Chess Isn't Tough Enough: Better Games for Mind-Machine Competition

Selmer Bringsjord & Adam Lally
Dept. of Philosophy, Psychology & Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
selmer@rpi.edu • lallya@rpi.edu
http://www.rpi.edu/~brings

Chess Isn't Tough Enough

One of us (Bringsjord, 1997b) recently wrote:

> That Strong AI is still alive may have a lot to do with its avoidance of true tests. When Kasparov sits down to face the meanest chessbot in town, he has the deck stacked against him: his play may involve super-computation, but we know that perfect chess can be played by a finite-state automaton, so Kasparov loses if the engineers are sufficiently clever ... [(Bringsjord, 1997b), p. 9; paraphrased slightly to enhance out-of-context readability]

This quote carries the kernel of the present (embryonic) paper, a robust version of which will incorporate discussion at the workshop.

We find it incredible that anyone would have wagered that computers of the future would not manage to play at a level well beyond Kasparov. (We confess to indecisiveness concerning which prediction — Simon saying three decades ago that thinking machines would be upon us within days;[1] or Dreyfus betting the farm that formidable chessbots would forever be fictitious — was the silliest.) After all, we know that there exists a perfect winning strategy for chess, and that strategy, at bottom, isn't mathematically interesting. Complexity, of course, is another issue: it's complexity that generates interestingness in this domain — but the bottom line is that if complexity is somehow managed, a human player has his or her hands full. Deep Blue versus Kasparov was proof of that.

[1] Here's that unforgettable quote:

> It is not my aim to surprise or shock you — but the simplest way I can summarize is to say that there are now in the world machines that think, that learn and create. Moreover, their ability to do these things is going to increase rapidly until — in a visible future — the range of problems they can handle will be coextensive with the range to which human mind has been applied.

Checkmate to Debate to S^3G

Instead of the checkmate game, we would prefer the debate game. Sit Selmer down across from Deep Debate, throw out a topic (how 'bout "Is cognition computation?"), and let's go at it. When Selmer "senses a new kind of intelligence across the table" in such a fight, well, then there may be something to write home about. We could of course ask the audience what they sense, if anything. We expect that they will be saying "Nada" for decades to come.

Maybe the debate game is *too* tough. (To draw an opponent with a fighting chance, perhaps we can let a human proponent of Strong AI oppose Bringsjord.) After all, the debate game is essentially a form of the Turing Test, and though we are quite sure that a reasonably parameterized version of TT will be passed by an AI of the future,[2] the advent of such an AI won't come in the *near* future. So here's an easier game. Russell and Norvig, in their excellent *Artificial Intelligence: A Modern Approach* (Russell and Norvig, 1995), which one of us (Bringsjord) uses to teach AI, take an approach that is now familiar to nearly all: the "agent approach." The beauty of this approach is that it unites a field that otherwise looks disturbingly disparate — but the approach also provides the substrate for games that go beyond classic strategy games of the sort so popular at AAAI and IJCAI. In *AIMA*, an agent is a mapping from percepts to behavior (see Figure 1). So let's build an agent to play the "Short Short Story Game" (S^3G): The percept to the artificial player in S^3G is one relatively simple sentence, say: "Barnes kept the image to himself, kept the horror locked away as best he could." (For a much better one, see the "loaded" sentence shown in Figure 2. When

[2] One of us is also sure that the TT, and variants thereof, is inadequate; see (Bringsjord, 1995).

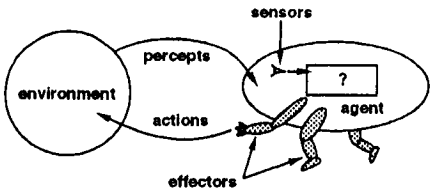

Figure 1: Russell and Norvig's Agent Scheme

will a machine give the Kafkas of this world a run for their money?) The same percept is given to the human player. Both must now fashion a short short story designed to be truly interesting; the more literary virtue, the better. The behavior in question, then, is simply producing the story (the length dimensions of which are specified, etc.) It seems to us that chessbots are arguably passé. So why not move to S^3G, or something similar, as the next frontier?

There are some rather deep reasons for moving from chess (and its cognates) to something like S^3G. Here are three:

1. Many cognitive scientists plausibly hold that narrative is at the very heart of human cognition. For example, in their lead target chapter in *Knowledge and Memory: The Real Story* (Wyer, 1995), Roger Schank and Robert Abelson boldly assert on the first page that "virtually all human knowledge" is based on stories.[3]

2. S^3G strikes right at the heart of the distinction between "Weak" and "Strong" AI. Humans find it impossible to produce literature without adopting the points of view of characters; hence human authors generate stories by capitalizing on the fact that they are conscious in the fullest sense of the word. Ibsen, for example, described in considerable detail how he couldn't write without feeling what it was like to be one of his characters. (We return to the notion of so-called "what it's like" consciousness below.) Chess, on the other hand, can *clearly* be played, and played *very* well, without the "Weak" vs. "Strong" split being touched. [For more on this second point, see (Bringsjord and Ferrucci, 1997).]

3. Despite the fact that our world is now populated with robots, softbots, immobots, and so on; despite the fact that AI continues to ascend — there remains a question, one that is on the minds of many of those who see our progress: namely: What about creativity? As many readers will know, Lady Lovelace famously pressed against Alan Turing and his "Turing Test" a short but powerful argument; charitably paraphrased, it runs as follows. "Computers can't create anything. For creation requires, minimally, *originating* something. But computers originate nothing; they merely do that which we order them, *via* programs, to do" (Turing, 1964). This argument seems to have bite against those who tout progress in checkers, chess, go and so on. It would seem likely to lose much of its force against a bot good enough to genuinely compete in S^3G.

