sun, and glimpsed the full perfect line of her arm and shoulder" will probably catalyze a number of images.

5. Readerly imaging can be particularly intense when voyeurism is involved.28

In our final Chapter, 6, we explain how points 1–4 are reflected in the BRUTUS architecture, and partially implemented in BRUTUS1, and we experience the narrative made possible by this architecture.

Chapter 3

Consciousness and Creativity

What is Brutus sick,
And will he steal out of his wholesome bed,
To dare the vile contagion of the night
And tempt the rheumy and unpurged air
To add unto his sickness? No, my Brutus;
You have some sick offence within your mind.
—Portia, in Julius Caesar

Given our objectives, the problem of consciousness for us, put colorfully, is this: How can we engineer things so that a certain zombie — BRUTUS1 — and, in fact, all future incarnations of the BRUTUS architecture — becomes a member of the literati? A less lively form of the question is: How can we build BRUTUS1 so that despite his lack of consciousness he can generate stories rich in consciousness — so rich that those who read them are inclined to ascribe genuine states of mind to BRUTUS1? Or put yet another way: How can something lacking a mind write compelling fiction about minds? (What Portia says about Shakespeare's Brutus — that his mind is afflicted — cannot be said about BRUTUS or BRUTUS1: they have no minds!) In this chapter we explain why we see these questions as the toughest ones facing those who aim to build genuinely creative agents (in the literary realm), and we provide the first phase of an answer. The second phase, more closely tied to actual implementation, is presented in Chapter 6.

28Another wonderful example can be found in Kafka’s unforgettable “Penal Colony.”

Can you follow it? The Harrow is beginning to write; when it finishes the first draft of the inscription on the man’s back, the layer of cotton wool begins to roll and slowly turns the body over, to give the Harrow fresh space for writing. Meanwhile the raw part that has been written on lies on the cotton wool, which is specially prepared to staunch the bleeding and so makes all ready for a new deepening of the script. Then these teeth at the edge of the Harrow, as the body turns further round, tear the cotton wool away from the wounds, throw it into the pit, and there is more work for the Harrow. ([123], p. 203)
3.1 BRUTUS as Zombie

When we say that BRUTUS is a zombie, we refer to philosophers’ zombies — not those creatures who shuffle about half-dead in the movies.¹ Philosophers’ zombies star in situations lifted directly out of the toolbox most philosophers of mind, today, carry with them on the job: Let’s imagine that due to cancer your brain starts to deteriorate and the doctors, desperate, replace it, piecemeal, with silicon chip workalikes, until there is only silicon inside your refurbished cranium.² John Searle claims that at least three distinct possibilities arise from this gedanken-experiment:

The Smooth-as-Silk Variation The complete silicon replacement of your flesh-and-blood brain works like a charm: same mental life, same sensorimotor capacities, etc.

The Zombie Variation “As the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are indeed losing control of your external behavior ... You have become blind, but you hear your voice saying in a way that is completely out of your control, 'I see a red object in front of me.' ... We imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same” ([206], pp. 66-67).

The Curare Variation Your body becomes paralyzed and the doctors, to your horror, give you up for dead.³

It’s the zombie variation that gives us the sort of creature BRUTUS, and his successors would seem to be: creations whose observable behavior is as impressive (at least in some domains) as that displayed by Homo sapiens sapiens, but whose “inner lives” are no different than those of common, ordinary rocks. Zombies act clever, but underneath there is no — to use Ned Block’s [15] recent term — P-consciousness. Here is part of Block’s explication of this term:

¹Actually, the zombies of cinematic fame apparently have real-life correlates created with a most creepy mixture of drugs and pre-death bruising; see [63], [29].
²For example, the toolbox is opened and the silicon supplantation elegantly pulled out in [55].
³This scenario would seem to resemble a real-life phenomenon: the so-called “Locked-In” Syndrome. See [181] (esp. the fascinating description on pp. 24-25) for the medical details.

