also tell whether what replaces 'p' translates s. (Davidson 1973, reprinted in Davidson 1984a, p. 134)

3. *If the theory were known to be true, would it be possible to interpret utterances of speakers of the language?* (Davidson 1973, reprinted in Davidson 1984a, p. 138)

Davidson gives two answers to Question 3. On the one hand if the situation is interpreting an isolated sentence or utterance, Davidson's answer is negative:

> A T-sentence does not give the meaning of the sentence it concerns: the T-sentences does fix the truth value relative to certain conditions, but it does not say the object language sentence is true *because* the conditions hold. (Davidson 1973, reprinted in Davidson 1984a, p. 138, italics by Davidson)

On the other hand if the situation is one of interpreting one sentence within the context of all the other sentences, then Davidson's answer is affirmative:

> We can interpret a particular sentence provided we know a correct theory of truth that deals with the language of the sentence. For then we know not only the T-sentence for the sentence to be interpreted, but we also 'know' the T-sentences for all other sentences. (Davidson 1973, reprinted in Davidson 1984a, p. 138)

Yet along with this affirmative answer Davidson admits that some indeterminacy is expected. (Davidson 1973 reprinted in Davidson 1984a, p. 139) That is, given a set of utterances by a speaker more than one set of interpretations could be given such that each are theoretically-valid. However, Davidson says he expects the amount of indeterminacy in his theory will be less than that of Quine's theory of radical translation. (Davidson 1973 reprinted in Davidson 1984a, p. 139)

1.2 Davidson theory of belief

Recall that in order for a radical interpreter to discover the possible interpretations of a speaker's utterance she will need to know what the speaker believes when making his utterances. Because of this Davidson has to provide two theories that simultaneously work together: a theory of interpretation and a theory of belief. His theory of interpretation is given above; his theory of belief briefly described below.

Davidson's theory of belief is a version of Bayesian decision theory that is derived from the decision theories of (Ramsey 1931) and (Jeffrey 1965). Davidson says that Ramsey used "an ingenious trick"[2] that created a decision theory that could take as input the preferences a subject has when choosing among various gambles and calculate the degrees of belief and the cardinal utilities of that subject.[3] (Davidson 1984b, p. 156) The degrees of

[2] This "trick" is explained in Chapter 2.

[3] A ordinal utility function gives an order on outcomes (for example, an ordinal utility function for John could say he prefers outcome A first, outcome D second, and K third) whereas a cardinal utility function gives specific numbers for outcomes (for example, a cardinal utility function for John could say John would pay €50 for A, €10 for D, and €1 for K).

belief are represented by subjective probabilities that a certain state of affairs would happen. That is, if a subject believed a certain event A had 75% chance of happening, then his subjective probability would be $P(A) = 0.75$. Davidson states he wants his theory of belief to operate along similar lines. (Davidson 1980, reprinted in Davidson 1984a, p. 155) However, Ramsey's theory is not suitable for Davidson's radical interpretation theory, because of two problems, each stemming from the fact that Ramsey's theory is fundamentally based on gambles. (Ramsey 1931, p. 183) The lesser problem is known as the presentation problem. Davidson claims "It is well known that two descriptions of what the experimenter takes to be the same option may elicit quite different responses from a subject." (Davidson 1974, p. 315) An example of this is the Asian-disease problem created by Tversky and Kahneman (Tversky and Kahneman 1981, p. 453), which is given in detail in Chapter 2, but briefly stated even though the two problems below describe identical outcomes, a majority of people choose option A in the first and option B in the second.[4] Theoretically it might be possible to devise a way to prevent any presentation problem from happening. Unfortunately, no such solution is available for the second problem.

Problem #1:

- If Program A1 is adopted, 200 people will be saved.
- If Program B1 is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

Problem #2:

- If Program A2 is adopted 400 people will die.
- If Program B2 is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

The second problem stems from the function of decision theory in Davidson's unified theory which is this: the speaker's cardinal utilities need to be derived so that the speaker's subjective probabilities (which are his beliefs) can be calculate so that possible interpretations can be given to the speaker's utterances. Therefore, the end of this process is to interpret utterances, but Davidson says, "[It is unreasonable] to imagine we can justify the attribution of preferences among complex options unless we can interpret speech behavior." (Davidson 1974, p. 315) In other words, the second problem is this: in a decision theory experiment the experimenter describes various gambles from which the subject chooses, which means the experimenter is quite far along the process of interpreting the subjects language, which contradicts

[4]The original letters in A and B for Problem #1 and C and D for Problem #2. I changed them to A1, B1, A2 and B2 for ease of reading.

the assumption that the radical interpreter (who will take the place of the experimenter) knows nothing at all about the speaker's language. Therefore, this problem is severe enough by itself to eliminate the possibility of using Ramsey's theory as-is, which is why Davidson turns to Jeffrey's version of decision theory.

Jeffrey presents a decision theory whose objects are propositions instead of gambles, which removes the problem of describing gambles. (Jeffrey 1965) Instead of calculating a subject's degrees of belief and the cardinal utilities based on his or her preference among gambles, Jeffrey's theory uses the subject's preference among propositions. But, just as Ramsey's theory could not be used as-is for Davidson's radical interpretation because of the hidden assumption of knowledge of the subject's language, neither can Jeffrey's theory be used as-is for the same reason: talk about propositions involves semantic notions that the radical interpreter cannot have about her speaker. Furthermore, Jeffrey's theory assumes that both the experimenter and subject understand sentences in the same way – which would mean the experimenter already knows a lot about the subject's language, which, as mentioned above, is knowledge the radical interpreter cannot be assumed to have. For this two-part problem, Davidson proposes a two-part solution, which he describes below.

> As Jeffrey points out, for the purposes of his theory, the objects of these various attitudes could as well be taken to be sentences. If this change is made, we can unify the subject matter of decision theory and theory of interpretation. Jeffrey assumes, of course, that sentences are understood by agent and theory builder in the same way. But the two theories may be united by giving up this assumption. The theory for which we should ultimately strive is one that takes as evidential base preferences between sentences - preferences that one sentence rather than another be true. The theory would then explain individual preferences of this sort by attributing beliefs and values to the agent, and meanings to his words. (Davidson 1974, p. 316)

Let us unpack the above quote to see the two parts of Davidson's solution and what they solve. The first part is to remove propositions and replace them with sentences, which prevents a radical interpreter from assuming any semantic knowledge of her speaker. The second part is to give up the assumption that the sentences (or utterances) are understood in the same way by the radical interpreter and her speaker. Doing this removes the hidden assumption that she already understands his language.

1.3 Merging the two theories

At the end of the last section our attention was focused on how to fix the problems that prevented Jeffrey's theory from being used as-is in Davidson's unified theory and on what Davidson's two-part solution was. While doing this, we may have missed the other benefits that his two-part solution accomplished. Below I repeat the two parts of Davidson's solution, and highlight the other benefits.

The first part of Davidson's solution is to remove propositions and replace them with sentences. And by making this change "we can unify the subject matter of decision theory and theory of interpretation". (Davidson 1974, p. 316) The second part is to give up the assumption "that sentences are understood by agent and theory builder in the same way", which allows "the two theories [to] be united". (Davidson 1974, p. 316) In other words, at this juncture Davidson has created his unified theory.

Now that Davidson has merged his theories of interpretation and of belief, how does he get this dual-theory machine to work? Below is a very brief description of how the different pieces of this machine operate.

- *Evidential base*: The evidence for both theories is preferences between sentences, which is based on sentences that the speaker holds true.
- *Decision theory*: Davidson's decision theory takes as input the preferences the speaker has among sentences, derives the cardinal utilities of the speaker, and from these calculates the speaker's subjective probabilities (beliefs).
- *Theory of interpretation*: Davidson's Tarski-style theory of truth (modified for natural language) takes as input the set of sentences that the radical interpreter has assumed the speaker held true at the time of utterance. The interpreter then tries for an interpretation on this set of sentences by making adjustments so as to maximize the number of utterances that are true (according to her).
- *Logical structure*: Either theory can derive the logical structure. For the theory of interpretation "this may mean reading the logical structure of first-order quantification theory (plus identity) into the language". (Davidson 1984a, p. 136) The decision theory can uncover the logical structure if it has all the logic connectives of the language.[5]

Thus, with the single attitude of a speaker holding a sentence true, Davidson can simultaneously run both of his theories. However, before the theory can

[5]Chapter 2 sketches how we might find the logic connective called the Sheffer stroke from which all the logic connectives can be derived.

run it has to be started, which leads to our next section.

1.4 Jump starting Davidson's dual-theory machine

How does the radical interpreter start this dual-theory machine? First she has to gather a large number of data consisting of the speaker's utterances, information about the environment, and his behavior (both linguistic and otherwise). Then she jump starts this dual-theory machine. How? The short answer is by attributing her beliefs to the speaker. The long answer includes the reason why she can attribute her beliefs to him. The principle of charity consists of two assumptions about the beliefs of a speaker. The first assumption is that his beliefs are consistent in a manner that her beliefs are consistent. The second assumption is that the speaker's beliefs correspond to the real world in a manner similar to how her beliefs correspond to the real world. Because both the speaker's and the interpreter's belief share these two qualities, the interpreter can jump start the dual-theory machine by initially attributing her beliefs to the speaker, seeing how well they fit the collected data, and adjusting the set of beliefs she holds for the speaker to come up with an interpretation (or more likely interpretation*s*), that maximize agreement by making her speaker right as much as possible. In this maximizing process she uses the information given by the decision theory about the speaker's beliefs.

1.5 Literature Review

Knowing what question it was that drove Davidson to create his thought experiment called radical interpretation will help us to evaluate what others wrote about his theory. Davidson himself tells us directly what this question was, for he wrote:

> I want to know what it is about propositional thought – our beliefs, desires, intentions, and speech—that makes them intelligible to others. (Davidson 1995, p. 133)

The reason why he chose to use the thought experiment of radical interpretation was that it could provide philosophical insights, for he says:

> The point of the [Unified Theory[6]] was not to describe how we actually interpret, but to speculate on what it is about thought and language that makes them interpretable. If we can tell a

[6]Davidson expanded and refined his radical interpretation into his Unified Theory, which Chapter 2 goes into great detail describing.

> story like the official story about how it is possible, we can conclude that the constraints the theory places on the attitudes may articulate some of their philosophically significant features. (Davidson 1995, p. 128)

To contrast the question he actually had with the linguistic questions others had mistakenly thought he was interested in, he wrote, "This [question] is a question about the nature of thought and meaning which cannot be answered by discovering neural mechanisms, studying the evolution of the brain, or finding evidence that explains the incredible ease and rapidity with which we come to have a first language."[7] (Davidson 1995, p. 133)

Clearly Davidson was interested in "propositional thoughts" and "what...makes them intelligible to others." Knowing this, we can see that complaints leveled against Davidson that say radical interpretation gives a wrong or inadequate account of certain linguistic phenomena miss the mark. A brief list of such complaints are that radical interpretation fails because:

- It does not give an accurate account about how field linguistics is actually performed. (Chomsky 1992, p. 99, Fodor and Lepore 1994, p.103)
- It cannot account for how children acquire their first language. (Chomsky 1992, p.102, Fodor and Lepore 1994, p.103)
- It assumes to exhaust the evidence available to an interpreter. (Fodor and Lepore 1994, p.105)
- It does not make use of information that is available to linguists. (Chomsky 1992, p.105, Fodor and Lepore 1994, p.105)

Since the basis of these criticism is how radical interpretation fails to adequately explain some linguistic phenomenon, we can safely exclude these and similar articles from our literature review.
In the years after Davidson introduced his conceptual experiment called radical interpretation, only two researches have made a formal representation of this theory, despite the fact that in 1980 Davidson sketched how to do this using the Bayesian decision theories of (Ramsey 1931) and (Jeffrey 1965), and Tarksi's theory of truth (Tarski 1944). The only two documents I found in the literature that give a formal representation of Davidson's radical interpretation are David Lewis' 'Radical Interpretation' (1974) and

[7] In same paragraph Davidson further drove his point home with this example: "Even if we were all born speaking English or Polish, it would be a question how we understand others, and what determines the cognitive contents of our sentences." (Davidson 1995, p. 133)

Marti's dotoral dissertation *Interpreting Linguistic Behavior with Possible World Models* (2016). Other such documents may exist in the literature. If so, they are very rare. The near absence of such documents indicates that a significant gap exists in the literature. Furthermore, one unanswered question has been whether Davidson's radical interpretation could work in the way he described. Davidson has repeatedly argued that it could work. (Davidson 1974, Davidson 1980, Davidson 1995) Obviously, if someone were to design a formal model that ran along the lines that Davidson described, then the mere existence of this model would answer this question in the affirmative. My thesis aims to help create such a formal model.

1.5.1 David Lewis' 'Radical Interpretation' (1974)

David Lewis in 'Radical Interpretation' (Lewis 1974, p. 337) gives a diagram for how to accomplish the goal of radical interpretation. However this diagram is drawn at a very abstract level as can been seen in Figure 1.1. In fact, it is so abstract that not much can be done with it in my thesis.

