Chapter 4
Mathematizing Betrayal

I kiss thy hand, but not in flattery, Caesar.
—Brutus, in Julius Caesar

The present chapter is the result of a conviction the two of us gradually came to share as we worked toward realizing a (seemingly) literally creative machine: namely, a good artificial storyteller must be in command of the innumerable themes that drive both belles-lettres and formulaic fiction. One such theme is betrayal; it is this theme, courtesy of Shakespeare’s immortal play, for which Brutus and Brutus are of course named. Other such themes are: unrequited (reckless, passionate, selfless, . . . ) love, fanaticism (ranging from the spiritual to the violently carnal), revenge, jealously, self-deception (which Brutus partially “understands”), infatuation (from the mild to the insane), hatred (from the murderous variety to forms of contempt well short of inclinations to harm), alienation, despair, triumph, and so on. Now, if an artificial storyteller is to — as we put it — “be in command” of such territory, then those who would engineer such a creature find themselves in a rather nasty pinch. The pinch is that such themes must be cast in terms that an AI can digest and process; that is, these themes must be, for want of a better word, mathematized.1 But mathematizing (say) love, by any metric, is a rather tall task, no? Well, tall or not, it’s a task we

1The alternative is to build a machine which learns about the themes in question, but for many reasons (some of which we shared with you in Chapter 1), we find such an approach to be profoundly unpromising. From the standpoint of sheer engineering, it seems to us silly to try to coax evolutionary computation along in a robot like COG to produce in it the capacity to discuss betrayal
CHAPTER 4. MATHEMATIZING BETRAYAL

intend to conquer. And we start here, in this chapter, by taming the concept of betrayal. We end by briefly considering the next theme we plan to fully formalize: self-deception. (A nascent formalization of this concept already makes it into BRUTUS's stories.) If everything goes according to plan, BRUTUS, will be in command of many of the themes in question. (Our plans in this regard are clearly reflected in what is called thematic concept instantiation in the BRUTUS architecture covered in Chapter 6.)

The technique we follow in the attempt to rigorously define betrayal is one that has been followed repeatedly by analytic philosophers. We begin armed with fairly firm pre-analytic intuitions about what betrayal essentially is and associated intuitions about whether or not certain fixed cases involve betrayal. We bring this prior knowledge to bear in the evaluation of thought-experiments involving proposed necessary and sufficient conditions for the truth of the location

Agent $s_r$ betrays agent $s_d$.

The upshot of this evaluation may be a new proposal; if so, the process iterates ... until we have the definitive account we seek.

One remark before we launch. It may be thought that a defective account of betrayal is sufficient to undergird the production of interesting narrative. For example, consider the following "definition."

Def$_B$ 0 Agent $s_r$ betrays agent $s_d$ iff (if and only if) $s_r$ kills $s_d$.

That Def$_B$ 0 doesn't cut the mustard is wholly unexceptionable. (Clearly, if Jones kills Smith because Smith is terminally ill and begs Jones to pull the plug, no betrayal has occurred.) Despite its inadequacy, however, this definition can give rise to interesting stories. Along these lines, consider the following dialogue.

Brown: Jones! I'm sick, I tell you, sick. I'm sick in my heart, in my soul. You have betrayed your friend.

4.1 FIRST STABS AT A DEFINITION

Jones: Betrayed? I saved him — saved him from these hellish tubes, from doctors and nurses sticking things into him with no more regard for his body than a pin cushion, from pain I pray neither of us will ever experience.

Brown: When you take a man's life, you betray him; you take from him the most sacred thing there is.

Jones: Smith wanted to die. What had life become for him? Calling it life was a cruel joke. What I did, I did as a friend. There is no betrayal here, and you know it.

This dialogue derives from tension about whether the mercy killing in question is a case of betrayal. It does show that it is entirely possible to base narrative on defective notions of betrayal. But — and this is our rebuttal to the objection that finding a correct account of betrayal is superfluous — in order to spin stories around tension like that seen in this dialogue, the storyteller must have command over what is and what isn't betrayal. In other words, if we can supply BRUTUS with an accurate account of betrayal, he will be able to write stories involving not only this account, but about variations on it that yield concepts falling short of betrayal, but concepts which are nonetheless sufficiently betrayal-like. Our approach in the present chapter is designed not only to eventuate in a definition that captures betrayal, but to produce along the way definitions that can in principle undergird narrative whose themes include those which are merely betrayal-like.

4.1 First Stabs at a Definition

Very well; where do we start? Actually, there's a natural starting place: Margaret Boden, in her The Creative Mind, opines that much progress has been and will be made on capturing literary, "consciousness-laden" concepts, including, specifically, betrayal. In the case of betrayal, the "definition" that gives rise to Boden's sanguinity is

---

3This rebuttal was devised in dialectic arising from a talk and demo we gave at Brown University, and we are indebted to many who were in attendance there.
4.2. BETRAYAL CALLS FOR AN EXPRESSIVE LOGIC!

Sun to work just fine. Though clauses 1–4 are satisfied, we would be disinclined to call this a case of betrayal. The problem, at least one of them, would seem to be that Dave hasn’t done anything. If the Sun does indeed malfunction, we would certainly agree that Selmer is unlucky; and maybe, just maybe, we’d grant that Dave is lucky. But Dave hasn’t betrayed Selmer. (He may of course wish Selmer ill luck, and for that he may be blameworthy. But he isn’t guilty of betrayal.)