So how should we point to P-consciousness? Well, one way is via rough synonyms. As I said, P-consciousness is experience. P-conscious properties are experiential properties. P-conscious states are experiential states, that is, a state is P-conscious if it has experiential properties. The totality of the experiential properties of a state are “what it is like” to have it. Moving from synonyms to examples, we have P-conscious states when we see, hear, smell, taste and have pains. P-conscious properties include the experiential properties of sensations, feelings and perceptions, but I would also include thoughts, wants and emotions. ([15], p. 230)

Block distinguishes between this brand of consciousness and A-consciousness; the latter concept is characterized as follows:

A state is access-conscious (A-conscious) if, in virtue of one’s having the state, a representation, of its content is (1) inferentially promiscuous, i.e., poised to be used as a premise in reasoning, and (2) poised for [rational] control of action and (3) poised for rational control of speech. ([15], p. 231)

As one of us has explained elsewhere [27], it’s plausible to regard certain extant, mundane computational artifacts to be bearers of A-consciousness. For example, theorem provers with natural language generation capability, for that matter perhaps any implemented computer program, would seem to qualify by Block’s definition.⁴ It follows from this, of course, that our very own BRUTUS is

⁴It follows that a zombie would be A-conscious. In [27] one of us (Bringsjord) argues that because (to put it mildly here) it is odd to count (say) ordinary laptop computers running run-of-the-mill Pascal programs as conscious in any sense of
3.2 The Conundrum

The conundrum is this: The construction of the genuine article (a genuine artificially intelligent storyteller) is possible, it seems, only if central human authorial techniques can be formalized in computational terms, but at least some of these techniques, to put it mildly, resist formalization. The one technique we’re especially concerned with is an author’s ability to adopt the point of view of his or her characters, to “feel what they feel on the inside.” Henrik Ibsen used this technique as often and as intensely as any writer. We heard from Ibsen in the previous chapter; let’s listen again: “Before I write down one word,” Ibsen said,

I have to have the character in mind though and through, I must penetrate into the last wrinkle of his soul. I always proceed from the individual; the stage setting, the dramatic ensemble, all that comes naturally and does not cause me any worry, as soon as I am certain of the individual in every aspect of his humanity. But I have to have his exterior in mind also, down to the last button, how he stands and walks, how he conducts himself, what his voice sounds like. Then I do not let him go until his fate is fulfilled. (reported in [90], p. xiv)

Suppose for the sake of argument that AI researchers can build a bona fide storyteller capable of Ibsen’s approach. Then, on the plausible principle that “if X adopts Y’s point of view, X is itself capable of having a point of view,” it would seem to follow that computers can have genuine conscious states — there can be something, to use Thomas Nagel’s [167] phrase, it’s like to be a computer, something it feels like to be a storytelling AI. But, as we discussed in the previous chapter, a number of thinkers ([115], [139], [167], [206], [40], [37]) have developed and endorsed a line of argument, the conclusion of which is that it’s impossible that inner, mental states of the human variety be formalized in any way. It seems a short step from this conclusion to the view that computers — which are themselves,

---

the term, ‘A-consciousness’ ought to be supplanted by suitably configured terms from its Blockian definition.

3.2. The Conundrum

mathematically speaking, formalizations — are incapable of having a point of view.

An informal version of the argument in question was presented in the previous chapter. At its heart was a thought-experiment involving Alvin, a scientist who works alone in a laboratory and then emerges to have the burst of P-consciousness generated by meeting a long lost friend for the first time. The argument, recall, runs as follows.