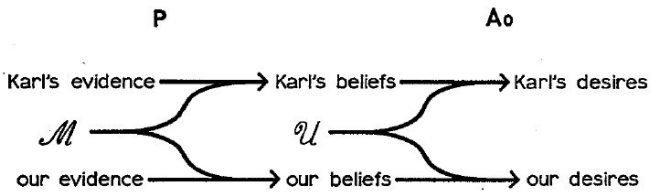


Figure 1.1: David Lewis (1974) diagram for radical interpretation

1.5.2 Marti's *Interpreting Linguistic Behavior with Possible World Models* (2016)

The purpose of Marti's dissertation is "to give an account of when a possible world model represents the beliefs of some subject and the meaning of her sentences that presupposes as little as possible prior knowledge about beliefs and meanings." (Marti 2016, p.10) To do this he uses the linguistic behavior of the subject to derive the beliefs of a subject as well as the meanings of that subject's utterances. His approach is similar that of decision theory that relies on the subject's choice behavior. However, unlike the approach of decision theory where subjective probabilities and the cardinal utilities of the subject can be calculated, Marti's approach can only achieve relative relationships between beliefs and desires. That is, his model cannot give specific numbers for the subject's degrees of belief or how much he desires

certain states of affairs.

To bridge the acceptance of sentences to possible world models Marti assumes the *acceptance principle*, which states, "The subject accepts a sentence if and only if she believes the proposition expressed by the sentence." (Marti 2016, p.10) Marti also defines three requirements listed below so that his model is able "to unambiguously and radically interpret all linguistic behaviors". (Marti 2016, p.16)

1. *Variety*: Every linguistic behavior that some subject might plausibly show should be interpretable. (2016:p.16)

2. *Determinacy*: A linguistic behavior should be interpretable by at most one model. (2016:p.17)

3. *Little-input*: The prior knowledge about the subject that is assumed by the account should be available to a radical interpreter. (2016:p.17)

While Marti's models are constructed using some of the ideas that Davidson used in constructing his radical theory, two of the elements in his model conflict with Davidson's radical interpretation. One is the determinancy requirement which limits the number of models that interpret the linguistic behavior to at most one. This conflicts with Davidson's radical interpretation which allows multiple valid interpretations on sets of utterances. The other is that the little-input requirement seems to imply that the radical interpreter in all of her models has some prior knowledge of her speaker, which like the previous two items stands in stark contrast with Davidson's assumption of no prior knowledge.[8] The reason for pointing out these conflicting items is that the formal model(s) for Davidson's radical interpreter that my thesis proposes to help created aim to be designed in a manner that aligns as much as possible with Davidson's assumptions about his radical interpretation. That is, while Marti has written a formal model that uses some elements of Davidson's radical interpretation, very little of the material can be used in my thesis.

1.5.3 Status of Charity Parts I and II (2006)

In the previous section the radical interpreter jump starts the dual-theory machine by attributing her beliefs to her speaker, which is justified by the assumption of the principle of charity. In 2006 Gluër and Pagin wrote a pair of articles about the status of charity. (Glüer 2006, Pagin 2006) Gluër

[8]Marti may have proposed models where the radical interpreter has no prior knowledge about the subject. However, I did verify that most of his models have a non-empty belief set B that represents the radical interpreter's prior knowledge.

wrote Part I and Pagin Part II. They were investigating the epistemic and metaphysical status of Davidson's principle of charity. That is, is charity a priori or is it a posteriori? And is it metaphysically necessary or not? The answer their investigation suggests is that Davidson's principle of charity is an a posteriori truth of law-like necessity.[9] On the one hand, the conclusions that Gluër, Pagin, or any other researcher come to about whether the principle of charity is justified does not bear on this thesis, because it is built based on what Davidson assumed. Yet on the other hand, the final answer on whether the principle of charity can be justified does bear on the use of any formal model that is created as a result of this thesis. Why does this matter? One, if charity in the end is not justified, then the claims made based on radical interpretation (whether as conceptual experiment in Davidson's case or as a formal model) are also not justified. Two, in Chapter 2 I claim that assumption of the principle of charity weaves itself throughout every sub-theory that Davidson uses to create his Unified Theory (which is an extended and refined version of radical interpretation) and because of this, the results that come from radical interpretation have something to say about us. That is, if the principle of charity is not justified, then neither are the claims about us justified.

1.6 The purpose of my thesis

My thesis presents possible ways to use Bayesian networks to formally represent the different parts of Davidson's unified theory. Then by way of an experiment with an imaginary radical interpreter and her speaker, I demonstrate how GeNIe, a Bayesian network software, can represent the radical interpreter attributing her beliefs to the alien, the belief revision process the interpreter goes through to refine her belief about his beliefs, and a way to derive T-sentences in a natural language equivalence of Tarski's Convention T.

1.7 Map of this paper

Chapter 1 introduces Davidson's radical interpretation, his theory of interpretation, and his theory of belief, explains how he merges these two (sub-)theories and jump starts them. Chapter 2 introduces Davidson's unified theory, which is a refined version of radical interpretation, and discusses

[9] Gluër wrote, "[Our] papers suggest an answer to the question of the epistemic and modal status of Donald Davidson's principle of charity: it is an a posteriori truth of nomological necessity." (2006:p.337)

the theory of belief part and the semantic theory of a language.[10] Chapter 3 introduces Bayesian networks, describes a Bayesian network software called GeNIe, and offers suggestions on how to use both to formally model Davidson's radical interpretation. Chapter 4 describes and gives the results of a number of experiments based on imaginary scenarios that use many of the suggestions offered in Chapter 3. Chapter 5 discusses the results the experiments and suggests possible avenues for further research.

[10]Davidson uses "theory of meaning" but his use is non-standard, so I use "semantic theory of a language" which is more accurate.

Chapter 2

Toward a unified theory

This chapter presents Davidson's expanded and refined version of radical interpretation, his Unified Theory of thought and speech (Davidson 1995, p. 125). The principle of charity is also closely examined and demonstrated to be an integral part of all the sub-theories Davidson uses to create his Unified Theory.

In the years since introducing radical interpretation Davidson wrote many articles on this subject. Four of these articles listed below he included in his book *Inquiries into Truth and Interpretation* (1984). These specific articles, Davidson says were "[addressed] to the question whether a theory of truth for a speaker can be verified without assuming too much of what it sets out to describe." (p. xvi) He also adds, "[A]ll of them, in one way or another, rely on the Principle of Charity." (p. xvii)

- Essay 9. Radical interpretation (1973)
- Essay 10. Belief and the Basis of Meaning (1974)
- Essay 11. Thought and Talk (1975)
- Essay 12. Reply to Foster (1976)

What is the Principle of Charity? It consists of the Principle of Coherence and the Principle of Correspondence.

- *Principle of Coherence*: The assumption that the beliefs of an agent are consistent to an extent in a manner like the interpreter.
- *Principle of Correspondence*: The assumption that the agent has correct beliefs that correspond with the world.

Davidson admits that the above four articles "rely on the Principle of Charity, one way or another". I add a further claim to this: the principles of

coherence and correspondence are integral parts of every sub-theory Davidson uses to create his unified theory. Most of these sub-theories have these principles in their original form and retain these principles after Davidson has modified them. Some of the sub-theories, on the other hand, may not have either principle in their original form (I refer specifically to Tarski's theory of truth), but after Davidson modifies them, they inherit both. As we proceed through the different sub-theories that go into Davidson's unified theory, periodically I will stop and point out how the principle of coherence and the principle of correspondence show up in the sub-theory under discussion.

Why is this important? My answer has two parts. The first relates to justifying the attribution of beliefs by the interpreter to her speaker. To jump start Davidson's unified theory the radical interpreter attributes her beliefs to her speaker. Davidson justifies this by the principle of charity (which is comprised of the principle of coherence and the principle of correspondence). I claim this attribution of beliefs is further justified because the principles in the principle of charity weave themselves through all of the machinery of Davidson's unified theory. If we want to be technical, she does not attribute her beliefs to her speaker, but rather this attribution of her beliefs is applied to the set of utterances along with the environmental facts she has gathered from her observations of her speaker. When framed this way, it is not hard to see how she could think, "My beliefs have something to say about all of this information about the speaker!" The second part of my answer reverses this last sentence: Because the principles of coherence and correspondence weave themselves throughout Davidson's unified theory, the information produced by his theory has something to say about us. What exactly? I do not know. This question will have to be saved for another research project.

The Dutch have a phrase *de rode draad* (literally "the red thread") which is used to refer to a theme, motif, or some other recurring thing that shows up in plays, literary works, and music. The principle of coherence and the principle of correspondence are like two strands of *de rode draad* that weave themselves throughout each sub-theory Davidson uses to create his unified theory. Periodically, as we go through the different sub-theories that go into the unified one, I will stop and point out how we can see these two strands in the sub-theory. In what follows I will briefly describe a theory of another researcher that Davidson uses for his unified theory, then stop and show that the assumption of the principles of coherence and correspondence are integral parts of the this other researcher's theory Davidson borrows from and that these assumptions survive the modifications Davidson applies to this other researcher's theory).

In 1980 Davidson presented his unified theory of meaning and action (Davidson 1980). This theory encompasses radical interpretation, includes more details, and provides a sketch for how to apply this theory to interpret every sentence in a language. Like his earlier theory, Davidson's unified theory is built from two sub-theories. His theory of belief continues to be built on the decision theories of Ramsey (1926) and Jeffrey (1965) and his theory of meaning remains a Tarski-style theory of truth.[1]

2.1 Davidson's unified theory: theory of belief

As mentioned above Davidson borrows from Ramsey's (1926) decision theory. Since Ramsey's theory is based on Bayes's Theorem, we begin with this theorem.

2.1.1 Bayes' Theorem

Bayesian decision theory is based on Bayes' Theorem, (3a), which can be interpreted as (3b), which can be read as follows: Suppose an agent has a degree of belief in A. That is, he assigns a certain probability that A is true. After this agent witnesses evidence B, he updates his belief in A given evidence B by multiplying the likelihood of his previous belief by the likelihood that A is true given that B is true.

(3a) $P(A|B) = P(A)$ x $P(B|A)/P(B)$.

(3b) posterior belief = (prior belief) x (likelihood).

De rode draad: Both the principles of coherence and correspondence are seen in Bayes's theorem. Using Bayes' Theorem to represent an agent's beliefs with subjective probabilities assumes that the agent's beliefs are consistent with the laws of probability. Hence, we have the principle of coherence. The fact that evidence from the real world is used to update an agents belief assumes that the agent has correct beliefs about the real world when he witnesses new evidence[2]. Hence we have the assumption of the principle of correspondence.

2.1.2 Ramsey's decision theory

Davidson says Ramsey (1926) uses "an ingenious trick" that allows him to take ordinal preferences that a subject has among possible gambles, convert

[1] At least since 1967, Davidson's use of the phrase "theory of meaning" is a non-standard one. A more accurate phrase is "a semantic theory for L". In light of this and for the sake of clarity for the rest of this thesis when directly quoting Davidson I will replace his words "theory of meaning" with "[a semantic theory of a language]".

[2] This assumption also includes the idea that the evidence is "real" and not just believed to be real.

these to cardinal utilities, which are used to calculate subjective probabilities, which we can interpret as degrees of beliefs that subject has for certain outcomes. (Davidson 1980, reprinted in Davidson 1984a, p. 156) This trick involves defining an event to which the subject is indifferent.

Using this trick and eight axioms Ramsey defines how to measure the value a subject has for a certain state of affairs. He also defines the degree of belief. With these definitions and axioms in place, Ramsey's decision theory takes as input the preferences a subject has among gambles, derives the subject's cardinal utilities, and calculates the subject's degrees of belief.

De rode draad: In the three quotes below Ramsey makes observations about degrees of belief and people in general that relate to the two strands of de rode draad. Ramsey makes these comments right after demonstrating the fundamental laws of probable belief match the laws of probability. In the first two quotes Ramsey focuses on consistency which ties into the principle of coherence:

1. These are the laws of probability, which we have proved to be necessarily true of any consistent set of degrees of belief. Any definite set of degrees of belief which broke them would be inconsistent in the sense that it violated the laws of preference between options, such as that preferability is a transitive asymmetrical relation, and that if α is preferable to β, β for certain cannot be preferable to α if p, β if not-p. If anyone's mental condition violated these laws, his choice would depend on the precise form in which the options were offered him, which would be absurd. He could have a book made against him by a cunning better [*sic*] and would then stand to lose in any event. (Ramsey 1926, p.182)

2. Having any definite degree of belief implies a certain measure of consistency, namely willingness to bet on a given proposition at the same odds for any stake, the stakes being measured in terms of ultimate values. Having degrees of belief obeying the laws of probability implies a further measure of consistency, namely such a consistency between the odds acceptable on different propositions as shall prevent a book being made against you. (Ramsey 1926, p.182)

In the third quote, Ramsey focuses on the fact that his decision theory is based on betting, which in an experiment involves an experimenter describing to a subject different possible states of affairs on which the subjects make bets. This, I claim, involves the principle of correspondence, because the subjects are expected to have correct beliefs about possible future states of the world, which is unreasonable unless you assume the subject has correct beliefs about the current actual world.