This suggests that we move to the following definition.

Def$_B$ 2 Agent $s_r$ betrays agent $s_d$ iff there exists some state of affairs $p$ such that

1. $s_d$ wants $p$ to occur;
2. $s_r$ believes that $s_d$ wants $p$ to occur;
3. $s_r$ agrees with $s_d$ that $p$ ought to occur;
4. $s_r$ intends that $p$ not occur;
5. $s_r$ believes that $s_d$ believes that $s_r$ intends that $p$ occur.

4.2 Betrayal Calls For an Expressive Logic!

Our attempt to formalize betrayal already reveals that the type of logic required must be extraordinarily expressive. In order to see this, let’s focus first on the implications arising just from clause 5 in Def$_B$ 2, which involves an iterated belief. Here, first, is a different, perhaps more colorful, way to see the extreme expressiveness that narrative can demand in connection with iterated beliefs.

4.2.1 Dizzingly Iterated Beliefs in Detective Stories

Consider a detective story pitting a villain against a hero. Suppose that initially the hero is given information that demonstrates that the villain is a villain. For example, perhaps the hero is shown demonstrative evidence that the villain murdered an innocent person. At this point in the narrative, the hero believes that the villain is a villain. Now suppose that the hero begins his investigation — at first in secret, but then in a way that (unknowingly, initially, to
the hero) reveals to the villain that the hero is on the job. At this point, the villain believes that the hero believes that the villain is a villain. Now, suppose that the hero discovers that the villain is aware of the hero's investigation. This implies that the hero believes that the villain believes that the hero believes that the villain is a villain. Next, suppose that the villain discovers that the hero is aware of the fact that his investigation is known to the villain. At this point, the villain believes that the hero believes that the villain believes that the hero believes that the villain is a villain! But we aren't done yet; not at all. Suppose that you read and understand a detective story conforming to the doxastic structure just described. (Many readily understood detective novels exemplify this structure. Consider, e.g., Gorky Park [217].) It follows from this that the reader believes that the villain believes that the hero believes that the villain believes that the hero believes that the villain is a villain. Next, it follows from the fact that Bringsjord has written this section that he believes that the reader believes (or at least can believe) that the villain believes that the hero believes that the villain believes that the hero believes that the villain is a villain. From this fact and the additional fact that Ferrucci has assimilated what Bringsjord has here written, we can conclude that Ferrucci believes that Bringsjord believes that the reader believes that the villain believes that the hero believes that the villain believes that the hero believes that the villain is a villain. Finally, some of the readers of this section will believe that Ferrucci believes that Bringsjord believes that the reader believes that the villain believes that the hero believes that the villain believes that the hero believes that the villain is a villain. (We said 'finally,' but do we dare point out that since Bringsjord wrote this section, and pondered it, he believes that readers of this section will believe that Ferrucci believes that Bringsjord believes that you can believe that the villain believes that the hero believes that the villain believes that the hero believes that the villain is a villain?)

What is the point of this dizzying iteration? The point is that fiction can rather easily generate, in the minds of readers, staggeringly complex beliefs—ones that first-order logic will at the very least have a devil of a time expressing. In first-order logic, even the sentence “Bringsjord believes that Ferrucci believes that the villain is a villain” is acutely problematic. If belief is represented by the two-place relation symbol $B$, and the proposition that the villain is a

villain is represented by the formula $\phi$, then this sentence becomes:

$$B(b, B(f, \phi))$$

But this formula is not grammatically acceptable in first-order logic. What we have devised so far in the attempt to mathematize betrayal—Def$_B$ 2—puts us firmly in the same boat that the detective scenario puts us in. In fact, as we will later see, better accounts of betrayal necessarily invoke third-order belief structures, that is, ones in which $x$ believes that $y$ believes that $z$ believes $\phi$.

### 4.2.2 OTTER Cannot Handle Betrayal

In Chapter 2 we introduced the theorem prover OTTER as an interesting tool to test logicist forays into story generation—as long as these forays demand nothing more than standard first-order logic. But despite the fact that our foray into the formal structure of betrayal has revealed that this structure moves us beyond first-order logic, some very interesting things happen when one implements Def$_B$ 2 in OTTER. To see them, consider first the formula

$$\forall x (P(x) \rightarrow Q(P(x))).$$

This formula is non-sensical on the standard grammar of first-order logic (which we gave, recall, in Chapter 2). The reason is that the antecedent in this universally quantified formula, $P(x)$, must be one that admits of truth or falsity. For the idea is that if it's true (for some instantiation to $x$), then the consequent, namely, that which is to the right of $\rightarrow$, must be true. (In technical terms, $P(x)$ is an atomic formula, not a term.) But this implies that the consequent consists of an atomic formula whose argument is itself an atomic formula, and this is ungrammatical and non-sensical in first-order logic. However, watch what happens when we give an input file to OTTER containing the relevant formula, along with the fact that $P(a)$, where $a$ is a constant, and the assumption for indirect proof, $\neg Q(P(a))$. Here is the input file:

```plaintext
set(auto).
formula_list(usable).
all x (P(x) -> Q(P(x))).
```

---

6For an introduction to the tools offered by logicist AI for tackling such recalcitrant sentences, see our [23].
CHAPTER 4. MATHEMATIZING BETRAYAL

\[ \text{P(a).} \]

% Assumption for contradiction:
\[-Q(P(a)). \]
\[ \text{end_of_list.} \]

