Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine

Selmer Bringsjord and David A. Ferrucci
(Rensselaer Polytechic Institute and IBM T.J. Watson Research Center)

Mahwah, NJ: Lawrence Erlbaum
Associates, 2000, xxxii+230 pp;
hardbound, ISBN 0-8058-1986-X, $59.95;
paperbound, ISBN 0-8058-1987-8, $27.50

Reviewed by
Ronald de Sousa
University of Toronto

BRUTUS is a program that tells stories. The stories are intriguing, they hold a hint of mystery, and—not least impressive—they are written in correct English prose. An example (p. 124) is shown in Figure 1. This remarkable feat is grounded in a complex architecture making use of a number of levels, each of which is parameterized so as to become a locus of possible variation.

The specific $BRUTUS_1$ implementation that illustrates the program's prowess exploits the theme of betrayal, which receives an elaborate analysis, culminating in a set

Betrayal in Self-Deception

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 Ph.D., 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 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, 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 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, saddened by Dave's failure. He tried to think of ways he could help Dave achieve his dream.

Figure 1
An example of a story by $BRUTUS_1$.

of necessary and sufficient conditions (Def_B 8, pp. 98–99) in terms of actions, goals, and beliefs of two agents, the betrayer and the betrayed. This illustrates the sort of **thematic knowledge** in which BRUTUS's stories are grounded. Thematic knowledge is part of the **knowledge level**, which also comprises the following:

1. More general **domain knowledge** pertaining to the kinds of things that may constitute the subject matter of stories: agents, events, beliefs, goals, actions, and reaction (pp. 175–179);
2. **Linguistic knowledge**, pertaining to morphology, syntax, paragraph, and discourse structure (pp. 167, 180–182);
3. **Literary knowledge**, incorporating principles of storytelling, including story-grammars, designed to achieve specific literary objectives such as triggering imaging in the reader, causing the reading to project personal consciousness onto the characters. Literary knowledge also includes rhetorical tropes, evaluative valences, analogies, and image associations (pp. 182–185);
4. A special level of **literary augmented grammars** (LAGs) brings the rhetorical knowledge of the literary knowledge level to bear on the syntax controlled by the linguistic level (pp. 185–189).

Story generation uses the knowledge of the kinds just listed in four types of developments through time at the **process level**:

1. **Thematic concept instantiation** sets the "stage" for a given story, exploiting thematic knowledge for a chosen theme;
2. **Plot generation** takes the characters and characteristics identified in the "stage"-setting phase and generates a scenario, using domain and thematic knowledge, particularly of action and reactive behavior. It results in a detailed scenario that generates consequences inferred from the knowledge level by production rules, and deploys a temporal sequence of events. The production rules follow ordinary first-order logic, but extensions are promised making use of temporal logic, conditional logic, and deontic logic, as well as "logics of action, deliberate action, intending etc." (p. 100).
3. **Story structure expansion** takes place on the basis of story grammars, at each choice-point of which sentence-types are produced.
4. Linguistic and literary knowledge are then used for **language generation**, that is, the production of specific linguistic structures down to the level of sentences, phrases, and words (pp. 194–197).

An example of a production rule used in the sample story is as follows (p. 179):

```
Rule committee_members_behavior
  IF
    Candidate is some person and
    Thesis is the thesis of Candidate and
    the committee of the Candidate includes Member and
    Request_To_Sign is some request and
```

```
    Member is the requestee of Request_To_Sign and
    The requester of Request_To_Sign is the chairman of the committee and
    Thesis is the document of subject of Request_To_Sign and
    Status of Request_To_Sign is pending
  THEN
    do answer(Member, Request_To_Sign) and
    do sign(Member, Thesis)
```

The literary themes at the heart of BRUTUS's stories are chosen for intrinsic interestingness: "The central thrust of BRUTUS's approach to story generation is . . . to begin with a well-established, interesting literary theme like betrayal, and then work down" (p. 199). So the interestingness is canned, and the complexity of the architecture makes for a startling air of creativity. But can interestingness be formalized? And is BRUTUS really creative?

Some—such as Hofstadter and the Fluid Analogies Research Group (1995)—have boldly tackled the challenge of endowing machines with creativity. Naysayers—such as Dreyfus (1992)—have claimed to show that the enterprise is impossible in principle. The former have tended to be workers in AI, the latter philosophers. Bringsjord and Ferrucci are unusual in that they belong to both camps at once. They "cheerfully operate under the belief that human (literary) creativity is beyond computation—and yet strive to craft the appearance of creativity from suitably configured computation" (p. 149).

I know of no reason to dispute the authors' claim (repeated on the website devoted to machine and book[1]) that BRUTUS is "the world's most advanced story generator." The approach is ingenious and thorough, and the results quite impressive. By contrast, the arguments against Strong AI, despite being couched in clean deductive form, remain unconvincing.