3. Some concluding remarks on this section may not be out of place. First, it is based fundamentally on betting, but this will not seem unreasonable when it is seen that all our lives we are in a sense betting. Whenever we go to the station we are betting that a train will really run, and if we had not a sufficient degree of belief in this we should decline the bet and stay at home. (Ramsey 1926, p.183)

Some of the benefits Davidson's unified theory taken from Ramsey's decision theory are: using one preference relation among some objects and deriving cardinal utilities and subjective probabilities from the preference relationship the subject has among the objects. Unfortunately, Davidson's unified theory cannot use Ramsey's theory as-is for two reasons. Both reasons have to do with the fact that in Ramsey's theory gambles are described using complex sentences. The first problem is what Davidson calls 'the presentation problem' and states "[This] problem is not merely theoretical: it is well known that two descriptions of what the experimenter takes to be the same option may elicit quite different responses from a subject." An example of a presentation problem is the framing effect, which is seen by the different responses given to the following two problems from a study by Tversky and Kahneman (1981).[3]

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume the exact scientific estimate of the consequences of the programs are as follows:
>
> - Problem 1 [N = 152]:
> - If Program A is adopted, 200 people will be saved. [72 percent]
> - If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. [28 percent]
> - Which program would you favor?
>
> - Problem 2 [N = 155]:
> - If Program C is adopted 400 people will die. [22 percent]
> - If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. [78 percent]
> - Which program would you favor?

[3]Note: '[N = 152]' means the number of respondents is 152. The bracketed percent such as '[72 percent]' indicates what percentage of the respondents voted for that program.

Tversky and Kahneman summarize the results of this experiment as follows:

> The majority choice in this problem is risk averse: the prospect of certainly saving 200 lives is more attractive than a risky prospect of equal expected value, that is, a one-in-three chance of saving 600 lives. [...] The majority choice in Problem 2 is risk taking: the certainty of death of 400 people is less acceptable than the two-in-three chance that 600 will die. The preferences in problems 1 and 2 illustrate a common pattern: choices involving gains are often risk averse and choices involving losses are often risk taking. (Tversky and Kahneman 1981, p.453)

Even if the presentation problem were to be solved among decision theorists, Ramsey's theory would still not be suitable for Davidson's unified theory, because of a second, more fundamental, problem. Davidson's unified theory is supposed to provide a theoretical framework within which his radical interpreter can discover what her speaker believes so that she can use this information to help in the process of interpreting what her speaker says. That is, she starts with no knowledge about the speaker's language and part of what she wants to do is to figure out some of the speaker's beliefs to help her interpret his utterances. The problem with presenting gambles using complex sentences is that we have to be quite far along in understanding the subject's language. To solve Davidson's gambling problem (ahem) he turns to the decision theory by Jeffrey (1965) (see below). However, one aspect about Ramsey's theory that Davidson keeps is preference a subject has among choices.

2.1.3 Holding true and preferring true

Davidson says, "The interdependence of belief and meaning is evident in this way: a speaker *holds a sentence to be true* because of what the sentence (in his language) means, and because of what he believes." (Davidson 1973, reprinted in Davidson 1984a, p. 133 (Italics mine)) And on this idea of holding-true, Davidson also says:

> A good place to begin is with the attitude of holding a sentence true, of accepting it as true. This is, of course, a belief, but it is a single attitude applicable to all sentences, and so does not ask us to be able to make finely discriminated distinctions among beliefs. It is an attitude an interpreter may plausibly be taken to be able to identify before he can interpret, since he may know that a person intends to express a truth in uttering a sentence without having any idea *what* truth. (Davidson 1973, reprinted in Davidson 1984a, p. 134 (Italics by Davidson))

Davidson (1973) advocates using the attitude of a speaker of holding a sentence true. Davidson (1980), however, extends and refines his original theory to using the speaker's preferring-true among sentences to not only derive what the speaker means by his utterances, but also what values he places on possible states of the world and what the speaker believes. Later in this chapter we will look at how Davidson uses a speaker holding an utterance true to gain insights into what this utterance might mean. Now, however let us return to the theory of beliefs of Davidson's unified theory.

2.1.4 Jeffrey's decision theory

Recall the primary reason Davidson could not use Ramsey's decision theory as-is in his unified theory was because gambles are described in terms of complex sentences, which has had the historical problem of conflicting choices by the same subjects to the same gamble described differently, and which has a deeper problem of assuming the experimenter understands the subject's language. Fortunately, Davidson has a solution. He says, "[Jeffrey's decision theory] eliminates some troublesome confusions in Ramsey's theory by reducing the rather murky ontology of the theory, which dealt with events, options, and propositions to an ontology of propositions only." (Davidson 1974, p.316) This means that "[p]references between propositions holding true then becomes the evidential base, so that the revised theory allows us to talk of degrees of belief in the truth of propositions, and the relative strength of desires that propositions be true." (Davidson 1974, p.316) Davidson extends the hold-true attitude which is only applicable to a single sentence or proposition to a preferring-true relationship among many sentence. With this, "Jeffrey has shown in detail how to extract subjective probabilities and values from preferences that propositions be true." (Davidson 1980, reprinted in Davidson 1984a, p. 160)

However, like Ramsey's decision theory, Jeffrey's version cannot be used as-is for Davidson's unified theory for similar reasons: the objects of the preference relationship are incompatible with assumptions in the unified theory. For propositions to be used in the preference relationship of the speaker, the radical interpreter will have to know a significant amount of semantics of her speaker, which is excluded by the assumption she knows nothing about his language. Another problem with Jeffrey's decision theory is that if it were used in the unified theory, then it would be assumed that the utterances of the speaker are understood the same way by the speaker and radical interpreter. Again, this is excluded by the assumption that she begins by knowing nothing about him or his language. For this two-part problem, Davidson proposes a two-part solution, which he describes as follows:

> As Jeffrey points out, for the purposes of his theory, the objects of these various attitudes could as well be taken to be sentences. If this change is made, we can unify the subject matter of decision theory and theory of interpretation. Jeffrey assumes, of course, that sentences are understood by agent and theory builder in the same way. But the two theories may be united by giving up this assumption. The theory for which we should ultimately strive is one that takes as evidential base preferences between sentences - preferences that one sentence rather than another be true. The theory would then explain individual preferences of this sort by attributing beliefs and values to the agent, and meanings to his words. (Davidson 1974, p. 316)

Let us unpack the above quote to see the two parts of Davidson's solution and what they solve. The first part is to remove propositions and replace them with sentences, which removes the hidden assumption that the radical interpreter knows a significant amount of semantics of her speaker. The second part is to give up the assumption that the sentences (or utterances) are understood in the same way by the radical interpreter and her speaker. Doing this removes the hidden assumption that she already understands his language.

De rode draad: The principle of coherence is present in Jeffrey's decision theory, because this theory is also built on the same assumption that Ramsey's theory is about the degrees of belief of a subject being consistent and in accord with probability theory. (Jeffrey 1965, p.49)

To show that the principle of correspondence is also in Jeffrey's theory is more involved. First, the bad news. In one sense Jeffrey's theory does not have any correspondence to the actual world. Davidson (1980) describes how to use the modified version of Jeffrey's theory to calculate degrees of beliefs and cardinal utilities of a speaker. Yet after this process is done Davidson says, "At this point the probabilities and desirabilities of all sentences have in theory been determined. But no complete sentence has yet been interpreted, though the truth-functional sentential connectives have been identified, and so sentences logically true or false by virtue of sentential logic can be recognized." In other words it is theoretically possible for someone to construct various "sentences" by stringing together random series of symbols and assigning preferences among these sentences to an imaginary speaker, and derive degrees of belief and cardinal utilities via Jeffrey's theory – all without having any connections to the real world.

Second, the good news. Jeffrey designed this theory to be used with real agents in our real world. We see this by the following excerpt from Jeffrey

(1965, p.172):

> We shall now consider cases in which the agent's belief function changes from *prob* to $prob_B$ as the result of an observation; where the agent's conclusive belief in B is caused by the observation; is unreasoned; and is justified by the consideration that the observation is of the paradigmatic sort which any normal speaker of the language in which B is expressed would respond by believing B, willy-nilly. (Jeffrey 1965, p.172)

The expression $prob_B$ refers to the Bayesian update based on evidence with B representing this evidence. This update is done using Bayes' theorem, which I argued does have the principle of correspondence built into it (under a Bayesian interpretation). Furthermore, Jeffrey's description above shows that not only is this an observation made in the real world, it is a true one since "any normal speaker of the language...would respond by believing B...". Hence, I claim that Jeffrey's decision theory does have the principle of correspondence built into it. Also, the fact that Davidson replaces propositions in Jeffrey's theory with sentences further strengthens the ties of this theory to the real world, since many of these sentences will be uttered by the speaker in response to changes in the world around him.

For now I will pause on discussing Davidson's theory of belief so we can turn our attention to his semantic theory of a language. Later, we will see Davidson's decision theory when he merges it with his semantic theory of a language.

2.2 Davidson's unified theory: semantic theory of a language

Recall Davidson's unified theory needs to provide a theory of belief and a [semantic theory of a language][4] that concurrently derive beliefs of a speaker as well as interpretations of his utterances. Davidson (1973, reprinted in Davidson 1984a, pp. 126-9) stipulates that a semantic theory of a language has to do the following:

(a) It has to provide the radical interpreter the resources to understand any sentence out of the "infinity of sentences the speaker might utter" while the theory at the same time has to be finite in form.

(b) It must "be supported or verified by evidence plausibly available to an interpreter."

[4]Recall that where Davidson uses the phrase "theory of meaning" I write "[semantic theory of a language]" for more accuracy.

(c) It must reveal significant semantic structure, e.g., "the interpretations of utterances of complex sentences will systematically depend on the interpretation of utterances of simpler sentences".

Davidson (1973, reprinted in Davidson 1984a, p. 129) claims "We have such theories, I suggest, in theories of truth of the kind Tarski first showed how to give." A theory of truth in a Tarski style entails for every sentence s in the object language a T-sentence of the form:

1. s is true (in the object language) if and only if p.

For Tarski instances of these T-sentences are created by replacing s with a canonical description of s and 'p' by a translation of p. Underlying this definition is the undefined notion of *satisfaction* which relates each sentence to an infinite number of objects in the object language. This notion of satisfaction is seen in Tarski's definition of the truth predicate:

- For all x, Tr(x) if and only if x is a sentence of LCC and every infinite sequence of subclasses satisfies x. (Tarski 1983b)

As with Ramsey's theory, and Jeffrey's, Davidson has to modify Tarski's theory in two ways to make it amenable for using it in his unified theory. First, Tarski formulated his semantic notion of truth for formal languages which do not have indexicals such as "I", "we", or "you", and which do not have demonstratives such as "this" or "that". Davidson's solution is simple: "The remedy is to characterize truth for a language relative to a time and a speaker. The extension to utterances is again straightforward." (Davidson 1973, reprinted in Davidson 1984a, p. 130) The second modification is reversing how Convention T is used. Tarski assumed we had the meaning of a sentence (that is, we can translate p) and from this Tarski defined truth. Davidson does the opposite, because he assumes the truth of a sentence (by assuming we know that a speaker holds this sentence true) and derives "the canonical description of s." (Davidson 1973, reprinted in Davidson 1984a, p. 129) Suppose, "The interpreter, on noticing that the agent regularly assigns a high or low degree of belief to the sentence 'The coffee is ready' when the coffee is, or isn't, ready will...try for a theory of truth that says that an utterance by the agent of the sentence 'The coffee is ready' is true if and only if the coffee is ready." (Davidson 1980, reprinted in Davidson 1984a, p. 165)

De rode draad: My answer to the question of whether Tarski's theory of truth has the principle of coherence in it is half-yes and half-no. Half-yes: A Tarski-style theory of truth assumes some logical structure because "the meaning of a sentence depends on the meaning of its part", which means

consistency runs throughout the theory and is assumed.[5] Half-no: Tarski's theory of truth was created for formal languages and therefore do not require a human subject, agent, or speaker. Therefore, the belief side of the principle of coherence is not guaranteed. So Convention T does assume consistency, but need not assume anything about the beliefs of a person.

Does a Tarski-like theory of truth assume the principle of correspondence? Davidson's answer is mixed. Davidson (1967) argues that Tarski's Convention T was a correspondence theory. Davidson (1990, p.304) goes into great detail of why "There is thus a serious reason to regret having said that a Tarski-style truth theory was a form of correspondence theory." Others have said Davidson is wrong in claiming it is not a correspondence theory. Fortunately, neither you nor I have to decide on this issue now. Why? Because the modifications that Davidson makes to Tarski's theory of truth make the new version inherit both principles of charity.

To accommodate the fact that natural language is replete with indexical features Davidson modifies Tarski's Convention T "to characterize truth for a language relative to a time and a speaker." (Davidson 1973, reprinted in Davidson 1984a, p. 130) This change brings with it both the assumption of the principle of coherence (because the speaker is holding a sentence true and consistency is assumed by Convention T) and the principle of correspondence (since the semantic content of an utterance held true depends on the events in the speaker's environment that influenced him to make his utterance).

2.3 Conclusion

The above sections presented Davidson's unified theory which is a refined version of his radical interpretation. What this thesis proposes to do is to offer suggestions on how to create a formal representation of Davidson's radical interpretation using the assumptions he made as well as the methods he proposes. So in a sense, this thesis aims to keep both the "spirit" of Davidson's radical interpretation as well as the "law" of it. That is, not only do we want to do things like he said, but also exactly what he said to do (as far as possible).

[5] Part of the definition of a theory within model theory is that it is built from a set of sentences that are *consistent*.

Chapter 3

Toward formal models

This chapter discusses in more technical details Bayesian networks, Bayesian network software, and possible ways to use these theories and tools to formally represent different parts of Davidson's unified theory.