And here is the proof from the output file:

\[ \text{--------------- PROOF ---------------} \]
1 \[ \neg P(x) \land Q(P(x)). \]
2 \[ \neg Q(P(a)). \]
3 \[ P(a). \]
4 \[ \text{[hyper,3,1]} \ Q(P(a)). \]
5 \[ \text{[binary,4,1,2,1]} \ \text{SF.} \]
\[ \text{----------- end of proof -----------} \]

OTTER doesn't bat an eyelash; hyperresolution and binary resolution kick in just fine. How can this be, given that the input is non-sensical? The answer is that OTTER interprets the first occurrence of \( P \) in

\[ \forall x (P(x) \rightarrow Q(P(x))). \]

to be a predicate symbol, but the second occurrence of \( P \) to be a functor. That is, \( P(x) \) gets interpreted as \( "x \ has \ property \ P," \) whereas \( Q(P(x)) \) get interpreted as \"the value of the function \( P \) taking \( x \) as argument has property \( Q." \)

OTTER will do the same thing with Def2 2. Here is a naïve translation of Def2 2 into OTTER, along with assumptions that the relevant conditions are instantiated for Dave and Selmer, and with the assumption for contradiction that Selmer doesn't betray Dave:

\text{set(auto).}
\text{formula_list(usable).}

% Defs-2 in OTTER:
\[ \forall x \forall y (\text{Betray}(x,y) \leftrightarrow \]
\[ (\exists z (\text{Wants}(y,z) \land \]
\[ \text{Believes}(x,\text{Wants}(y,z)) \land \]
\[ \text{Agrees}(x,y,z) \land \]
\[ \text{Intends}(x,z) \land \]
\[ \text{Believes}(x,\text{Believes}(y,\text{Intends}(x,z))))). \]

% Pretend facts of the case:
\[ \text{Wants(adave,agraduate).} \]

4.2. BETRAYAL CALLS FOR AN EXPRESSIVE LOGIC!

\text{Believes(aseselmer,Wants(adave,agraduate)).}
\text{Agrees(aseselmer,adave,agraduate).}
\text{Intends(aseselmer,agraduate).}
\text{Believes(aseselmer,Believes(adave,Intends(aseselmer,agraduate))).}

% Assumption for indirect proof:
\[-\text{Betray(aseselmer,adave)}. \]
\[ \text{end_of_list.} \]

And here is the resulting proof:

\[ \text{--------------- PROOF ---------------} \]
6 \[ \neg \text{Betray}(x,y) \land \neg \text{Wants}(y,z) \land \neg \text{Believes}(x,\text{Wants}(y,z)) \land \]
\[ \neg \text{Agrees}(x,y,z) \land \neg \text{Intends}(x,z) \land \]
\[ \neg \text{Believes}(x,\text{Believes}(y,\text{Intends}(x,z))). \]
7 \[ \neg \text{Betray}(adaveselmer,adave). \]
8 \[ \text{Wants(adave,agraduate).} \]
9 \[ \neg \text{Believes}(aseselmer,Wants(adave,agraduate)). \]
10 \[ \text{Agrees(aseselmer,adave,agraduate).} \]
11 \[ \text{Intends(aseselmer,agraduate).} \]
12 \[ \neg \text{Believes}(aseselmer,\text{Believes(adave,Intends(aseselmer,agraduate))).} \]
13 \[ \text{[hyper,12,6,8,9,10,11]} \ \text{Betray(aseselmer,adave).} \]
14 \[ \text{[binary,13,1,7,1]} \ \text{SF.} \]
\[ \text{----------- end of proof -----------} \]

So here we have a perfectly acceptable proof, apparently of the fact that Selmer betrays Dave. The important thing to note is that this proof is generated despite the fact that the knowledge-base in this case (composed by the formulas in the OTTER input file) doesn't express the facts we would like it to express. For example, we would like the knowledge-base to include a formula expressing the fact that Selmer believes that Dave believes that Selmer intends that Dave graduate. But the formula

\text{Believes(aseselmer,Believes(adave,Intends(aseselmer,agraduate))).}

really expresses something like the proposition that Selmer believes that the value of the function \text{Believes} applied to the argument \text{adave}, and the argument that is the value of the function \text{Intend} applied to the arguments \text{aseselmer} and \text{adave}. This proposition seems to literally make no sense. What sense does it make to say that someone believes that \( z \), where \( z \) is a name for some object? For example, what sense does it make to say that Clinton believes that Moscow?
CHAPTER 4. MATHEMATIZING BETRAYAL

What OTTER does with DefB 2 is actually something that any so-called "syntactic approach" in AI will do to verbs like believes, hopes, fears, and so on. For example, consider the syntactic approach to such verbs taken in Russell and Norvig's popular Artificial Intelligence [201]. In this approach, "mental objects," which are the things, for example, that we have beliefs about, are expressed as strings. So for example if Selmer believes that Dave believes that Brown is a villain, we might write something like this:

\[ B(s, B(d, (V(b)))) \]

where whatever is flanked by " " is a string, not a term or a subformula. The problem with this approach is that if as an AI researcher one takes oneself to be building a genuine mind (or to be contributing to the eventual building of a genuine mind), the approach is inadequate. The reason is simple: When I have a belief about you I don't have a belief about a string; I have a belief about you. And after all, some people have beliefs about things even though they can't write down strings in any language. Their communication may be exclusively oral. But on the syntactic view, such people can't have beliefs.