\[ \text{Arg}_3 \]

\begin{align*}
(1) & \quad \text{To know everything knowable about a psychological state is to have complete first- and third-person knowledge of it.} \\
(2) & \quad \text{Alvin, prior to his first first-person long-lost-friend experience, knows everything knowable about meeting long lost friends from a third-person symbolic perspective.} \\
(3) & \quad \text{To know everything knowable about meeting long lost friends from a first-person perspective implies knowing what it’s like to meet a long lost friend in the flesh.} \\
(4) & \quad \text{Alvin, prior to his first-person long-lost-friend experience, doesn’t know what it’s like to meet a long lost friend in the flesh.} \\
(5) & \quad \text{If what-it’s-like-to-X information is captureable in some symbolic representation scheme, then [if Alvin, prior to his first-person long-lost-friend experience, knows everything knowable about meeting long lost friends from a third-person symbolic perspective, then, prior to his first first-person long-lost-friend experience, he knows everything knowable about meeting long lost friends].} \\
(6) & \quad \text{Alvin, prior to his first-person long-lost-friend experience, doesn’t know everything knowable about meeting long lost friends.} \\
(7) & \quad \text{What-it’s-like-to-X information isn’t captureable in some symbolic representation scheme.} \\
\end{align*}
This argument is obviously formally valid (as can be shown when it's symbolized in first-order logic). The only formidable objection we know of comes from Daniel Dennett [70]. In a nutshell, Dennett thinks that Arg's (1), despite strong intuitions to the contrary, is unimaginable, that, at best, we can reasonably affirm only

(2') Alvin, prior to his first first-person long-lost-friend experience, knows everything currently knowable about meeting long lost friends from a third-person computational perspective.

One of us has treated Dennett's position on these matters elsewhere [37]. Jacquette [116] disarms Dennett's attack in a different but no less effective manner; here we encapsulate this treatment. First, it must be conceded that the argument arising from supplanting (2) with (2') is unsound: this much is uncontroversial. The reader should satisfy himself or herself that (5') is false, or at least such that there is no apparent reason whatever to affirm it. (Note that we refer to the "primed" versions of (1) through (7): They are the result of propagating the change indicated in (2') through the rest of the argument.) Now let's take a look at Dennett's own words on the matter (adapted so as to apply to Arg):

The image [of Alvin] is wrong; if [seeing Alvin make a discovery] is the way you imagine the case, you are simply not following directions! The reason no one follows directions is because what they ask you to imagine is so preposterously immense, you can't even try. The crucial premise is that "[Alvin] has all the physical [computational] information." This is not readily imaginable, so no one bothers. They just imagine that [he] knows lots and lots — perhaps they imagine that [he] knows everything that anyone knows today about the neurophysiology etc. of [such psychological states]. But that's just

5 Actually, a formalization of Arg in first-order logic would have to be considerably naïve. The first premise might be symbolized as

\[ \forall x \forall y (sK^4 \phi \implies (sK^1 \phi \land sK^1 \phi)). \]

This would be to deploy separate three-place predicates — for knowing everything about something, for knowing everything about something from the first-person perspective, and for knowing everything about something from the third-person perspective. The core rule of inference (used twice) would be modus tollens. Quantifier manipulation would involve existential instantiation and universal elimination. One sophisticated way to formalize Arg would involve knowledge operators rather than predicates.

3.2. THE CONUNDRUM

a drop in the bucket, and it's not surprising that [Alvin] would learn something if that were all [he] knew. ([70], p. 399)

Dennett has been conveniently struck by an uncharacteristic failure of imagination. Why does he say that Alvin's having all the information in question cannot be imagined? He provides no argument for this view; and the absence of such an argument is especially odd because Dennett's position seems flatly inconsistent with logical and mathematical knowledge and reasoning: How can it be that in logic and math we conceive of situations involving infinite information, but suddenly when it comes to Alvin we cannot, even in the realm of thought and thought-experiment, arm him with enough knowledge? The thought-experiment in question, and the corresponding argument, can both be put in austere terms wholly divorced from the particular sciences of today — terms no different, really, than those involved in formal pursuits: Imagine that Alvin works with some fixed, austere model for computation and/or symbolic representation — say that provided by Turing machines. Accordingly, for Alvin to have complete third-person knowledge about meeting long lost friends is for him to have assimilated some collection of Turing machines. What, pray tell, is unimaginable about Alvin assimilating n Turing machines, for any natural number n? (Is there anything unimaginable about a person counting up to n, for all n? Given that TMs can be coded as natural numbers, if the answer to this question is an affirmative one, it would seem to follow immediately that Alvin, in principle, can pull things off.) We would even go so far as to say — though for purposes of emerging victorious from the present dialectic we needn't — that it seems perfectly imaginable that Alvin assimilate an infinite number of TMs (perhaps he works faster and faster in the fashion of so-called "Zeus machines" [21]).