3.1 Bayesian Networks

3.1.1 Math, probability, and graphs

Bayes' Theorem

Below are the the basic probability axioms from which Bayes' theorem can be derived.[1]

Axiom 1. $0 \leq P(A) \leq 1$, with $P(A) = 1$ if A is certain.

Axiom 2. If events $(A_i)(i = 1, 2, \ldots)$ are pairwise incompatible, then $P(\cup_i A_i) = \sum_i P(A_i)$.

Axiom 3. $P(A \cap B) = P(B|A)P(A)$.

Bayes' Theorem can be derived as follows. From Axiom 3 and $P(A \cap B) = P(B \cap A)$, we have 3.1, from which we can derive 3.2.[2]

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A). \tag{3.1}$$

[1] Axiom 3 uses both unconditional and conditional probabilities, whereas the standard account which follows Kolmorogov (1950) takes the unconditional probability as primitive. Cowell et al. (1999, p.13) states, "any 'unconditional' probability is only really so by appearance, the background information behind its assessment having been implicitly assumed and omitted from the notation", which this thesis follows.

[2] This succinct proof comes from Cowell et al. (1999, p.14). About Axiom 2 Cowell says, "There is continuing discussion over whether the union in Axiom 2 should be restricted to finite, rather than countably infinite, collections of events (Finetti 1975). For our purposes this makes little difference, and for convenience we shall assume full countable additivity." (Cowell 1999, p.14)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3.2}$$

Suppose we begin by assigning a probability to A, giving us $P(A)$, which is our *prior probability*, which represent our belief that A will occur. Then suppose we witness B which is now evidence. We then calculate our *posterior probability* $P(A|B)$ which is our revised belief about A. This calculation is done by multiplying $P(A)$ by $\frac{P(B|A)}{P(B)}$.

> This can be interpreted as follows. Suppose we are interested in A and we begin with a prior probability $P(A)$, representing our belief about A before observing any relevant evidence. Suppose we then observe B. By (2.2), our revised belief for A, the posterior probability $P(A|B)$, is obtained by multiplying the prior probability $P(A)$ by the ratio $P(B|A)/P(B)$. (Cowell 1999, p.14)

3.1.2 Bayesian networks

A Bayesian network is a probabilistic model that uses Bayes' theorem to revise the values of its probabilities after some probability values have changed in it. This model consists of a structure and its parameters. The structure is determined by the conditional probabilities among the variables and usually is represented as a directed acyclical graph (DAG) such that for every node i in the graph a variable X_i is assign to it. If the probability of a variable X_i is conditioned on other variables, then the DAG contains arrows from the nodes of these other variables (called parent nodes of X_i) to the node for X_i. The parameters for a Bayesian network are the values of the probabilities. The joint probability distribution of the Bayesian network is the product of the conditional probability distributions (see Equation 3.3).

$$P(x_1, \ldots, x_n) = \prod_i^n P(x_i|x_{(\pi_i)}). \tag{3.3}$$

Below are two examples showing how to go from a particular factored form of a joint probability among four variables to its DAG.

1. Supposed we had (a) for the factorization of a joint probability that had four variables. To draw the directed acyclical graph (DAG), which is the structure for this Bayesian network, we draw arrows from every variable listed on the right of the conditional bar (that is, variables Y, Z, and W) to the variable on the left of the conditional bar (that is, X) as seen in Figure 3.1.

(a) $P(X, Y, Z, W) = P(X|Y, Z, W)\ P(Y)\ P(Z)\ P(W)$

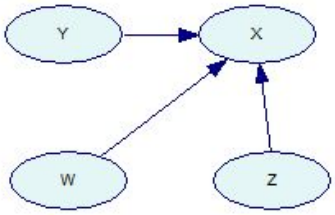


Figure 3.1: Example of a Bayesian network

1. However if the factorization was (b), we would follow the same procedure of drawing an arrow from the node that matches every variable on the right of a conditional bar to the node that matches the variable on the left of that conditional bar as seen in Figure 3.2.

(b) $P(X, Y, Z, W) = P(X|Y, Z, W)\ P(Y|Z, W)\ P(Z|W)\ P(W)$

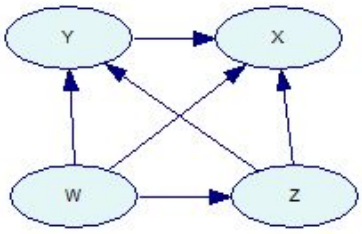


Figure 3.2: More complicated example of a Bayesian network

When probabilities of nodes change, the Bayesian network updates the other nodes using Bayes' theorem. If Bayes' theorem is interpreted as a belief an agent has, then the network is a (Bayesian) belief network.

Reducing complexity

In addition to determining a structure and set of parameters for a Bayesian network, the designer has to address the problem of complexity. The marginal values[3] of a variable x_i can be calculated from 3.4. However, the calculations can become intractable since adding a variable increased the complexity exponentially. To calculate the unconstrained joint probability distribution for n binary variables requires $O(2^n)$ probabilities. For example suppose there are 30 binary variables, then calculating the joint distribution would be 2^{30} (over one billion), which is intractable.

[3]Also known as expected values.

$$P(x_i) = \sum_{1} \cdots \sum_{i-1} \sum_{i+1} \cdots \sum_{N} P(x_1, \ldots, x_N) \tag{3.4}$$

However, the property of conditional independence greatly reduces the number of calculations needed. First conditional independence will be explained, then the above example will be revisited.

Two events A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$. That is, C is known, then knowledge that A has occurred provides no information on the likelihood of B occurring and vice versa. In a Bayesian network this means the conditional probability of a variable depends only on the variables that are parent to the variable. So Equation 3.4 above reduces to where $x_{(\pi_i)}$ are the variables that are direct parents of x_i.

$$P(x_i) = P(x_i|x_{(\pi_i)}). \tag{3.5}$$

Let us return to the example with 30 binary variables. Suppose that each of these variables has at most 4 parent variables, then calculating the joint probability distribution involves only 480 probabilities.

However, despite the complexity reducing benefits of using conditional independence, the computational cost of Bayesian networks is NP-hard because adding a node causes the complexity of the graph to grow exponentially. Furthermore, one probability distribution $P(x_1, \ldots, x_n)$ can have multiple Bayesian networks that fit it. Therefore, one of the goals for designing a Bayesian network is to find DAGs for the network that are close to the simplest possible for the problem to be modeled. Finding a suitable simple DAG can be accomplished in a manner similar to building a Bayesian network. One strategy on the expert information side is to leverage causal relationships that exists among some of the variables. The reasoning is as follows: if certain events (read: variables or nodes) are in fact causally related in the "real world", then the simplest way (graph-wise) to represent these relationships in a belief network is to connect the nodes based on the causal relationship. Note: we can use *causal relationships* between events/nodes/variables to help create a simpler or even one of the simplest belief network, but we cannot make *causal inferences* from the belief network (unless it is a causal network (see Pearl 2000) for more information). A further benefit from using causal relationships to design a belief network is that the resulting DAGs are more meaningful to the modeler (Pearl 2000).[4]

[4]Causal networks are not appropriate for the project of this thesis because such networks model ontological objects, and the objects for the Bayesian belief network of this thesis are epistemological. Also, the causal networks by Pearl (2000) require more structure, namely exogenous variables, and an intervention process called *doing*. Robert van

Possible error correcting benefits

When someone wishes to design a Bayesian network, she has a two basic options that can be combined. One option is she can gather information for the structure and/or the probabilities from experts who are familiar with the problem she wants to model. A second option is to use algorithms to analyze data to derive possible structures, parameters, or both. She can also combine these two options. Bayesian networks that were designed using information both from experts and data provide a three ways of error correction. One comes directly from using Bayes' theorem by way of the updating process. An agent can begin with a wrong probability, but by repeatedly updating his beliefs based on evidence, the agent can derive the (or a) correct probability. Similarly, a Bayesian network can begin with wrong probabilities and update itself to correct ones. The second possible correction comes from the expert. For example, it is possible that a structure learning algorithm could be given data that contains information about when an agent chose to take an umbrella, and the algorithm derive a DAG like Umbrella $\rightarrow$ Rain, which means the choice to take an umbrella affects the chance of rain. In this situation, information from the expert can prevent the algorithm from considering this possibility. The third possible correction is an algorithm correcting wrong information from the expert. That is, if an expert provides wrong information, some algorithms can override these restrictions based on the strength of the data. An additional benefit of combining expert and data information when designing a Bayesian network is that one may have missing information that the other can provide.

3.2 Software for Bayesian networks

This section introduces GeNIe, which is a Bayesian network software by BayesFusion, LLC. "GeNIe Modeler is a development environment for building graphical decision theoretic models. It was created and developed at the Decision Systems Laboratory, University of Pittsburgh between 1995 and 2015." (Version 2.4.R1, Built on 8/5/2019, BayesFusion, LLC, p.32). While the main idea behind this thesis is to suggest ways to use *a* Bayesian network software to formally represent Davidson's radical interpretation, for ease of exposition I will refer to GeNIe, with the understanding that other Bayesian network softwares may work as well or even better. Furthermore, since the experiments in the next chapter primarily deal with only the structure of these Bayesian belief networks (that is, either the DAGs or graphs) only three features and two algorithms will be described, primarily because

Rooij (p.c,) also adds 'In a causal network, instead, the link are determinate, and the probabilities come in via probabilities over (extra) exogenous variables. According to Pearl, determinate links are much more stable than the probabilistic links of Bayesian networks."

with these few tools many parts of Davidson's radical interpretation can be modeled (as will be seen in the next chapter).

- *Data*: GeNIe can load data files that are in Comma Separated Values formatting. GeNIe converts each column of the file into a node in the graph with the name of the node corresponding to the entry at the top of the column. The values of the node can be discrete or continuous or the node can be a set of states with names (e.g., Low, Medium, High).[5] A data file can contain over 1000 variables.

- *Background Knowledge*: Expert knowledge about the structure of the system being modeled can be entered in the Background Knowledge feature. GeNIe uses this information, which is entered in the form of a graph (with or without directed arrows) as a guide for what structures to include or exclude while deriving the set of graphs that fit the data. Note, however, if the strength of that is strong enough, the structure learning algorithm can override the graph structure that was entered.

- *Structure learning algorithms*: GeNIe has the feature Learn New Network in which the user can specify which kind of algorithm GeNIe should use when learning the structure of the data. The PC algorithm and Greedy Thick Thinning algorithms are described below. When the Learn New Network feature is used the algorithm derives from the data a set of possible graphs or DAGs that fit the data.

Algorithms for Structure learning

Suppose you have a set of data for the problem that you want to build a Bayesian network. The structure for the Bayesian network can be learned from the data. One method is to use *score-based structured learning* which is to find the DAG that best fits the data based on some score function. (Scanagatta, Salmerón, and Stella 2019, p.427) Another method is *constraint-based* structure learning that involves algorithms using statistical methods to calculate conditional hypothesis tests to determine what the independent variables are in the data. (Ibid., p.429) Based on these constraints DAGs are built.

Greedy Thick Thinning: "The Greedy Thick Thinning (GTT) structure learning algorithm is based on the Bayesian Search approach [where] GTT starts with an empty graph and repeatedly adds the arc (without creating a cycle) that maximally increases the marginal likelihood $P(D|S)$ until no

[5] All the structure learning algorithms in GeNIe can work with discrete values. Only the PC algorithm work with data that has both discrete and continuous variables. For all the other algorithms, the continuous data has to be converted into a discrete form (and GeNIe has a feature that can do this).

arc addition will result in a positive increase (this is the thickening phase). Then, it repeatedly removes arcs until no arc deletion will result in a positive increase in $P(D|S)$ (this is the thinning phase)." This algorithm uses a score-based one. (Version 2.4.R1, Built on 8/5/2019, BayesFusion, LLC, p.252)

PC algorithm: Scanagatta (2019, p.429) states, "The state-of-the-art approach is the PC algorithm (named after its authors, Peter and Clark)". The PC algorithm assumes the worse case scenario – that is, it begins with a complete, undirected graph – then recursively performs the tests on the data to determine which edges to delete. The end result is a set of undirected or partially directed graphs which the modeler of the network can choose from. If the actual underlying DAG is sparse, the run time is polynomial. The PC algorithm is a constraint-based one.

For the reader interested in learning more about other Bayesian network softwares, Scanagatta (2019) provides a list.

3.3 Using GeNIe to formally represent radical interpretation

As previously stated I will present these suggestions in terms of GeNIe, with the understanding that other Bayesian network softwares may work as well or better. Also two lists of suggestions will be given. One has suggestions that have actually worked in at least one of the experiments described in the next chapter. The other list has suggestions for possible ways to do parts of Davidson's radical interpretation, but have not been confirmed to work. With all of these suggestions I will say *what* to do, say a little about *why*, and not much about *how* to implement the given suggestion. The reason for presenting the suggestions this way is that a lot of details are given for the experiments in the next chapter. That is, a lot of *how*'s are given there.

- Suggestions that have been verified

    1. *DAGs*: Use DAGs to represent beliefs, because they are the structure of a belief networks and can represent what an agent believes about how his or her world runs.
    2. *Data*: Use data files to represent the set of observations the radical interpreter has made of her speaker.
    3. *Background Knowledge*: Enter the DAG of the belief network for the radical interpreter into the Background Knowledge feature to simulate her attributing her beliefs to her speaker.