However, though it's true that syntactic approaches, such as the inferencing by OTTER displayed earlier, and Russell and Norvig's approach, fail to abide by the meaning of propositional attitudes like 'believes,' we find these approaches well-suited to BRUTUS, for reasons we explain in section 4.2.4. Astute readers can probably guess why we have this attitude.

4.2.3 Cervantes' Don Quijote for Skeptics

Some readers will be skeptical. They will insist that we are exploiting idiosyncratic aspects of narrative (i.e., iterated beliefs) in order to support the claim that truly representing standard narrative calls for non-standard logics. Such readers couldn't be more mistaken. They fall prey to a temptation that has seduced many, many in AI. The temptation is to search out and conquer those parts of a problem domain that are clearly computable (checkers, chess, medical diagnosis, natural language in straightforward newspaper stories, etc.). When AI confronts creativity, this temptation is deadly, because creativity pushes one immediately and unavoidably toward at least the possibility that here is something in the human sphere that exceeds computation. Since we intend to eventually engineer a system that, at least in terms of its output, competes head-on with human authors, we are forced to consider what human authors concoct; we are forced to consider this material in its full, uncompromising glory. The detective example was abstract; here is a specific, concrete example of such glory, one that goes all the way back to the very first modern novel: Don Quijote, penned by the literary genius Miguel De Cervantes. The example comes from Chapter 51 of the second book of Cervantes' immortal book. Sancho Panza is governor of an island and must preside as judge over some rather nasty cases, one of which is presented to him as follows:

My lord [Sancho Panza], a broad river separates the two parts of a single domain ... Now, there's a bridge over this river, and at one end there stands a gallows and a court building, in which four judges usually preside, applying the law formulated by the lord of this river, this bridge, and this entire realm, which runs as follows: "Anyone passing over this bridge, from one section of this domain to the other, must first declare under oath where he is coming from and where he is going, and if he swears truly, he shall be allowed to pass, but if he lies, he shall be hanged from the gallows standing nearby, without any appeal or reprieve allowed." ... Well, it happened, one day, that a man came and swore the required oath, saying among other things that he had come to be hanged on that gallows, and for no other purpose. The judges considered his oath, saying: "If we simply let this man cross the bridge, his oath will be a lie, and then, according to the law, he ought to die, but if we hang him, the oath he swore about being hanged on this gallows will be true, and then the same law decrees that he be allowed to cross over in peace." Please consider, my lord governor, your grace, what the judges should do with this fellow, for even now they remain anxious and unsure how to proceed, and, having been made aware of your grace's keen mind and sublime understanding, they have sent me to improve your grace to tell them how you view this singularly complicated and puzzling case. ([45], 629)
4.2. BETRAYAL CALLS FOR AN EXPRESSIVE LOGIC!

4.2.4 What Shall We Then Do?

We have established that the demands placed upon a logic-based story generation system by such literary themes as betrayal are severe. There really are only two types of moves that can be made in the face of this. One is to search out and embrace logical systems, and corresponding implementations, that would in theory give an artificial agent built upon them the ability to possess and manipulate formal structures that in some real sense capture the full meaning of betrayal in the human sphere. The second possible move is to engineer a system, on the strength of a "humbler" logic, that, because of the cleverness of the engineering, *appears* to understand the full meaning of betrayal. As you can probably guess, the move we make is the second one, which is why we said at the conclusion of section 4.2.2 that in our story generation work we embrace syntactic approaches to propositional attitudes — despite the demonstrable fact that such approaches are philosophically inadequate. (One of us has taken the first approach — finding a more powerful logic — in a slightly different context: In *What Robots Can and Can't Be* [40], the highly expressive class of quantified modal logics is used to express arguments about the foundations of AI.)

BRUTUS_1 is built with tools having no more expressive power than first-order logic, for BRUTUS_1, as we reveal in Chapter 6, is built using FLEX and Prolog. How can this be? Why haven't we used tools that are designed for greater expressivity? Why haven't we turned to systems designed to express such things as iterated beliefs? The answer is that for now we are content with FLEX and Prolog code in which iterated beliefs can be typed, for there is nothing to stop us from writing Prolog code that includes

\[ \text{believes(bringsjord, believes(reader, believes, ...?}, \]

even if this code doesn't correspond to any desirable formal semantic model. (In other words, we can use Prolog here as we used OTTER earlier to reason about Def 2.) More sophisticated incarnations of BRUTUS may well need to be engineered with more expressive knowledge representation and reasoning tools.

We now return to our quest for an acceptable formal definition of betrayal. We will express candidate definitions as the first two have been expressed: in an English-like language that is neutral on what particular logical system or knowledge representation and reasoning

---

92 CHAPTER 4. MATHEMATIZING BETRAYAL

Here, in a novel published in 1605, is a passage written in the belief that if readers believe that Sancho Panza believes that this would-be bridge-crooser has made a claim about the future (as opposed to merely making a claim about his plans), then these readers will grasp the fact that Sancho cannot possibly uphold the law in question.\(^8\) Many other such tricks are performed by Cervantes. For example, in the Prologue to the book, Cervantes himself appears, and says that he writes only by following the literary advice of a friend (who also makes an appearance in the Prologue). And in Part Two of the novel, Don Quijote comes upon characters who recognize the knight as the hero of Part One, which had been read widely!