Bringsjord and Ferrucci may seem to have stacked the deck against BRUTUS (and in favor of their philosophical thesis) by insisting that all computer storytelling must proceed exclusively by means of the traditional tools of databases and algorithms. Their "approach is a thoroughly logic-based one. Neural nets and dynamical systems are nowhere to be found in this volume" (p. 26). Their "explanation" of this strategy "is based on two stories, one involving Tchaikovsky, and the other involving Sherlock Holmes." The first notes that Tchaikovsky convinced audiences that his sixth Symphony was worth listening to by renaming it *Pathétique*, and talking of the range of emotional experiences he intended it to express. No robot composer, the authors assert, could provide that sort of gloss. The second alleges that even if some connectionist robot descended from MIT's COG were to emulate the powers of Sherlock Holmes, it could never explain how it accomplished its feats of inference without using language and deductive logic.

Take the latter "explanation" first. Most of Sherlock Holmes's "deductions" are actually inductions, or crucially rest on prior—and often dubious—inductive inferences (such as "dogs always bark at strangers," in their example drawn from the story of Silver Blaze, p. 31). But suppose the detective's reasoning were exclusively logical. Bringsjord and Ferrucci stress the fact that the nature of this reasoning couldn't be "gleaned from neural nets" any more than our own could be read off a brain scan (p. 30). Quite, but so what? That doesn't show that the logical reasoning didn't super-

1 http://www.rpi.edu/dept/ppcs/BRUTUS/brutus.html

vene on the lower-level activity of both nets and neurons.

In this connection, Bringsjord and Ferrucci make a revealing observation: unless we can analyze "what cognizers think and say" in logical terms, then the robot's "success will be impenetrable, and will thus fail to advance our understanding of how detectives do what they do". And they conclude: "Since we desire not only to build literarily creative agents but to understand them, our use of logic goes without saying", though of course "it's fine with us, . . . if this computation takes the form of neural networks in action" (p. 31).

This remark about understanding is reminiscent of the old joke about looking for your keys under the lamp, because that is where the light is. If that just is the way the brain works, and logical inferences, like conscious states and creative innovations, are just supervenient on those unintelligible processes, then consciousness, creativity, and logic will all be emergent properties of a system the inner workings of which are simply too complex to be understood in detail. That may well be regrettable, but this hardly constitutes a good reason for thinking it's not true.

In fact, what neural nets notoriously seem to be good at is precisely inductive learning and pattern recognition. These are not all that Sherlock Holmes needs, but they are certainly key items in his toolbox. So why should neural nets be relegated to the implementation level of the deductive components of reasoning? Bringsjord and Ferrucci's answer rests on the fact that there can in principle be no difference between the logical powers of neural nets and those of symbolic systems: "Neural nets can be rendered as cellular automata, cellular automata can be rendered as . . . [Turing machines]"—and Turing machines are in principle equivalent to traditional symbolic computers (p. 38).

The problem is that this argument can as easily be stood on its head. Despite a long tradition of invoking Gödel to the contrary (Penrose 1994, pp. 171–209), there is no consensus that our own brains cannot be "rendered" as neural nets, and therefore as Turing machines. So the argument simply begs the question. It gives me no reason to reject the following alternative argument:

1. Humans are creative.
2. Human brains are neural nets.
3. Therefore neural nets can be creative.
4. Neural nets are logically equivalent to Turing machines.
5. Therefore Turing machines can be creative.

The same strong whiff of question-begging affects the authors' endorsement of well-known arguments against interpreting even the most spectacular success of their own program as a vindication of Strong AI. Let's look at just two of these arguments.

First, what I'll call the Direct Argument (DA). It appeals to the idea of novelty contained in creativity. DA rests on the plausible thought that creativity must involve the capacity to produce something new. But what is new? DA purports to show that no deterministic machine can ever count as creative:

1. Only something that merely implements algorithms counts as a genuine machine.
2. Nothing can count as creative if it produces nothing new.
3. Algorithms produce nothing new.

4. Nothing that merely implements algorithms can count as creative (2, 3).

5. Conclusion: No machine can be creative (1, 4).

DA seems impossibly short. Its premises beg important questions. Proposition (4) would entail that no standard mathematical proofs can be creative, since a mathematical proof typically implements algorithms. Premise (3) is sometimes advanced as the thesis that the conclusion of a valid deductive argument is "contained" in the premises, so that deductive arguments produce no new knowledge. Yet both mathematicians and those who "know no mathematics" would be puzzled by the claim that there is nothing the former know that the latter do not.