Given Arg, the argument that expresses the problem we face can also be put in deductive form:

6 In connection with this issue, see [29].
CHAPTER 3. CONSCIOUSNESS AND CREATIVITY

ARG₄

(8) If computers (or computer programs, or artificial agents, ...) can adopt the points of view of creatures of fiction, then computers can have genuine “there’s something it’s like to be a” conscious states.

(9) If computers can have genuine “there’s something it’s like to be a” conscious states, then such states are formalizable in computational terms (= in third-person symbolic representation schemes).

(10) “There’s something it’s like to be a” conscious states are not formalizable in computational terms.

\[ \text{from ARG₃} \]

(11) Computers can’t adopt the points of view of creatures of fiction.

.(12) If computers can’t adopt the points of view of creatures of fiction, they can’t generate sophisticated fiction.

(13) Computers can’t write sophisticated fiction.

(11), (12)

This chain of reasoning, like its predecessor, is formally valid. But are the premises true? Premise (9) would seem to be uncontroversial; it could presumably be defended with help from a formal account of computerhood (which one of us has developed: [31]). Premise (8), it would seem, is plausible (in large part because we accept the principle cited earlier, viz., “if X adopts Y’s point of view, X is itself capable of having a point of view”). Premise (10), again, is the conclusion of ARG₃ — an argument which may be vulnerable, but is surely not easily dodged, as Dennett will hopefully learn. This leaves premise (12). Why is this proposition plausible?

In short, (12) seems plausible because compelling narrative, at least in this day and age, and in this culture (and we certainly find ourselves inescapably designing BRUTUS and building BRUTUS₁ both in the present and in this culture), needs to have both a “landscape of action” and a “landscape of consciousness” [44]. The former is typically composed of temporally sequenced “action events reported in the third person with minimal information about the psychological states of the protagoniste” ([87], p. 2). In the landscape of action, there is little concern about how things are perceived, felt, intended, and so on; “things happen or they don’t” ([87], p. 2). In

3.2. THE CONUNDRUM

the landscape of consciousness things are quite different: Narrative is driven in large part by how the cast of characters feels about the world from their own points of view. Most successful modern fiction (again, in our culture) generates both types of landscapes, with an emphasis on the landscape of consciousness. Some narrative shrinks the landscape of action down to nearly nothing (as in, e.g., Joyce and Beckett). And some narrative, for example folktales, generates next to nothing in the realm of consciousness [186].

So, if some such proposition as

(14) Sophisticated fiction requires both a landscape of consciousness and a landscape of action.

is true, and if it’s true as well that

(15) If computers can’t adopt the points of view of creatures of fiction, they can’t generate a landscape of consciousness.

(12) follows immediately.

Why might some resist this argument? Well, it might be said that (15) is false because computers can generate sophisticated fiction without adopting the points of view of the characters in that fiction. Perhaps (so the story goes) storytelling AIs will use some other technique to generate narrative with the same “depth” as, say, Ibsen’s immortal A Doll House.

This reaction would seem to be fueled by some rather wishful thinking — because when one ponders the power of the knowledge-based techniques by which a computer might write “three-dimensional” characters without anything as mysterious as adopting a character’s point of view, there is little reason to believe that these techniques will rival the tried-and-true one of which Ibsen is a consummate practitioner. And yet here we stand, dreaming of giving a zombie, BRUTUS₁ (or at least BRUTUSₙ), the ability to create a landscape of consciousness. How in the world are we going to pull it off?