The point is that even the very first modern novelist exploited techniques that cutting-edge AI, supposedly about to burst boldly into the new century, has little idea how to accurately represent, and *a fortiori* little idea how to get a machine to reason about. As we discussed in the Preface, it is therefore exceedingly peculiar that seemingly smart men like Hans Moravec would predict that within a few decades all our cleverness will be as debris in the wake of super-smart robots.

\(^8\)Cervantes has given us a maddening variation on the famous Liar Paradox, the simplest version of which is generated by the following sentence: "This sentence is false." Denote this sentence by \(\lambda\). We know, by the law of the excluded middle, that \(\lambda\) is either true or false. Suppose it's true. Then since it says that it's false, it is false; so if \(\lambda\) is true it's false. Suppose, on the other hand, that \(\lambda\) is false. Then since \(\lambda\) says that it's false, it's actually true. But then if \(\lambda\) is false it's true. So, \(\lambda\) is true if and only if it's false, which is a contradiction. The paradox isn't solved in the least by banning self-referential sentences, both because such sentences can be thoroughly meaningful and innocuous (e.g., "This sentence starts with the letter 'T'."), and because the paradox can be generated without such sentences. For example, let \(\lambda_1\) be the sentence to the left in the following table, and let \(\lambda_2\) be the sentence on the right.

<table>
<thead>
<tr>
<th>(\lambda_1)</th>
<th>(\lambda_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The sentence to the right is false.</td>
<td>The sentence to the left is true.</td>
</tr>
</tbody>
</table>

Here, \(\lambda_1\) is true if and only if it's false; contradiction redux. Even giving up the law of the excluded middle doesn't prevent the paradox from rising up: see the *Strengthened Liar Paradox* presented by Robert Martin in his Introduction to *Recent Essays on Truth and the Liar Paradox* [152].
system is used for actual implementation. If we do our job well, then
an acceptable definition, expressed in this manner, will be useful if
we sustain our “hackerish” use of first-order logic, and if, down the
road, we decide to move to more expressive frameworks.

4.3 On To More Sophisticated Accounts

Defn 2 is certainly better than its predecessor, but it too can be over-
thrown by a relatively simple counter-example: Suppose that Horace
wants President Clinton to make a trip to Moscow; and suppose as
well that Joe believes that Horace wants Clinton to make this trip,
and that Joe agrees with Horace that Clinton ought to go. However,
assume in addition that Joe intends that Clinton not go — but takes
no action toward that end. In this case it seems that since Joe does
nothing (relevant), even if Clinton fails to go, there is no betrayal in
the picture.

This suggests certain changes be made to clauses 4 and 5, namely:

Defn 3  Agent $s_r$ betrays agent $s_d$ iff there exists some state of affairs $p$
such that

1. $s_d$ wants $p$ to occur;
2. $s_r$ believes that $s_d$ wants $p$ to occur;
3. $s_r$ agrees with $s_d$ that $p$ ought to occur;
4′ there is some action $a$ which $s_r$ performs in the belief that thereby $p$ will not occur;
5′ $s_r$ believes that $s_d$ believes that there is some action $a$ which $s_r$ performs in the belief that thereby $p$ will occur.

At this point perhaps it’s safe to say we’re making genuine progress. But we’re not home yet, for the following example shows
down Defn 3:

Doris is a monomaniacal woman intent on nothing more than
crossing the street. You are to assume that Doris’ mental life, save
for this desire, is quite empty. Now suppose that a mugger agrees
with Doris that crossing the street is a good idea — but in “helping”
her across is only looking for an opportunity to slam Doris on the
head with a tire iron and dash off with her purse. In this case it
appears that Doris’ barren mentation precludes betrayal. Of course,
we can all agree that a crime has been perpetrated, but betrayal
doesn’t seem to be part of the scenario.

Now some may find it hard to grasp the notion of “barren men-
tation.” We urge such thinkers to substitute for the mugger case the
following variation. Dorrie is a dog. a collie to be exact; and Dor-
rine very clearly wants to cross the street. (Perhaps there is another
dog on the other side; Max, perhaps.) Jones, seeing what Dorrie
wants, and seeing that she has a collar, bends down, utters some of
those niceties we reserve for the canine population (“Oh, what a nice
dog you are, yes” — petting Dorrie — “yes, you’re a good dog.”),
and then “offers” to guide Dorrie across the street to Max. In reality,
Jones is a cruel person (worse than Cruella DeVille) who plans to get
Dorrie out in the road so that the collie can be run down by a truck.
Suppose that Jones pulls off his fiendish plan. Has Jones betrayed
Dorrie? No. And the reason, generally speaking, is that collies are
incapable of been “duperd”; they don’t have the right “mind set.”

It would seem that the problem can be remedied by adding to
Defn 3 an appropriate new clause:

Defn 4  Agent $s_r$ betrays agent $s_d$ iff there exists some state of affairs $p$
such that

1. $s_d$ wants $p$ to occur;
2. $s_r$ believes that $s_d$ wants $p$ to occur;
3. $s_r$ agrees with $s_d$ that $p$ ought to occur;
4′ there is some action $a$ which $s_r$ performs in the belief that thereby $p$ will not occur;
5′ $s_r$ believes that $s_d$ believes that there is some action $a$ which $s_r$
    performs in the belief that thereby $p$ will occur;
6 $s_d$ believes that $s_r$ intends that $p$ occur.