A crucial underlying assumption is that we ourselves do not implement algorithms when we think creatively. Bringsjord and Ferrucci admit that "raw origination . . . may well be impossible," but they claim to give examples of "a type of creativity in which something utterly and completely new is produced (e.g., non-Euclidean geometry)" (p. *xix*). But the example undermines the thesis. The elaboration of non-Euclidean geometries required little more than a change of perspective after Giovanni Girolamo Saccheri's purported proof of the parallels postulate by reductio ad absurdum (Saccheri and Halsted 1986). All that was needed was to notice that the "absurd" consequences entailed no contradiction. Bolyiai's and Lobachewski's creative recognition of that fact could certainly be attained by formal methods, but required the choice of a crucial change in attitude.

In the light of this example, the best prospect for creative machines seems to me to lie in selectionist systems. If variation and selection was good enough for God, it should be good enough for AI. Or are the products of biological evolution merely algorithmic and therefore not creative enough for Bringsjord and Ferrucci? I'm not sure: the issue seems to hang, for them, on whether there is an algorithm to effect the selection task. Their fifth chapter is devoted to showing that interestingness, like creativity itself, while *recognizable* by humans, is not recognized by means of any computable procedure. To be sure, a human organism brought up on multifarious experience can learn to recognize creativity *without ever becoming fully conscious of the procedures responsible* for that recognition. Since neural nets are good at learning to recognize patterns even when we don't really understand how they do it, the argument seems insufficient to establish that machines couldn't learn to recognize creativity and (in the context of a selectionist system) *thereby* learn to be creative.

Bringsjord and Ferrucci's Arg_4 (p. 74) purports to derive the conclusion that "Computers can't write sophisticated fiction." Crucial to this argument is the proposition:

10. "There's something it's like to be a" conscious states are not formalizable. . .

which is the conclusion of Arg_3 and rests, in turn, on this premise:

4. Alvin, prior to his [first ever] first-person long-lost-friend experience, doesn't know what it's like to meet a long-lost friend in the flesh. (p. 71)

In the much-discussed form in which it was put by Frank Jackson (1982), the place of Alvin is taken by Mary, a brain specialist who has been brought up in a purely black-and-white environment and knows, by description, "everything there is to know" about color. It is somehow supposed to be obvious that when suddenly confronted with the first-person experience of colors, she learns something she didn't know before

(p. 53). But why should telling that science-fiction story compel us to believe that? The story is a thought experiment, and thought experiments don't tell us what must be the case, but only what we think we would think if something were the case. There is no manifest logical incoherence involved in giving the story a different conclusion, in which the third-person knowledge would be sufficient to generate first-person knowledge. That there is such a logical incoherence is what Bringsjord and Ferrucci purport to prove; so they are not entitled to stuff this claim into a premise.

Towards the end of the book, Bringsjord and Ferrucci write:

> Computer programs are not human; they are not affected by the emotional elements that are part and parcel of human drama. In fact, it is hard to even *imagine* how a computer could discover, from a sea of swimming 0s and 1s, what might be compelling to a living, feeling human being. (p. 198)

Well, yes, it's hard to imagine, but then in the same sense it's hard to imagine how a brain could discover anything from a sea of swimming neurotransmitter drops, phosphorus ions, and the rest of it. We don't really know how it happens. But our ignorance and the failure of our imagination should not be confused with logical impossibility. Perhaps, indeed, the right kind of ignorance lies at the heart of what passes for creativity. The methods used by BRUTUS look mechanical when they are used on only a relatively small knowledge domain. But Bringsjord and Ferrucci have failed to show that *merely scaling up* could not make all the difference between merely simulated creativity and the real thing. Appeals to uncomputability in the explanation of human powers, like nineteenth-century appeals to entelechies and vital spirit, underestimate the power of well-placed ignorance to create key illusions. BRUTUS may not be creative yet, because we see too well, thanks to its authors' excellent exposition, just how it works. But once we raise the degree of its complexity and the diversity of its knowledge base, our skepticism may be harder to maintain. The key to creativity may just be nothing more than just enough of the right kind of ignorance in the face of complexity.

References

Dreyfus, Hubert L. 1992. *What Computers Still Can't Do: A critique of artificial reason*. The MIT Press, Cambridge, MA.

Hofstadter, Douglas R., and the Fluid Analogies Research Group. 1995. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books, New York.

Jackson, Frank. 1982. Epiphenomenal Qualia. *Philosophical Quarterly*, 32: 127–136.

Penrose, Roger. 1994. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press, Oxford and New York.

Saccheri, Girolamo and George B. Halsted. 1986. *Girolamo Saccheri's Euclides vindicatus*. Edited and translated by George B. Halsted. Chelsea Publishing Co., New York.

Ronald de Sousa is the author of *The Rationality of Emotion* (The MIT Press, 1987). His research interests include the modeling of mind and rational processes, and the philosophy of language, biology, cognitive science, and artificial life. De Sousa's address is: Department of Philosophy, University of Toronto, Toronto, Ontario, Canada M5S 1A1; e-mail: sousa@chass.utoronto.ca.