4. *Learn New Network*: Run the Learn New Network on the data file mentioned in 1 to derive the set of graphs that fit the data, then do the following:
   (a) Verify whether the attributed DAG is an element of the set of possible graphs that fit the data. This corresponds to the case that the radical interpreter beliefs are similar enough to make sense of all the speaker data she gleaned from her observations. If not, got to 4b.
   (b) Use a measurement function to determine which graphs in the set of possible graphs she disbelieves the least, and choose this graph(s) as the one to attribute to her speaker.[6]
5. *Restructure data file*: This suggestion is hard to describe without a concrete example, which Experiment 2 and 3 of the next chapter provide us. In light of this I will forgo explaining the what to do in GeNIe. However, I will say that this suggestion derives T-sentences from the speaker data.

- Unverified suggestions

1. *Restructure data file (Part II)*: This suggestion proposes using the same procedure as the previous suggestion, but has not been verified. But, using the procedure it might be possible to derive T-sentences for more theoretical sentences. That is, for sentences whose holding-true value depends solely on other sentences and not states or events in the environment.
2. *Restructure data file (Part III)*: Same set up procedure as before. This should be able to derive the T-sentences for complex-theoretical sentences. That is, sentences that are composed of other sentences using AND, OR, NOT, or IF-THEN. Again, this should derive T-sentences for these utterances.
3. *Restructuring data file (Part IV)*: Same set up as before. This suggestion has not been fully worked out conceptually like the previous two. This *might* provide a way for deriving a preference relationship for the speaker among his utterances. The motivating idea is that Ramsey (1926) and Jeffrey (1965) used a proposition and its negation to form complex gambles or propositions, respectively, to determine the preferences a subject had among choices.

[6]The fact that her DAG is not in the set of possible graphs means she believes something that with respect to the speaker data she believes something false. That is, for her to ask which one she believes the most would be a contradiction. Thus, we reverse the question: which set of beliefs would she disbelieve the least.

Again the above are suggestions on how to use a Bayesian network software like GeNIe to formally represent Davidson's radical interpretation. One apparent discrepancy needs to be pointed out and explained. Davidson repeatedly recommended using *Bayesian decision theory* while all of my recommendations involve *Bayesian networks*. So where did the decision theory go? It is still present, because in the the experiments of the next chapter I frequently have to step in and either take information generated by GeNIe and analyse it or simply put it back into GeNIe for it to do the analysis. For example, I check whether the DAG that the radical interpreter attributed to her speaker (by entering this DAG into Background Knowledge) is contained in the set of possible graphs the PC algorithm derived from the data. This is a *decision*. At other times I needed to determine which of the possible graphs were the best to select for the belief revision. Again this is a *decision*. The places where I stepped in and manually did something with or to GeNIe, some form of *decision(s)* was made. It may be possible to have GeNIe make all of these decisions with a Bayesian decision theory network called an influence diagram or influence network.

Interested readers can look up influence diagrams and how they are used.

Chapter 4

The Experiments

This chapter describes experiments based on an imaginary scenario to examine how much and how well a Bayesian network software[1] can model different parts of Davidson's unified theory based on the scenario. The beliefs of the radical interpreter and of the speaker in this scenario differ significantly in one respect.

The focus of the first set of experiments was on beliefs. The primary questions for these experiments were whether GeNIe can model (1), (2), and (3). These questions are primarily about the structural differences among Bayesian networks that represent beliefs of agents and not about the particular numerical values within the structures. For this reason when discussing various Bayesian networks usually only the graphs are shown. That is, generally the associated the conditional probability tables are not shown.[2]

1. The radical interpreter attributes her beliefs to her speaker.
2. The radical interpreter discovers the beliefs she attributed do not fit the speaker's beliefs.
3. The radical interpreter correctly discovers the speaker's beliefs.

The focus of the second set of experiments was on deriving information on the speaker's utterances. The primary questions were whether it is possible to accomplish (1), (2), and (3):

1. Correctly derive the different conditions under which various utterances were held true by the speaker?[3]

[1] Specifically, the Bayesian network software *GeNIe* by BayesFusion LLC.

[2] However, when investigating a possible way to form T-sentences using GeNIe, conditional probability tables are examined various conditional probability tables. These tables will be listed in this section.

[3] Another way to say this is: Can GeNIe correctly derive T-sentences for most or all of the speaker's utterances?

2. Ignore random noise in the data.
3. Derive a preference order among the speaker's utterances using information from GeNIe.

Based on the results from the first set of experiments, I claim the following was accomplished:

- GeNIe successfully modeled for the following parts of Davidson's unified theory:
  - The beliefs agents have about causal relationships in their environment.
- GeNIe successfully modeled the radical interpreter:
  - Attributing her beliefs to her speaker.
  - Discovering that these attributed beliefs did not match her speaker's.
  - Discovering the speaker's beliefs.

Based on the results from the second set of experiments, I claim the following was accomplished:

- GeNIe successfully did the following:
  - Correctly derived the conditions under which the speaker held specific utterances true.
  - Ignored random information in the data.

Another possible result is that a method *may* have been discovered for how to calculate a preference order. However, these results may be due to specific elements of the imaginary scenario. These results may not be generalized.

4.1 Scenario behind Experiment 1

This section provides the scenario upon which Experiment 1A and 1B were based. The underlying idea is to have a problem that has to be fixed before correct possible interpretations can be derived. The problem is that the beliefs of the radical interpreter and of the speaker differ significantly about the effect low air pressure has on the chance of rain. No utterances are involved in Experiment 1A or 1B. Utterances are included in Experiment 2.

First, a fact about air pressure and the chance of rain. In the northern region of the United States low air pressure causes the chance of rain to rise significantly, but in other regions (such as Florida), the air pressure has no measurable effect on the chance of rain.

Backstory: A radical interpreter grew up in a region of the world where air pressure has no noticeable effect on the chance of rain. On the other hand, the speaker grew up in a region where low air pressure raises the chance of rain to 60%. Recently she moved to his region and has not discovered that her beliefs about air pressure and rain do not match what happens in her new location. She wants to learn the speaker's language. Our radical interpreter under the principle of charity has attributed her beliefs her speaker and now observes him to discover whether she needs to revise her belief about his beliefs.

The daily scenario: Every, morning unbeknownst to the speaker, the radical interpreter observes him as he does his morning ritual before leaving. First, he examines a barometer. Second, he looks out the window. And third, if he believes it probably will rain, he takes an umbrella; if not, then no umbrella is taken. Then he leaves. Our interpreter observes and records this daily routine 1000 times. Then she analyzes the data, revises her attributed beliefs if needed, then calculates his subjective beliefs and his preferences.

4.2 Formally representing the scenario

Beliefs: Since the only difference between their beliefs is about how low air pressure does or does not increase the chance of rain, the beliefs of the radical interpreter and her speaker are (1) and (2).

1. $B_{interpreter} = \{P(\text{rain}|\text{air pressure low}) = P(\text{rain}|\text{air pressure not low})\}$.
2. $B_{speaker} = \{P(\text{rain}|\text{air pressure low}) = 60\%\}$.

Technically, if our network only represents the beliefs of an agent, then the umbrella node should not be part of the network, because even if it was raining extremely hard, the speaker could choose to leave his umbrella behind. However, for the sake of simplicity the umbrella node is included with arrows from the rain and air pressure nodes, based on the fact that any time the speaker believes it probably will rain, he takes an umbrella. This means that the arrows from the air pressure and rain nodes to the umbrella one imply a Bayesian decision theory network (in fact it is a hidden influence diagram). For similar reasons the Bayesian networks that represent the beliefs of the radical interpreter and the one she has for her speaker also have the umbrella node with arrows going to it from other parent nodes.

Actions: The set of actions consist of only one element $A = \{a_1\}$, where a_1 is "take an umbrella".

4.3 Experiment 1

Experiment 1 consists of two parts. Experiment 1A has a belief network for the radical interpreter that structured the nodes in a temporal sequential manner (specifically, the air pressure node came temporally before the one for rain which came before the one for umbrella). Experiment 1B removes this temporal restriction.

4.3.1 Experiment 1A

Representing beliefs

The Bayesian network that represents the radical interpreter's and her speaker's beliefs in Experiment 1 are shown Figures 4.1 and 4.2 and are explained in detail below.

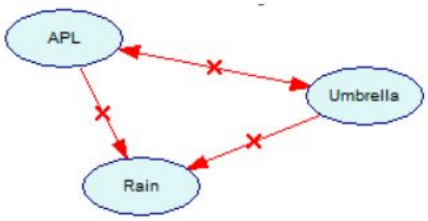


Figure 4.1: DAG of interpreter's beliefs

The Bayesian network for the radical interpreter is shown below by way of her DAG seen in Figure 4.1. The red arrow from *APL* to *Rain* indicates her belief that low air pressure has no affect on the chance of rain. The other two red arrows depict some common sense knowledge about the world. The double-headed red arrow with an X in the middle indicates that *APL* and *Umbrella* are probabilistically independent of each other (and therefore have no causal effect on each other). The red arrow from *Umbrella* to *Rain* indicates she believes no probabilistic (and therefore no causal) relationship exists between whether someone takes an umbrella and the chance of rain.

The Bayesian network for the speaker is shown in Figure 4.2. In contrast to the one for the radical interpreter, the network for the speaker has no red arrows that forbid probabilistic relationships, this diagram has three solid blue arrows, each indicating a probabilistic relationship. The blue arrow from APL to Rain indicates that a change in the APL node (say, the air

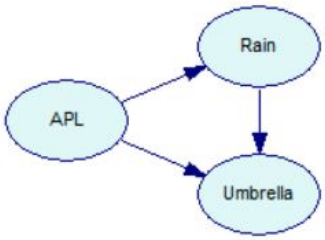


Figure 4.2: DAG of the speaker's beliefs

pressure gets low) can cause a change in the Rain node (say, the chance of rain increases significantly). Incidentally, if the state of the APL node is not low (say, it is high), then this causes no change to the Rain node. That is, just because an arrow goes from one node to another does not mean every time the value or state changes in the parent node, a change will happen in the child node. Similarly a blue arrow goes from the Rain node to the Umbrella one. This fits our scenario because if it is raining at the time the speaker leaves his house, he will take an umbrella.

Attributing beliefs

To formally represent the radical interpreter attributing her beliefs to her speaker, the following was done in GeNIe. Under the Data menu is the *Learn New Network* option. When this option is selected this brings up a dialogue box in which the user can choose which algorithm will be used to learn the structure of the Bayesian network from the data. In all of the experiments in this thesis, the structure that GeNIe tries to learn from the data is the belief network for the speaker. This dialogue box also has *Background Knowledge* which is a feature that allows the user to enter expert knowledge and other information outside of the data. This information may help GeNIe find the correct graphs[4] for the data. The DAG in Figure 4.3 represents the radical interpreter's beliefs. Entering this DAG into the Knowledge Editor of the feature Background Knowledge is equivalent to her attributing her beliefs to her speaker.

When the algorithm analyzes the data from the speaker, it uses the radical interpreter's DAG as a guide for which structures to include or exclude. Also, if the data is strong enough the algorithm will override the DAG that was entered (which is a result we want) because when GeNIe overrides the DAG for the radical interpreter it means her beliefs conflict with the choices

[4]The word 'graphs' is intentionally plural to align with the indeterminancy Davidson discussed. That is, different valid interpretations for the utterances in a set are possible partly because different valid belief networks are possible based on the same speaker data.

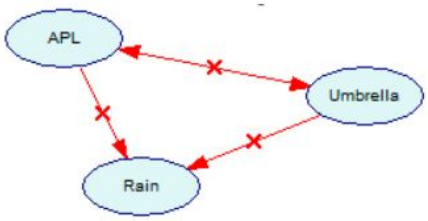


Figure 4.3: DAG of interpreter's beliefs

her speaker has made. Thus, we have a formal mechanism by which we can tell that the attributed beliefs of the radical interpreter conflict with the speaker's data.

Observing the speaker

To formally represent the radical interpreter observing her speaker, a data file was created containing the information the radical interpreter would record (or memorize) from her observations. In the scenario the radical interpreter observes the morning ritual of her speaker for 1000 days. She records information about the environment and what he did. Specifically, she records the reading on the barometer, whether it was raining at the time, and whether he took an umbrella. For Experiment 1 Microsoft Excel was used to randomly generated the data for these 1000 days. Below is a description of what this Excel file reflects.[5]

1. It rains 40% of the time the air pressure is low;
2. If the air pressure is low, then the chance of rain is 60%;
3. If the air pressure is not low, then the chance of rain is 30%; and
4. If it is raining or the air pressure is low, then the speaker takes an umbrella.

Discovering whether the attributed beliefs are adequate

To formally represent the radical interpreter discovering her attributed beliefs are false the following was done. Unlike the two previous formal representations that were solely done within GeNIe, this one involved the experimenter (who in this case is me) analyzing graphs that GeNIe derived from the data. This process of analysis can easily be done with a computer, since it involves verifying whether a particular graph belongs to a set of graphs.

[5]Appendix C contains a step-by-step description of how Excel was used to create this file.