But things are not so easy; clause 6 would seem to be too strong:
Suppose that Jones tells Smith that he will work to get Smith elected
to some public office. Smith, knowing Jones’ mixed record on fol-
lowing through with “backroom” agreements, doesn’t believe that
Jones will work. (Smith does very much want Jones to live up to

8If you’re not convinced by the dog variation of the thought-experiment, sub-
stitute a lower animal, or even a simple robot capable — as its designers say —
of having wants, etc.
4.4 A Branch Point

At this point we hit a rather significant snag. In order to see it, consider the movie *The Man Without a Face*, starring Mel Gibson. Gibson plays McCloud, a man with a disfigured face and a shadowy past who befriends a younger boy. At no time does this boy agree with McCloud that no rumors ought to be spread about how McCloud came to be disfigured. They never have a discussion about the issue; not a word is said about it. However, the first part of the film gets its energy from the fact that the boy betrays McCloud — by spreading a heartless rumor about how McCloud came to acquire such a gruesome countenance. The audience knows this is betrayal, McCloud knows it's betrayal; and the boy knows it as well. But clause 3 is never satisfied. (Recall the similar defect in Def$_B$ B.)

So we have to face a complicating issue: It appears that an acceptable definition of betrayal must incorporate two possibilities, viz., (i) there may be no agreement between $s_r$ and $s_d$, in which case 6 should be replaced by

---

Chapter 4. Mathematizing Betrayal

his obligation, and Smith is in dire need of Jones' support.) Actually, Smith doesn't commit either way: He suspends judgment on whether or not Jones will follow through. Here most of us would still be inclined to say we have a case of betrayal, despite the fact that clause 6 isn't satisfied.

Perhaps the solution is to "tweak" 6, yielding:

Def$_B$ 5 Agent $s_r$ betrays agent $s_d$ iff there exists some state of affairs $p$ such that

1. $s_d$ wants $p$ to occur;
2. $s_r$ believes that $s_d$ wants $p$ to occur;
3. $s_r$ agrees with $s_d$ that $p$ ought to occur;
4. there is some action $a$ which $s_r$ performs in the belief that thereby $p$ will not occur;
5. $s_r$ believes that $s_d$ believes that there is some action $a$ which $s_r$ performs in the belief that thereby $p$ will occur;
6. $s_d$ wants that there is some action $a$ which $s_r$ performs in the belief that thereby $p$ will occur.

---

4.4 A Branch Point

- $s_d$ wants that there is no action $a$ which $s_r$ performs in the belief that thereby $p$ will not occur.

On the other hand, (ii) there may well be an agreement (as classic literary cases of betrayal reveal), in which case 6 should remain as is. Given this, we can build a disjunction into our definition:

Def$_B$ 6 Agent $s_r$ betrays agent $s_d$ iff there exists some state of affairs $p$ such that

1. $s_d$ wants $p$ to occur;
2. $s_r$ believes that $s_d$ wants $p$ to occur;
3. $s_r$ believes that $s_d$ wants $p$ to occur;
4. there is some action $a$ which $s_r$ performs in the belief that thereby $p$ will not occur;
5. $s_r$ believes that $s_d$ believes that there is some action $a$ which $s_r$ performs in the belief that thereby $p$ will occur.

Perhaps you'll agree that Def$_B$ 6 really isn't half bad. At this point in the process we were ourselves quite optimistic. And then, you guessed it: There arose yet another thought-experiment (devised by Dave, to Selmer's chagrin), one just as destructive as the previous ones:

Harriet Truism is the mother of a cocaine-addicted teenager, Billy. Billy realizes his secret stash is dry, and his desperation is so great that he goes to his own mother to ask if she can acquire some cocaine out on the streets of LA. Billy gives his mother the name of a dealer, Scrum; he tells her that Scrum will be coming around soon to make a delivery to Billy, but that he will surely die long before Scrum arrives. Harriet isn't sure what to do. Eventually, she agrees to obtain the coke for Billy, and sets out alone into the night, heading for the most dangerous part of town. However, Harriet's intention is to purchase cocaine from Scrum only to prevent the drug from reaching her son (who she plans to somehow get to professional help).

In this case Harriet is courageous, self-sacrificial, perhaps foolhardily, but she is not betraying Billy. And that presents a problem, because Def$_B$ 6's definition is satisfied in this case. The solution, it
would seem, is to move to a slightly more complicated construction, viz.,

Def 7 Agent \( s_r \) betrays agent \( s_d \) iff there exists some state of affairs \( p \) such that

1. \( s_d \) wants \( p \) to occur;
2. \( s_r \) believes that \( s_d \) wants \( p \) to occur;
3'. \((3 \land 6') \lor\)

6''. \( s_d \) wants that there is no action \( a \) which \( s_r \) performs in the belief that thereby \( p \) will not occur;

4'' there is some action \( a \) such that:

4''a \( s_r \) performs \( a \) in the belief that thereby \( p \) will not occur; and
4''b it's not the case that there exists a state of affairs \( q \) such that \( q \) is believed by \( s_r \) to be good for \( s_d \) and \( s_r \) performs \( a \) in the belief that \( q \) will not occur;

5' \( s_r \) believes at \( t_j \) that \( s_d \) believes that there is some action \( a \) which \( s_r \) will perform in the belief that thereby \( p \) will occur.