It’s important to be clear on how difficult the problem is. One reason it’s so difficult is that our goals seem to be unreachable in the absence of an understanding of the nature of creatures of fiction. It’s hard to see how one can engineer a machine with the capacity to occupy the point of view of a creature of fiction if one doesn’t know what a creature of fiction is: this is just plain old commonsense. So how does BRUTUS₁, and artificial agents generally, have a chance
if the human engineers of such systems fail to grasp the nature of creatures of fiction? Well then, what are creatures of fiction? How do they differ from us? In what sense are they unreal, exactly — given that we can have all sorts of utterly determinate beliefs about them? (Is it not true, e.g., that Sherlock Holmes is brilliant, and that Brutus underestimated the oratorical skills of Mark Antony in Julius Caesar, and that Dave Striver is betrayed in BRUTUS’s stories about him, and . . .?) If nothing else, if one could come to understand the nature of creatures of fiction, then it would at least seem that one would be in a better position to evaluate the objection that a computer can write three-dimensional characters without slipping into the shoes of these characters. Unfortunately, as things stand today, no one knows what the ontological status of creatures of fiction is — or at least there’s no consensus about their status. There are proposals. The most plausible account of fictional characters we know of is Bill Rapaport’s [192], which builds upon four prior, influential accounts ([241], [173], [147], [46]) and attempts, specifically, to indicate the nature of an AI’s beliefs about creatures of fiction. Rapaport’s seminal work is carried out in the context of a well-known knowledge representation and reasoning system — SNePS — with which he has been intimately involved for a long while. The basic idea behind this work is to try to solve some difficult problems that arise by supposing that a computational mind — CASSIE — has beliefs about not only the “real” world, but about fictional worlds as well. Unfortunately, from the perspective of Ibsen’s threat to story generation, CASSIE doesn’t provide much hope. This is so because though CASSIE promises to offer, among other things, unprecedented ways of keeping an AI’s beliefs about fiction and reality from getting muddled, the objects of CASSIE’s beliefs, at bottom, are symbolic entities, formulas in a powerful intensional logic (SNePS, however, is traditionally displayed as a network; we interact with it through SNePSLOG, a logic-based interface). And this brings us straight back to Alvin. For how can command over purely symbolic information, whether that information is represented by Turing machines or by formulas in some logic, give to its owner the Ibsenian power to get inside the soul of a character? The problem is exceedingly difficult.

Let’s take a closer look at the problem.

Suppose CASSIE has beliefs in 1997 about Bill Clinton and Tolstoy’s Count Vronsky, that Clinton is overweight and that Vronsky is dashing, and countless others about this unlikely duo. Suppose that CASSIE’s having these beliefs is the presence of certain formulas in CASSIE’s belief base. So picture a list

\[ L = \phi_1, \phi_2, \phi_3, \ldots \]

of some formulas in some logic. Now suppose that you assimilate \( L \). Would your new knowledge suffice to allow you to occupy Vronsky’s point of view, so that you could add an adventure to Tolstoy’s Anna Karenina? With all due respect, we doubt it — because despite having digested \( L \), you will need to do something else to write a scene suitable for inclusion in perhaps the greatest landscape of consciousness ever created: You will have to try — armed with \( L \), we readily agree — to feel what it’s like to be the Count. In other words, what you get by virtue of grasping \( L \) is the sort of declarative information an author can impart to you about one of his or her characters, but you don’t thereby get the ability to adopt the point of view of one of these characters (though of course you may have that ability for reasons independent of \( L \)).