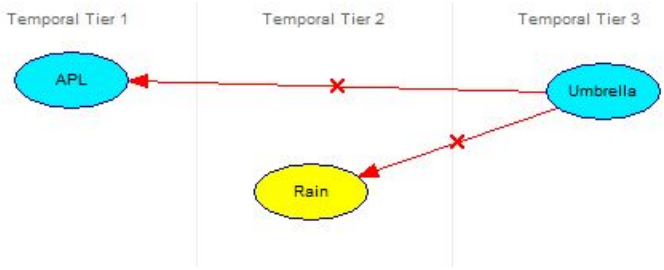


Figure 4.4: Before structure learning with attributed beliefs (A)

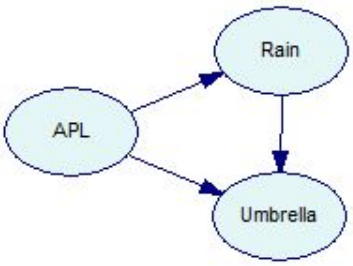


Figure 4.5: After structure learning on attributed beliefs (A2)

Using the data from the Excel file described above and the radical interpreter's DAG that had been entered into Background Knowledge, GeNIe, using the PC algorithm, derived a set of possible graphs that fit the data. In this case, the set contained only one graph, which is Figure 4.5. Formally we can conclude that the attributed beliefs will not work because the Bayesian network for the radical interpreter is not in the set of possible graphs that the PC algorithm derived.

Revising her belief about his beliefs

To formally represent the radical interpreter revising the set of beliefs she has for her speaker a process must be defined so that it selects the correct graph from the set mentioned in the last section. As in the previous section, the experimenter (who again in this case is me) has to manually revise the set of beliefs that the radical interpreter uses for her speaker. But, because in this case the set contains only one graph, this is the only choice. It turns out that, in fact, this directed acyclical graph (DAG) indeed represents the speaker's beliefs.[6]

[6]As we will see in Experiment 1B, more than one graph can be suggested for a set of data. In Experiment 1B, I offer a possible algorithm for dealing with more than one graph.

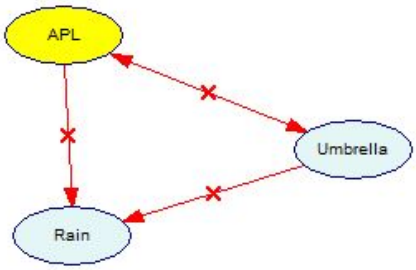


Figure 4.6: Before structure learning with attributed beliefs (B)

4.3.2 Results of Experiment 1A

On the positive side the results of Experiment 1A demonstrate that GeNIe was able to formally represent the beliefs of the radical interpreter, simulate her attributing her beliefs to the speaker, finding out that they are wrong, and discovering the beliefs of her speaker. On the negative side, the Bayesian network that was used to represent the beliefs of the radical interpreter had some unwarranted temporal restrictions. Because the scenario specifies that she grew up in a region of the world where air pressure has no noticeable affect on the chance of rain, she has no reason to believe that changes in the air pressure precede changes in the chance of rain. Therefore, using using temporal restrictions among the nodes for air pressure and rain in the network that represents her beliefs was unjustified. Removing the temporal restrictions from the Bayesian network for the radical interpreter made this model more accurate. This change was the only one made for Experiment 1B.

4.3.3 Experiment 1B

The set up of Experiment 1B matches Experiment 1A in every respect except that a different belief network was defined for the radical interpreter, Figure 4.6. The set of possible graphs that GeNIe derived for the speaker by running the Learn New Network using the PC algorithm on the data is Figure 4.7.

Discovering attributed beliefs inadequate

Because the Bayesian network for the radical interpreter's belief is not in the set of possible graphs that GeNIe derived using the PC algorithm, we can formally conclude that her attributed beliefs do not fit the speaker's data.

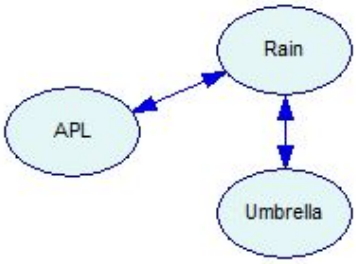


Figure 4.7: After structure learning on attributed beliefs (B)

Revising her belief about his beliefs

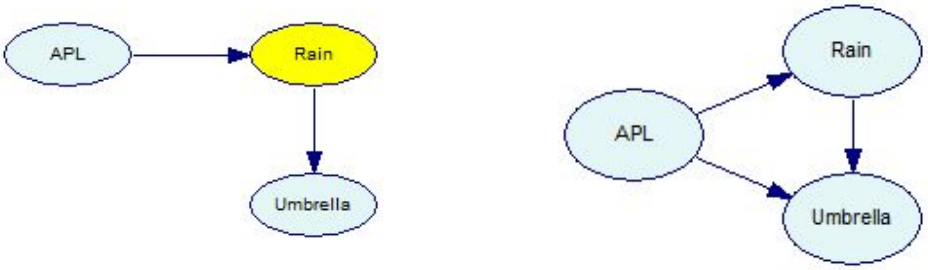
The set of possible graphs the PC algorithm derived is Figure 4.7, which represents four possible DAGs. To see whether GeNIe could derive the DAG that represents the speaker's belief, GeNIe re-analyzed them using each of the four DAGs as the Background Knowledge. The resulting graphs are shown next to the origial DAG in Figures 4.8 to 4.9. On the left is the original DAG; on the right the set of graphs that the PC algorithm derived.

At least three methods exist within a Davidsonian framework for interpreting these results. One falls under the heading *indeterminancy*, another under *the principle of charity*, and the third under *best fit*. These three methods are:

- *Indeterminancy*: Because different sets of beliefs can fit the same data, keep the eight graphs from the four sets (the set from DAGs #3 and #4 contain two DAGs each), and use them when interpreting the set of utterances of the speaker. This means at least eight interpretations of the sets of utterances are possible.

- *Principle of charity*: Because of the assumption that the speaker's beliefs share much in common with the radical interpreter's beliefs, use her beliefs to put the possible graphs in order from more likely to less likely to happen. This can be done by putting in order all the parent-to-child components of the DAG for the radical interpreter and create a numerically measurement for how important this parent-child relationship is to her.

  An example might help. Suppose we have the same scenario described at the beginning of this chapter. Also suppose that the scale of disbelief goes from zero (which means no disbelief) to minus one (which means the impossible state – impossible to believe). Recall that in Bayesian networks an arrow from one node to another means that changes in



(a) DAG #1

(b) Set of graphs from DAG #1

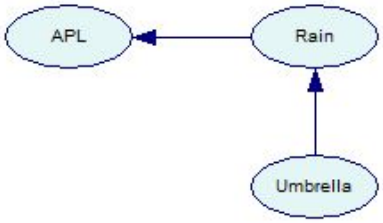
Figure 4.8: DAG #1 of set from PC algorithm

the state or value of the first node correlate to changes in the second node.[7] So an arrow going from APL to Rain means changes in the air pressure correlate with changes in the chance of rain; whereas an arrow from Umbrella to Rain means taking an umbrella correlates with changes in the chance of rain. Let the difference between having no arrow from APL to Rain and adding one is -0.25; and the difference between having no arrow from from Umbrella to Rain and adding one is -0.9.[8] And further suppose our formal model has to put in order two DAGs that possibly represent the speaker's beliefs. One has APL $\rightarrow$ Rain and the other has Umbrella $\rightarrow$ Rain. With respect to our radical interpreter the degree of disbelief for the first DAG is -0.25 and -0.9 for the second, so the first DAG is ordered before the second. However, no DAGs are deleted in the process to allow for indeterminancy.

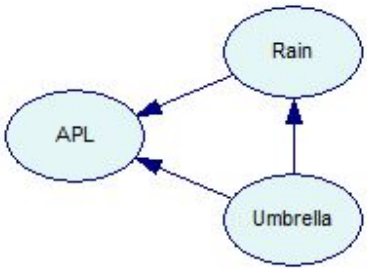
- *Best Fit*: Suppose GeNIe learns the structure of the same set of data twice using two different DAGs as Background Knowledge and both times the set of possible graphs has only one graph. If one of these possible graphs has more arrows that the other, this means, the original DAG entered for the Background Knowledge fits the data. With this in mind, the best fit method is to select from the sets of possible graphs the ones that are either DAGs or are the closest to being DAGs, then calculate an order on these using the algorithm described in the principle of charity method.

[7] Recall: correlation does not mean causation.

[8] This example involves counterfactual reasoning, which is beyond the scope of this thesis. The radical interpreter has to do counterfactual thinking, e.g., "Which is harder for me to believe? That air pressure affects the chance rain? Or That the act of taking an umbrella affects the chance of rain?

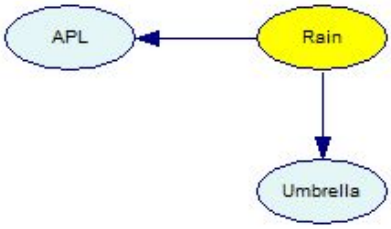


(a) DAG #2

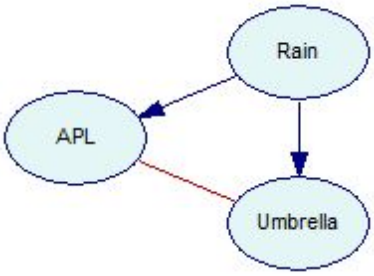


(b) Set of graphs from DAG #2

Figure 4.9: DAG #2 of set from PC algorithm

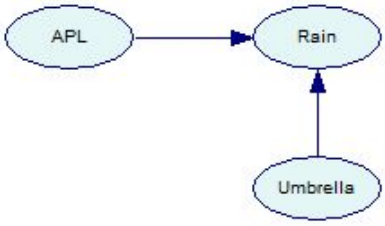


(a) Dag #3

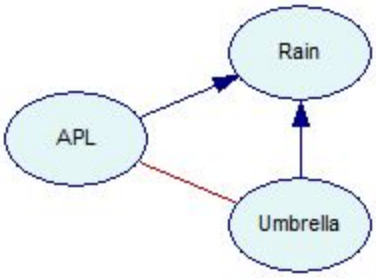


(b) Set of graphs from DAG #3

Figure 4.10: DAG #3 of set from PC algorithm



(a) DAG #4



(b) Set of graphs from DAG #4

Figure 4.11: DAG #4 of set from PC algorithm

4.3.4 Analysis of the results of Experiment 1B

Experiment 1B had all the positive benefits that Experiment 1A had without the problem of including a temporal order in the Bayesian network that represents the radical interpreter's beliefs. In addition to this, the formal process of Experiment 1B resulted in a set of four possible graphs that fit the speaker's data and three methods were created for either justifying keeping all of the proposed graphs or putting the graphs in order using the DAG that represents the beliefs of the radical interpreter. Thus the following were accomplished using GeNIe to formally represent different parts of Davidson's unified theory:

1. The radical interpreter attributes her beliefs to her speaker.
2. The radical interpreter discovers the beliefs she attributed do not fit the speaker's beliefs.
3. The radical interpreter correctly discovers the speaker's beliefs.

However, the above experiments did not involve any utterances by the speaker. In a sense, they were like silent movies. The next set of experiments include sound, i.e., include utterances.

4.4 Experiment 2

The focus of Experiment 2 is to see how much and how well GeNIe can formally represent a scenario where a radical interpreter interprets the utterance of a speaker "from scratch". The purpose of these experiments is not to create a fully working formal model, but to see how much can be modeled. Since the focus of these experiments is on utterances, it is assumed that the radical interpreter and her speaker share the same correct beliefs about their environment (that is, they both believe that low air pressure significantly raises the chance of rain and this is indeed the case in the environment where they live).

The element of noise was added to the system to see whether and how well GeNIe could filter out this noise and not include it in the information about the utterances. This was done by simulating the random flips of a coin.

4.4.1 Scenario behind Experiment 2

The scenario on which Experiment 2 is based is below.

> *Backstory*: A radical interpreter and her speaker both grew up in a region of the world where low air pressure raises the chance

of rain to 60% and they both correctly believe this. Recently she moved to his city and wants to learn the speaker's language.

The daily scenario: Every morning unbeknownst to the speaker, the radical interpreter observes him as he does his morning ritual before leaving. First he examines a coin. Second he looks at the barometer. Third, he looks out the window. Fourth he makes an utterance. And fifth, if he believes it probably will rain, he takes an umbrella; if not, then no umbrella is taken. Then he leaves. Our interpreter observes and records this daily routine 1000 times. Then she analyzes the data, revises her attributed beliefs if needed, then calculates his subjective beliefs and his preferences. For each of her speakers individual utterances she determines what was present in the environment each time that particular utterance was made.

4.4.2 Formally representing the scenario

Beliefs: Since their beliefs are the same I put (1) for both of their beliefs.

1. $B_{speaker} = B_{interpreter} = \{P(\text{rain}|\text{air pressure low}) = 60\%\}$.

Technically, if our network only represents the beliefs of an agent, then neither the umbrella nor the utterance nodes should be part of the network, because even if it was raining extremely hard, the speaker could choose to leave his umbrella behind or choose to utter any of the four available utterances in this model. However, for the sake of simplicity the umbrella and utterance nodes are included with arrows from the rain and air pressure nodes, based on the fact that any time the speaker believes it probably will rain, he takes an umbrella and chooses a context-appropriate utterance. This means that the arrows from the air pressure and rain nodes to the umbrella and utterance nodes imply a Bayesian decision theory network (in fact it is a hidden influence diagram). For similar reasons the Bayesian networks that represent the beliefs of the radical interpreter and the one she has for her speaker also have the umbrella and utterance nodes with arrows going to it from other parent nodes.