Def 8 7 seems pretty sharp — until one ponders a bit further. For starters, it seems entirely possible that one betray a dead person. (When alive, Jones asks Smith to bury him in a certain fashion, and toward that end gives Smith the necessary cash. After Jones dies, Smith spends the cash on his own hedonic pursuits, and leaves the body for others to deal with. Would we not unhesitatingly say in such a case that Smith has betrayed his dead friend?) Can cases like this one be handled? We think so. Indeed, to some degree our overall approach may be vindicated by the apparent ease with which Def 7 can be refined to handle betrayal of the dead: The trick is to include temporal parameters and appropriate tense changes:

Def 8 Agent \( s_r \) betrays agent \( s_d \) at \( t_k \) iff there exists some state of affairs \( p \) and \( \exists t_i, t_s, t_j \) \((t_i \leq t_s \leq t_j \leq t_k)\) such that

1. \( s_d \) at \( t_i \) wants \( p \) to occur;
2. \( s_r \) believes that \( s_d \) wants \( p \) to occur;
3'. \((3 \land 6') \lor\)

6''. \( s_d \) wants at \( t_k \) that there is no action \( a \) which \( s_r \) performs in the belief that thereby \( p \) will not occur;

4.5 Two Popular Objections

When we presented parts of this chapter at various conferences and at various colloquia, two objections were often raised against us. The first objection runs like this: "Why do you strive to find the one perfect definition? As you yourselves have noted, interesting narrative can involve instantiations of what you would regard to be defective accounts of betrayal. For example, your thought-experiment involving Harriet might make for a promising kernel of a story, despite the fact that Harriet doesn't betray her son." We answered this objection at the beginning of this chapter; that answer may make more sense now that the definitions have been presented, so we repeat it. The answer is that once we have arrived at what we regard to be the account of betrayal, one immune to counter-examples, there is no reason why the definition cannot be systematically modified (say, by dropping a clause), and then instantiated in connection with some knowledge-base in order to produce an interesting story. All such modifications should be reachable from the one solid definition, whereas if one stops short of reaching such a definition, any number of variants of the solid definition may be unobtainable. To put the point a bit more formally, if you have defined \( P \) to be true just in case \( Q_1 \land Q_2 \land \ldots \land Q_n \), then you can easily consider many concepts related to but short of \( P \), e.g., \( Q_3 \land \ldots \land Q_n \). But if you have stopped short, and have only a "defective" definition of \( P \), say \( Q_3 \land Q_4 \), then you can never consider the instantiation of \( Q_2 \land \ldots \land Q_n \).

The second objection we've heard is this one: "But do you really think that human authors reason with frameworks as complex as Def 8? They do no such reasoning. What you have produced in this chapter, therefore, is psychologically unrealistic." Of course, this objection reflects an utter lack of understanding of what the BRUTUS
4.6 Miles to Go Before We Sleep

Unfortunately, there is a problem with Def$_B$ 8: it doesn't deal with a case we touched upon earlier: the case in which a parent betrays a very young child, a child too young to instantiate the psychological properties ascribed to $s_d$ in this definition. The solution here is no doubt in the direction of considering what $s_d$ would want and believe if certain things (e.g., normal physical and psychological development) were to come to pass. Such a solution, if fully formalized, would require a formalism sensitive enough to represent sophisticated conditionals. Of course, this comes as no surprise. For a fully formalized version of Def$_B$ 8, one that doesn't try to capture only the surface structure of this definition in some variant of first-order logic, would alone require the machinery of not only doxastic logic (for beliefs, as we discussed above), but also

- Logics of action, deliberate action, intending, etc.
- Temporal logic
- Conditional logic
- Deontic logic

4.6. MILES TO GO BEFORE WE SLEEP

In the future we expect to move toward such logics. For now, we rest content with Def$_B$ 8 as part of the BRUTUS architecture, and with mapping it to an implementation in Prolog and FLEX. This mapping will be discussed in Chapter 6. As we indicated previously, after betrayal, our plan is to proceed to the concept of self-deception. BRUTUS$_1$ currently works with a rough-and-ready account of self-deception (due essentially to a less formal account propounded by Harold Sackheim and Ruben Gur [202]) that can be set out via three conditions:

Def$_{SD}$ 1 $s$ is self-deceived iff there is a proposition $p$ and a time $t$ such that

1. $s$ both believes at $t$ that $p$ and believes at $t$ that $\neg p$;
2. $s$ is unaware of one of the beliefs in clause 1;
3. There is an action $a$ which determines which of the beliefs in question is subject to $s$'s awareness, and $a$ is motivated.

Though as Alfred Mele [158] has recently pointed out, the empirical evidence traditionally taken to show that Def$_{SD}$ is routinely instantiated in humans is questionable, the fact remains that Def$_{SD}$ 1 makes for some interesting narrative, and that is good enough for BRUTUS.\footnote{Interestingly enough, Def$_{SD}$ 1 is affirmed largely due to the fact that certain psychologists can use it to tell a coherent story about what is happening in certain experiments. See [158].}

We conclude by presenting two more stories from BRUTUS$_1$'s repertoire. This first is one in which the account of self-deception

\footnote{This is as good a place as any to mention another literary theme with which we have been experimenting: evil. Currently we are working with a formalization of M. Scott Peck's [175] characterization of evil, which consists in a person exemplifying four properties, viz.,

1. consistent destructive, scapegoating behavior, which may often be quite subtle;
2. excessive, albeit usually covert, intolerance to criticism and other forms of narcissistic injury;
3. pronounced concern with a public image and self-image of respectability, contributing to a stability of lifestyle but also to pretentiousness and denial of hateful feelings or vengeful motives; and finally,
4. intellectual deviousness, with an increased likelihood of a mild schizophrenic-like disturbance of thinking at times of stress.}

\begin{itemize}
  \item consistent destructive, scapegoating behavior, which may often be quite subtle;
  \item excessive, albeit usually covert, intolerance to criticism and other forms of narcissistic injury;
  \item pronounced concern with a public image and self-image of respectability, contributing to a stability of lifestyle but also to pretentiousness and denial of hateful feelings or vengeful motives; and finally,
  \item intellectual deviousness, with an increased likelihood of a mild schizophrenic-like disturbance of thinking at times of stress.
\end{itemize}
CHAPTER 4. MATHEMATIZING BETRAYAL

isn't instantiated. Once again, note that the point is that these stories can be generated by BRUTUS courtesy of formalisms that we have crafted. BRUTUS did not and does not originate the following stories.