What we have just described schematically can be fleshed out and experimented with. We urge you to try it for yourself. Have a friend give you a list of propositions about some person who you don’t know. Your friend can follow the templates we follow for story generation, which are enhancements of ones appropriated from the movie industry. In these templates (which all aspiring scriptwriters are taught to use when starting out), you need to write down a person’s age, their educational background, a physical description, where they grew up, their short- and long-term goals, their hobbies, their occupation, where and what they studied in school, their friends (and the occupations and avocations of these friends), key personality traits (and here there are a host of instruments from psychology that are useful), the five best things that have happened to them and the five worst things that have happened to them, their family members (and the occupations and avocations of these people), and so on. Now take this information from your friend and assimilate it. Now try to write a good short story about the person that is described. We predict that one of two things will happen. Either you will write an exceedingly poor story that trades mechanically directly off of the information you have absorbed, or you will take respectable steps toward impressive narrative. If you manage to take the second route, you will have used the declarative information
given you to trigger in your mind that which it's like to X, where X might be 'have a father who was a brain surgeon.' We can't for the life of us see how a machine could travel this route. BRUTUS$^1$ takes the first, shallow route — on the strength of declarative information upon which it **mechanically** capitalizes.

So, once again, it seems as if Alvin's lesson is a knife that is being driven ever deeper into our hopes of building an AI that produces sophisticated narrative: Having a point of view certainly seems to be something over and above having symbols on hand, so having a point of view (not to mention having the capacity to occupy another's point of view) is something we'll have trouble imparting to a computer — which after all is nothing more than a device for manipulating symbols.

Are we paralyzed, then? No. Given that BRUTUS$^1$, is to be **believed** to be as creative as Ibsen, rather than literally being another Ibsen, there are moves to be made in the face of the problem we have described. Unsurprisingly, these moves all derive from a careful study of those texts which evoke landscapes of consciousness in the minds of readers. (This chapter, you now see, ends on the same note as did the previous one.) These texts are invariably marked by heavy usage of mental verbs — thinking, supposing, intending, knowing, feeling, fearing, believing, hating, betraying, etc.\(^7\) In fact, there is a body of fascinating research that grows out of minimizing and maximizing the landscape of consciousness in a given story by adjusting the frequency and intensity of mental verbs \([87]^{8}\). Our methods for implementing the calculated modulation of mental verbs (and other linguistic objects) in order to allow BRUTUS$_1$ to successfully confront the problem expressed in this chapter are described in Chapter 6.

We end this chapter with a "non-conscious" version of the story presented in Chapter 1. Readers may find it profitable to compare it with the "conscious" version seen earlier. The difference is that in

\(^7\)In *Julius Caesar*, the seeds of betrayal are cultivated by Cassius in Brutus by the *action of sending to Brutus a note that sparks the right sort of consciousness*. The note is quite brief: "Brutus, thou sleep'st: awake and see thyself. Shall Rome, etc. Speak, strike, redress..."

\(^8\)In one of these studies, for example, two versions of Joyce's story "Eveline" [120] are presented to subjects, the original one, and one (the so-called "non-conscious" version) whose landscape of consciousness is contracted by deleting and downplaying mental verbs.

3.2. THE CONUNDRUM

the non-conscious version the frequency of mental verbs is decreased. Again, know well that this is a change that we make in BRUTUS$_1$; the change does not originate with

"**Betrayal in Self-Deception**" (non-conscious)

Dave Striver loved the university — its ivy-covered clocktowers, its ancient and sturdy brick, and its sun-splashed verdant greens and eager youth. The university, contrary to popular opinion, is far from free of the stark unforgiving trials of the business world: 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. This was a test Professor Edward Hart enjoyed giving.

Dave wanted 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, and Hart had often said — to others and to himself — that he was honored to help Dave secure his well-earned dream.

Well before the defense, Striver gave Hart a penultimate copy of his thesis. Hart read it and told Dave 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 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 looked at him.

"Edward, are you going to sign?"

Later, Hart sat alone in his office, in his big leather chair. He tried to think of ways he could help Dave achieve his goal.