Actions: The set of actions consist of only one element $A = \{a_1\}$, where a_1 is "take an umbrella".

Coin: The set of states for a coin consists of two element $A = \{H, T\}$, where H is "heads" and T is "tails".

Utterances: The set of utterances the speaker can choose from is $\Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$, where the different ϕ_i's are defined below.

ϕ_1 = "It probably will rain."

ϕ_2 = "It will rain."

ϕ_3 = "It probably will not rain."

ϕ_4 = "It will not rain."

4.4.3 Experiment 2A

Representing beliefs

The Bayesian networks for the radical interpreter and speaker are the same and is shown in Figure 4.12. The three arrows from the *APL* node to the *Rain*, *Umbrella*, and *Phi* nodes indicate a probabilistic relationship exists between low air pressure and the chance of rain, the choice to take an umbrella, and the choice of an utterance. Similarly the two arrows from the Rain node to the *Umbrella* and *Phi* nodes indicate the speaker believes a causal relationship exists between the rain and the choice of taking an umbrella and the choice of making an utterance.

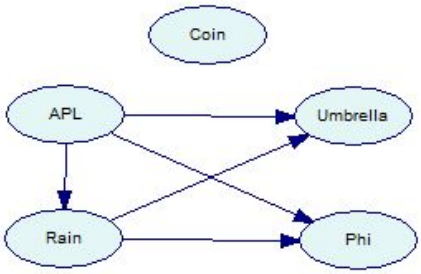


Figure 4.12: DAG of the speaker's beliefs

Attributing beliefs

To formally represent the radical interpreter attributing her beliefs to her speaker the DAG in Figure 4.13 is entered in the Background Knowledge so GeNIe can use this when deriving possible Bayesian network structures from the speaker data. Formally speaking since the DAG for the radical interpreter is in the set of possible beliefs the PC algorithm derived, the attributed beliefs do fit the speaker data.

Observing her speaker

To formally represent the radical interpreter observing her speaker, a data file was created containing the information the radical interpreter would

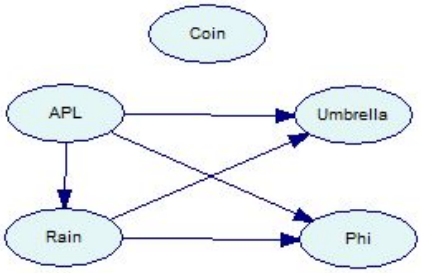


Figure 4.13: Experiment 2A: Attributing DAG of interpreter's beliefs

record (or memorize) from her observations. In the scenario the radical interpreter observes the morning ritual of her speaker for 1000 days. She records what he uttered and information about the environment. Specifically, she records the reading on the coin, the barometer, whether it was raining at the time, what utterance the speaker made, and whether he took an umbrella. For Experiment 2 Microsoft Excel was used to randomly generated the data for these 1000 days. Below is a description of what this Excel file reflects.[9]

1. It rains 40% of the time the air pressure is low;
2. If the air pressure is low, then the chance of rain is 60%;
3. If the air pressure is not low, then the chance of rain is 30%; and
4. If it is raining or the air pressure is low, then the speaker takes an umbrella.
5. If it was raining, the speaker uttered ϕ_2 = "It will rain."
6. If it was not raining and the air pressure was low, the speaker uttered ϕ_1 = "It probably will rain."
7. If it was not raining and the air pressure was not low, the speaker uttered ϕ_3 = "It probably will not rain."

Discovering whether the attributed beliefs are adequate

In this experiment not only is the DAG for the radical interpreter's belief contained in the set of possible graphs that the PC algorithm derives from the speaker data, it is the only one in the set, which is what we expect because the data was generated based on the Bayesian network for the speaker's beliefs.

[9]Appendix D contains a step-by-step description of how Excel was used to create this file.

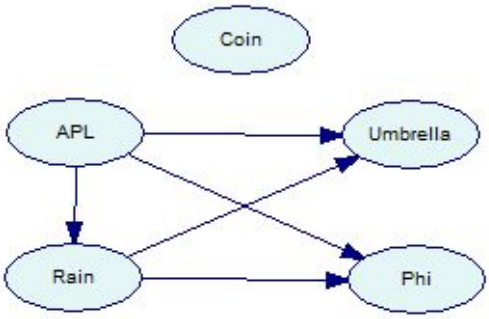


Figure 4.14: Experiment 2A: The attributed beliefs before structure learning

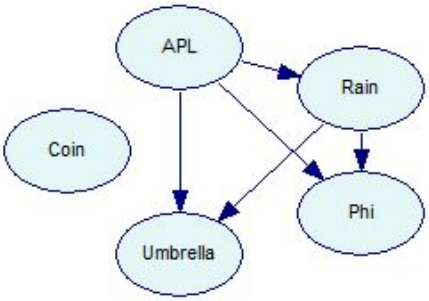


Figure 4.15: Experiment 2A: After structure learning on attributed beliefs

Revising her belief about his beliefs

Because the DAG that represents the beliefs of the radical interpreter is an element in the set of possible graphs that the PC algorithm derived from the speaker data, no belief revision is necessary.

4.4.4 Results of Experiment 2A

On the positive side the results of Experiment 2A demonstrate that GeNIe was able to formally represent the beliefs of the radical interpreter, simulate her attributing her beliefs to the speaker, her finding out that her beliefs do fit the data from the speaker, and her discovering the beliefs of her speaker. On the negative side, information about what conditions hold in the environment when the speaker makes various utterances is not clear. This information may exist in the hidden conditional probability tables. This led to considering how to make environmental variables that are associated with the various utterances clearer. Thus, we have Experiment 3.

4.5 Experiment 3

The focus of Experiment 3 is to see how much and how well GeNIe can formally derive T-sentences from the speaker data for the various utterances of the speaker. For this experiment it is assumed that the radical interpreter and her speaker share the same correct beliefs about their environment (that is, they both believe that low air pressure significantly raises the chance of rain and this is indeed the case in the environment where they live). If GeNIe successfully accomplishes this – that is, it correctly derives from the speaker data the environmental conditions that were present consistently when each specific utterance was made – then progress would be made toward GeNIe deriving T-sentences from the data.

The element of noise is retained from Experiment 2 to see whether and how well GeNIe could filter out this noise and not include it in the information about the utterances. This was done by simulating the random flips of a coin.

4.5.1 Scenario behind Experiment 3

The scenario on which Experiment 3 is the same as Experiment 2.

4.5.2 Restructuring the nodes for action and utterances

All the formal representations for Experiment 3 are the same as Experiment 2, except for nodes related to utterances and the umbrella choice. Because of this only these two will be discussed below. The key idea is to make into

a node each element in the set of a node. For example instead of having one node $\Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$, have four individual nodes each containing an element of Φ. That is, nodes Φ_1, Φ_2, Φ_3, and Φ_4. Each of these nodes would have two states. For example, Φ_1 would have ϕ_1 and $\neg(\phi_1)$, which can be represented as a 1 when the speaker has chosen to utter ϕ_1 and 0 when he has made some other choice.

Actions: Two sets or nodes represent the choices the speaker makes as to whether to take an umbrella.

- TakeUmbrella = $\{take\_umbrella, 0\}$
- LeaveUmbrella = $\{leave\_umbrella, 0\}$

Utterances: The sets of utterances the speaker can choose from is Φ_1, Φ_2, Φ_3, and Φ_4 as defined below:

$\Phi_1 = \{\phi_1, 0\}$, where ϕ_1 is "It probably will rain."

$\Phi_2 = \{\phi_2, 0\}$, where ϕ_2 is "It will rain."

$\Phi_3 = \{\phi_3, 0\}$, where ϕ_3 is "It probably will not rain."

$\Phi_4 = \{\phi_4, 0\}$, where ϕ_4 is "It will not rain."

4.5.3 Reformatting speaker data

The observation that the radical interpreter makes in Experiment 3 is the same as in Experiment 2. However, to reflect the changes in the action and utterance nodes described above, the data in the Excel file has to be rearranged. The main difference in the formatting of this data file is that, instead of having one Φ-column and one *Action*-column, there are four Φ_i columns and columns for TakeUmbrella and one for LeaveUmbrella.

4.5.4 Structure learning with the reformatted data file

GeNIe learned the structure of the reformatted Excel data file using the DAG in Figure 4.16 for the radical interpreter's beliefs for the Background Knowledge. All of the red arrows are in this DAG because of our assumption (and the radical interpreter's beliefs) that choosing an utterance and choosing whether to take an umbrella do not cause each other (at least in this case) and neither do choices among utterances and taking umbrellas affect air pressure, the chance of rain, or the flip of a coin. The blue arrow from *APL* to *Rain* reflects her true belief that low air pressure does affect the chance of rain.

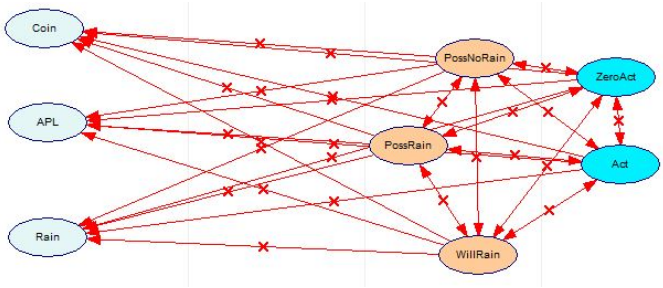


Figure 4.16: Experiment 3: DAG for radical interpreter's belief

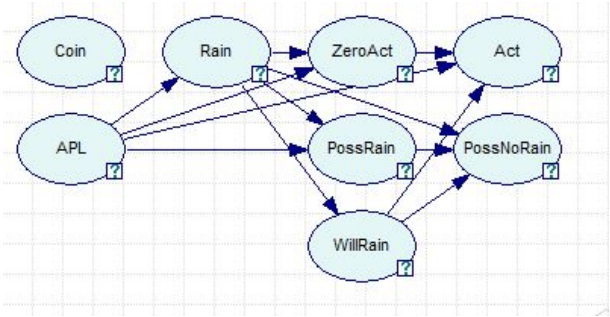


Figure 4.17: Experiment 3: After structure learning on attributed beliefs

The DAG that GeNIe derives from this using the Greedy Thick Thinning algorithm is seen in Figure 4.17. All of the arrows match our intuitions. For instance a blue arrow goes from *APL* to *Rain*, *ZeroAct*, *ProbRain*, and *ProbNoRain*(via *ProbRain*). However, no arrow goes from *APL* to the utterance node *WillRain*, which makes sense because the only time that the speaker utters "It will rain" is when it currently is raining and this holds no matter what the state of *APL* is.

4.5.5 Analysis of the environmental context of utterances

The focus of Experiment 3 was to see whether and how well GeNIe could extract information about the individual utterance. More specifically, whether it could show under what conditions in the environment the different utterances are made. In the previous section we see that, node-wise, GeNIe seems to be getting the big picture correct. Now let us examine the details of the conditional probability tables for the three utterances *PossNotRain*, *ProbRain*, and *WillRain*. Before doing so, one note is in order. When GeNIe analyzes data that contains no labels it assigns for each column (or node) $S0$ and $S1$ for State-0 and State-1. The situation with our reformatted data is a half-and-half one. That is, one label is given and the other is the number zero, e.g., TakeUmbrella = $\{take\_umbrella, 0\}$. In this example, $S0$ stands for "a choice other than take-umbrella". Similarly for every node X, $S0$ in

APL	Low				NotLow			
Rain	IsRaining	IsNotRaining	IsRaining	IsNotRaining	IsRaining	IsNotRaining	IsRaining	IsNotRaining
WillRain	SO	MustRain	SO	MustRain	SO	MustRain	SO	MustRain
S0	0.500	0.998	0.997	0.500	0.500	0.997	0.001	0.500
MightNotRain	0.500	0.002	0.003	0.500	0.500	0.003	0.999	0.500

Figure 4.18: Experiment 3: Conditional probability table for *PossNotRain*

APL	Low		NotLow	
Rain	IsRaining	NotRaining	IsRaining	NotRaining
SO	0.998	0.003	0.997	0.999
MightRain	0.002	0.997	0.003	0.001

Figure 4.19: Experiment 3: Conditional probability table for *ProbRain*

the table represents "the choice other than X".

It probably won't rain.

In Figure 4.18 is the conditional probability table for the *ProbNoRain*, which is the utterance "It probably will not rain." From the table below we see that whenever the speaker has made this utterance, it was not raining (Rain = NotRaining) and the air pressure was not low (APL = NotLow). This matches our intuition.

It might rain

In Figure 4.19 is the conditional probability table for the *ProbRain*, which is the utterance "It probably will rain." From this table we see that whenever the speaker has made this utterance, it was not raining (Rain = NotRaining) and the air pressure was low (APL = Low). This matches our intuition.

It will rain

In Figure 4.20 is the conditional probability table for the *WillRain*, which is the utterance "It will rain." From this table we see that whenever the speaker has made this utterance, it was raining (Rain = IsRaining) and no other environmental conditions apply. This matches the intution with the understanding that if it is raining right now, then in five minutes, it will be raining. Thus, the reason why the utterance "It will rain" is held true when right now it is raining.