"Simple Betrayal" (no self-deception; conscious)

Dave Striver loved the university. He loved its ivy-covered clocktowers, its ancient and sturdy brick, and its sun-splashed verdant greens and eager youth. He also loved the fact that the university is free of the stark unforgiving trials of the business world — only this isn't a fact: academia has its own tests, and some are as merciless as any in the marketplace. A prime example is the dissertation defense: to earn the PhD, to become a doctor, one must pass an oral examination on one's dissertation.

Dave wanted desperately to be a doctor. But he needed the signatures of three people on the first page of his dissertation, the priceless inscriptions which, together, would certify that he had passed his defense. One of the signatures had to come from Professor Hart.

Well before the defense, Striver gave Hart a penultimate copy of his thesis. Hart read it and told Striver that it was absolutely first-rate, and that he would gladly sign it at the defense. They even shook hands in Hart's book-lined office. Dave noticed that Hart's eyes were bright and trustful, and his bearing paternal.

At the defense, Dave thought that he eloquently summarized Chapter 3 of his dissertation. There were two questions, one from Professor Rodman and one from Dr. Teer; Dave answered both, apparently to everyone's satisfaction. There were no further objections.

Professor Rodman signed. He slid the tome to Teer; she too signed, and then slid it in front of Hart. Hart didn't move.

"Ed?" Rodman said.

Hart still sat motionless. Dave felt slightly dizzy.

"Edward, are you going to sign?"

Later, Hart sat alone in his office, in his big leather chair, underneath his framed PhD diploma.

The second story is based on an instantiation of the concept of self-betrayal:

4.6. MILES TO GO BEFORE WE SLEEP

"Self-Betrayal" (no self-deception; conscious version)

Dave Striver loved the university — at least most of the time. Every now and then, without warning, a wave of ... well, it was true: a wave of hatred rose up and flowed like molten blood through every cell in his body. This hatred would be directed at the ghostly gatekeepers. But most of the time Striver loved — the ivy-covered clocktowers, the ancient and sturdy brick, and the sun-splashed verdant greens and eager youth who learned alongside him. He also loved the fact that the university is free of the stark unforgiving trials of the business world — only this isn't a fact: academia has its own tests, and some are as merciless as any in the marketplace. A prime example is the dissertation defense: to earn the PhD, to become a doctor, one must pass an oral examination on one's dissertation.

Dave wanted desperately to be a doctor. He had been working toward this end through six years of graduate school. In the end, he needed the signatures of three people on the first page of his dissertation, the priceless inscriptions which, together, would certify that he had passed his defense. One of the signatures had to come from Professor Hart.

Well before the defense, Striver gave Hart a penultimate copy of his thesis. Hart read it and told Striver that it was absolutely first-rate, and that he would gladly sign it at the defense. They shook hands in Hart's book-lined office. Hart's eyes were bright and trustful, and his bearing paternal.

"See you at 3 p.m. on the tenth, then, Dave?" Hart said.

At the defense, Dave eloquently summarized Chapter 3 of his dissertation. His plan had been to do the same for Chapter 4, and then wrap things up, but now he wasn't sure. The pallid faces before him seemed suddenly nauseating. What was he doing?

One of these pallid automata had an arm raised.

"What?" Striver snapped.

Striver watched ghostly eye at each other. A pause.

Then Professor Teer spoke: "I'm puzzled as to why you prefer not to use the well-known alpha-beta minimax algorithm for your search?"

Why had he thought so earnestly about inane questions like this in the past? Striver said nothing. His nausea grew. Contempt, fiery and uncontrollable, rose up.
Chapter 5

The Narrative-Based Refutation of Church’s Thesis

5.1 Can Interestingness Be Formalized?

As Chapter 4 makes plain, we have set ourselves the goal of capturing in logic such lofty literary themes as unrequited love, and of then implementing this logic in a theorem-prover-based story generator. Some will find this to be a laughably unreachable goal, despite our success with betrayal and BRUTUS. Actually, there was a time, about a decade ago, when one of us (Bringsjord) thought that perhaps the entire ball game could be won using something like this method. His idea was this: We would figure out what makes a story interesting from the standpoint of logic, proceed to formalize interestingness in some logical system, and then — as in the case of our plans for betrayal and other literary themes — code this formalization so as to yield a program that generates interesting stories. This idea proved to be a painful dead end. (Many of you are probably not surprised.) Whereas in the case of betrayal, counterexamples to proposed definitions of the concept led the way to improved definitions that promised to eventuate in the correct definition, nothing of the sort happened in dialectic at Rensselaer about what conditions are necessary and sufficient for a story’s being interesting. In fact, it began to occur to Bringsjord that perhaps the set of all interesting