Rain	IsRaining	NotRaining
SO	0.001	0.999
MustRain	0.999	0.001

Figure 4.20: Experiment 3: Conditional probability table for *WillRain*

The above results demonstrate, at least for this experiment that GeNIe was able to correctly derive the different environmental factors that held whenever each of the speaker's utterances were made. I claim that GeNIe was able to derive from the speaker data a T-sentences for each of three utterances the speaker used that gave the correct environmental conditions under which each of the utterances was held true by the speaker.

4.5.6 Results of Experiment 3

On the positive side, in Experiment 3 GeNIe correctly derived from the speaker data the environmental factors that had to be in place for the speaker to make one of the four utterances and hold the utterance true.

4.6 Next steps...

On the one hand, we could focus on improving the current model to make it more realistic. For example, we could modify the barometer variable to have more than two states (say, Low, Medium, and High) or to be a continuous variable so that it can be any real number within a specified range. There are many other ways to improve the current model – add a temporal element, have the speaker correct his beliefs in the middle of the observation period (which would give the radical interpreter mixed data), rework GeNIe so that data about old utterances have less weight than data about new utterances, etc.

On the other hand, if we decided what is the one thing such that if we get it, the most amount of progress would be made or made possible. And this one thing is the missing preference order. Once we get a way to systematically put a preference order on the speaker's utterances, we will have a way to derive the cardinal utilities of the speaker from which we can derive his subjective probabilities (beliefs) which we can use with the T-sentences to try to maximise agreement (according to the radical interpreter's lights). And how might we get from here to a preference order? By focusing on the unverified suggestions from the last chapter, which are copied below.

1. *Restructure data file (Part II)*: This suggestion proposes using the same procedure as the previous suggestion, but has not been verified.

But, using the procedure it might be possible to derive T-sentences for more theoretical sentences. That is, for sentences whose holding-true value depends solely on other sentences and not states or events in the environment.

2. *Restructure data file (Part III)*: Same set up procedure as before. This should be able to derive the T-sentences for complex-theoretical sentences. That is, sentences that are composed of other sentences using AND, OR, NOT, or IF-THEN. Again, this should derive T-sentences for these utterances.[10]

3. *Restructuring data file (Part IV)*: Same set up as before. This suggestion has not been as fully worked out conceptually as the previous two. This *might* provide a way for deriving a preference relationship for the speaker among his utterances. The motivating idea is that Ramsey (1926) and Jeffrey (1965) used a proposition and its negation to form complex gambles or propositions, respectively, to determine the preferences a subject had among choices.

It seems the best course of action is to focus on realising the above suggestions.

[10]If this Part III is successful, we may be able to use this to determine what the logical connectives are in the speaker's language. These would (should?) be equivalent to AND, OR, NOT, and IF-THEN. Davidson (1980, reprinted in Davidson 1984a, p. 163) claims that once all these connectives are discovered (or the Sheffer stroke is, which then would allow us to derive the other connectives) Jeffrey's decision theory would uncover the logical structure of the speaker's language.

Chapter 5

Discussion

This section offers two possible areas of research to which a formal model of Davidson's radical interpretation (once it is created) can be used to investigate.

5.0.1 Epistemic-ontological interface

On the one hand the following seems obvious, yet on the other it may contain subtle nuances: Changes in the environment cause changes in beliefs. Likewise, changes in beliefs can cause changes in the environment (through action based on intentional choices.[1]) Below are quotes by Jeffrey (1965) and Davidson (1980) that argue for causal relationships between the real world (ontology) and what an agent believes (epistemology) and vice versa.

Quotes about agents' observations of their environment changing their belief:.

> We shall now consider cases in which an agent's belief function changes from *prob* to $prob_B$ as a result of an observation; where the agent's conclusive belief in B is caused by the observation; is unreasoned; and is justified by the consideration that the observation is of the paradigmatic sort to which any normal speaker of the language in which B is expressed would respond by believing B, willy-nilly. (Jeffrey 1965, Chapter 12, p. 172)

Davidson (1980, reprinted in Davidson 1984a, p. 158) says:

> What the interpreter has to go on, then, is information about what *events in the world cause an agent to prefer that one rather than another sentence be true.* (Davidson 1980, reprinted in Davidson 1984a, p. 158 (italics mine))

[1] Note I leave open whether the goal of the intention behind the choice to act is realized. My emphasis is that an act that is intentionally chosen affects the environment.

In his outline of how to find the logical structure of a speaker's language and Davidson (1980, reprinted in Davidson 1984a, p. 165) also says:

> Further steps in interpretation will require some elaboration of the empirical basis of the theory; it will be necessary to attend, not just to the agent's preferences among sentences, but also to *the events and objects in the world that cause his preferences (and hence also his beliefs)*. Thus it will be the observable circumstances under which an agent is caused to assign high or low probabilities to sentences like "It is raining".... (Davidson 1980, reprinted in Davidson 1984a, p. 165 (italics mine))

It seems that at least three kinds of causal relationships exists:

1. Events in the real world causing change in other real world events.
2. Events in the real world causing an agent to change his beliefs.
3. An agent's intentional act causing a change in the real world.

Furthermore, the formal model created in the experiment used all three: The event of low air pressure caused an increase in the chance of rain (Item 1); the event of low air pressure caused the speaker to believe it probably would rain (Item 2); and the agent's intentional action of taking an umbrella prevented rain from falling on him (Item 3).

Therefore, once formal models of Davidson's unified theory have been created, they may help us tease apart different aspects of causality and intentional action.

5.0.2 Prequel to game theory

The original goal of this thesis was to formally represent Davidson's radical interpretation in a game theoretic model. However, Bayesian networks worked better because of the uncertainty in the subjective probabilities and the belief updating process via Bayes' theorem. Yet one question about the players in game theoretic games has been "Where do the homogeneous players in a game theoretic model come from?" The reason for this question is either all of the players in a game theoretic model are copies of each other (for instance they have identical utilities) or each player knows exactly what the other players value (including whether the player was risk-averse or risk-taking).[2] Davidson (1973, reprinted in Davidson 1984a, p. 134) originally

[2]Also in this mix of question was the problem that classical game theory had with the case where a player makes an off-path choice – that is, a choice that one player thought the other player would never consider.

proposed his radical interpretation for radically interpreting a community of speakers. It seems that if each speaker were a radical interpreter within this community, then a common understanding of what each person values, how each views risk, and what each would choose could become common knowledge. In other words, applying Davidson's unified theory in a community of radical interpreters could transform it to a community of game theoretic players.

Bibliography

[Cow99] Robert G. Cowell. *Probabilistic Networks and Expert Systems Exact Computational Methods for Bayesian Networks*. eng. 1st ed. 1999. Information Science and Statistics. New York, NY: Springer New York, 1999. ISBN: 1-280-14578-1.

[Dav67] Donald Davidson. "Truth and Meaning". eng. In: *Synthese* 17.3 (1967), pp. 304–323. ISSN: 00397857.

[Dav73] Donald Davidson. "Radical Interpretation". In: *Dialectica* 27.3-4 (1973), pp. 313–328. DOI: 10.1111/j.1746-8361.1973.tb00623.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1746-8361.1973.tb00623.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1746-8361.1973.tb00623.x.

[Dav74] Donald Davidson. "Belief and the basis of meaning". eng. In: *Synthese* 27.3 (1974), pp. 309–323. ISSN: 0039-7857.

[Dav80] Donald Davidson. "Toward a Unified Theory of Meaning and Action". In: *International Journal for Analytic Philosophy* 11 (1980), pp. 1–12. eprint: https://doi.org/10.5840/gps19801120.

[Dav84a] Donald Davidson. "Inquiries into truth and interpretation". eng. In: *Inquiries into truth and interpretation*. Oxford: Clarendon Press, 1984. Chap. 9. ISBN: 019824617X.

[Dav84b] Donald Davidson. "Inquiries into truth and interpretation". eng. In: *Inquiries into truth and interpretation*. Oxford: Clarendon Press, 1984. Chap. 10. ISBN: 019824617X.

[Dav94] Donald Davidson. "Radical Interpretation Interpreted". eng. In: *Philosophical Perspectives* 8 (1994). ISSN: 1520-8583. URL: http://search.proquest.com/docview/1300235773/.

[Dav95] Donald Davidson. "Could there be a science of rationality?" In: *International Journal of Philosophical Studies* 3.1 (1995), pp. 1–16. DOI: 10.1080/09672559508570801. eprint: https://doi.org/10.1080/09672559508570801. URL: https://doi.org/10.1080/09672559508570801.

[Fin75] B. de. Finetti. *Theory of Probability (Volumes 1 and 2)*. eng. Wiley series in probability and mathematical. New York: John Wiley, 1975.

[FL94] Jerry Fodor and Ernie Lepore. "Is Radical Interpretation Possible?" eng. In: *Philosophical Perspectives* 8 (1994), pp. 101–119. ISSN: 15208583.

[Glü06] Kathrin Glüer. "The Status of Charity I: Conceptual Truth or A Posteriori Necessity?" eng. In: *International Journal of Philosophical Studies: Donald Davidson (1917-2003)* 14.3 (2006), pp. 337–359. ISSN: 0967-2559. URL: http://www.tandfonline.com/doi/abs/10.1080/09672550600858320.

[Jef65] Richard Jeffrey. *The logic of decision*. eng. McGraw-Hill series in probability and statistics. (NL-LeOCL)810651750. New York: McGraw-Hill, 1965.

[Lew74] David Lewis. "Radical interpretation". eng. In: *Synthese* 27.3 (1974), pp. 331–344. ISSN: 0039-7857.

[Mar16] J.F. Marti. "Interpreting linguistic behavior with possible world models". eng. In: (2016).

[Pag06] Peter Pagin. "The Status of Charity II: Charity, Probability, and Simplicity 1". eng. In: *International Journal of Philosophical Studies: Donald Davidson (1917-2003)* 14.3 (2006), pp. 361–383. ISSN: 0967-2559. URL: http://www.tandfonline.com/doi/abs/10.1080/09672550600868683.

[Pea00] Judea. Pearl. *Causality : models, reasoning, and inference*. eng. Cambridge: Cambridge University Press, 2000. ISBN: 0521773628.

[Ram31] Frank Ramsey. "Truth and Probability". In: *Foundations of Mathematics and other Logical Essays*. Ed. by R.B. Braithwaite. New York: Harcourt, Brace and Company, 1931, pp. 156–198.

[SSS19] Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. "A survey on Bayesian network structure learning from data". eng. In: *Progress in Artificial Intelligence* 8.4 (2019), pp. 425–439. ISSN: 2192-6352.

[Tar44] Alfred Tarski. "The Semantic Conception of Truth: and the Foundations of Semantics". In: *Philosophy and Phenomenological Research* 4.3 (1944), pp. 341–376. ISSN: 00318205. URL: http://www.jstor.org/stable/2102968.

[TK81] A Tversky and D Kahneman. "The framing of decisions and the psychology of choice". eng. In: *Science (New York, N.Y.)* 211.4481 (1981), pp. 453–458. ISSN: 0036-8075.

Appendices

C Experiment #1: Description of data

In Excel the following was done:

1. Column 2 had random numbers generated between 0 and 1, which were copied and re-pasted using Past-Special: number so that the value of each cell would not change.

2. Column 1 had numbers 1 to 1000 in order to be used at the end to put the data generated back into random order.

3. In the APL column I had Excel assign 1 (= Air Pressure Low) 40% of the time using a random number generator.

4. I sorted the APL column so that the 0's were together and also the 1's were together.

5. I then selected the part of the Rain column that had APL = 1 in the cell to the left. For these I had Excel assign 1 to Rain 60% of the time.

6. I then selected the part of the Rain column that had APL = 0 in the cell to the left. For these I had Excel assign 1 to Rain 30% of the time.

7. For the Umbrella column I had Excel assign a value 1 (= Take An Umbrella) whenever Rain or APL had a 1 and zero for all the other cases.

8. Then I resorted the rows of the APL, Rain, and Umbrella columns by putting them in order by the original order of Step 2 of this list.[3]

I did go through and check the zeros and ones in the various columns and they appeared to be in the right amount (e.g., 394 cells of 1000 APL cells had 1 which is close to the 40% it needed to be).

D Experiment #2: Description of data

To create the data file for Experiment #2, I took the data file from Experiment #1 and did the following:

1. Added a column for COIN and had Excel randomly generate values between zero and one, then had Excel assign 1 for Heads if the value was less than 0.50 and 0 for Tails if over 0.50.

[3]The first time I created data for my miniature-model I didn't do Steps 2 or 7 of this list, which meant APL was zero for the first 600 or so cells, then one for the remaining 400 or so. This may have caused the structure learning to be skewed in some way because some of the variables had the same value for the first number of runs then switched to a different value the rest of the runs.

2. Added a column for PHI with the following IF-THEN formula:

    - If Rain = 1, then PHI was 2 (="It will rain.")
    - If Rain = 0 and APL = 0, then PHI was 3 (="It probably will not rain.")
    - If Rain = 0 and APL = 1, then PHI was 1 (="It probably will rain.")

3. Added a column for Umbrella with the following IF-THEN formula.

    - If Rain + APL greater than 0, then Umbrella = 1 (= "take umbrella")
    - If Rain + APL = 0, then Umbrealla = 0 (="do not take umbrella")