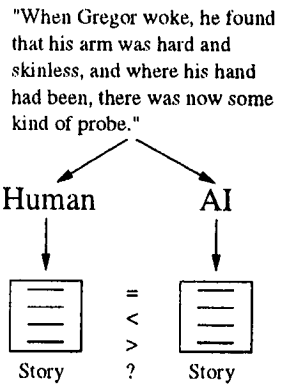

Figure 2: S^3G

How do machines fare in S^3G? How *will* they fare? Bringsjord may be in a good position to ponder such questions. With help from the Luce Foundation, Apple Computer, IBM, and the NSF, he has spent the past seven years working (along with a number of others, most prominently Dave Porush, Dave Ferrucci and Marie Meteer) to build a formidable artificial storyteller.[4] The most recent result of this toil is the agent BRUTUS$_1$, soon to debut in conjunction with the publishing of *Artificial Intelligence and Literary Creativity: The State of the Art* (Bringsjord and Ferrucci, 1997) from Lawrence Erlbaum. BRUTUS$_1$ is a rather interesting agent; he is capable of writing short short stories like the following.

"Betrayal in Self-Deception" (conscious)
by BRUTUS$_1$

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

[3] An insightful review of this book has been written by Tom Trabasso (Trabasso, 1996).

[4] The project is known as *Autopoeisis*, and now falls within a recently launched larger investigation of machine creativity undertaken by the *Creative Agents Group* at RPI.

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 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's book-lined office. Dave noticed that Hart's'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 Rogers and one from Dr. Meteer; Dave answered both, apparently to everyone's satisfaction. There were no further objections.

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

"Ed?" Rogers 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.

But such near-belletristic feats are possible for BRUTUS$_1$ only because he (we use 'he' rather than 'it' in order to remain sensitive to BRUTUS$_1$'s intimate relationship to the late, corporeal Brutus, who was of course male) has command over a formalization of the concept of betrayal.[5] (BRUTUS$_1$ also has a quasi-formal account of self-deception, and provisional accounts of evil[6] and voyeurism.) In order to adapt BRUTUS$_1$ to play well in S^{3}G, he would certainly need to "understand" not only betrayal, but other great literary themes as well — unrequited love, revenge, jealousy, patricide, and so on. Though our intention is to craft a descendant BRUTUS$_n$, for some $n > 1$, that "understands" all these literary concepts (and a lot more), perhaps S^{3}G is still a bit too tough. (At the workshop, Adam Lally can report on his attempt to build, from scratch, an agent capable of meaningfully playing S^{3}G.) Hence we briefly discuss a third type of game: infinite games.

"McNaught" and Infinite Games

Seeing as how there is insufficient space to set out all the mathematical niceties (they will have to wait for the full version of this paper), let's dive in and play an infinite game — a game we call, in deference to some recent investigations carried out by Robert McNaughton, "McNaught" (McNaughton, 1993). McNaught isn't a game like chess, mind you: chess, as we've noted, is after all a finite game, one handled quite well by ordinary computation, as even Dreyfus must now admit. We're talking about an *infinite* game; here's how it works. You will need a place-marker (a dime will do nicely), and the graph shown in Figure 3 (across which you will slide the dime). We will be black, you will be red. Notice that the nodes in the graph of Figure 3 are divided in half: three are red (r) nodes; three are black (b) nodes. If the dime is on an r node, then it's your turn, as red, to move; if the dime is on a b node, it's our turn. Here's how the game proceeds. The dime is placed randomly on one of the nodes, and then we take turns with you, sliding it from node to node, making sure that a move is made in accordance with a connecting arrow. So, if the dime is initially upon r_1, you would move, and your options are b_1 and b_2. If you slid

[5] The following definition gives a sense of the relevant formalization:

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

1 s_d at t_i wants p to occur;

2 s_r believes that s_d wants p to occur;

3′ $(3 \wedge 6') \vee$

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″ there is some action a such that:

4″a s_r performs a at t_b 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 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.

[6] In the case of evil, BRUTUS$_1$'s knowledge is based upon M. Scott Peck's description of this phenomenon as a species of psychiatric illness (Peck, 1983).

the dime to b_2, our only option would be r_3, and so on. Now, here's the thing: you and the two of us are going to take turns back and forth for an infinite amount of time. Since you may complain at this point that you are mortal, we want you to assume for the sake of the game that the three of us, like super-machines, can in fact take turns *forever*. [Super-machines are those with more power than Turing Machines. Super-*minds* are beings having, among other things, information-processing power above TMs. For more on super-computation in general, including an introduction to the Arithmetic Hierarchy, see (Bringsjord, 1997a). For a sustained defense of the view that human persons are indeed super-minds, see the forthcoming book *Super-Minds*: (Brinsjord and Zenzen, 1997).] Okay, now notice that nodes b_1 and r_1 are double-circles; this is because these two are "winning" nodes. We win, as black, if and only if either r_1 and b_1 are both visited only finitely many times or are both visited infinitely often. You, red, win if and only if *one* of these two nodes is visited infinitely often and the *other finitely* often. Got it? Okay, now: What is your strategy? What is your *best* strategy? What is *our* best strategy? If we both play our best, who will win? And supposing we play only for a finite amount of time, how could a referee predict a winner?

⋆*Don't read this paragraph if you intend to tackle these questions.*⋆ Only black has an invincible strategy, viz., from b_3 move to r_2 if b_1 has never been visited or if r_1 has been visited since the last time b_1 was visited; in all other circumstances move to r_1. So there was really no way for you to beat us! It is remarkable that ordinary computation can find this strategy when presented with the game in question (McNaughton, 1993). (No ordinary computer can literally play the game, of course.) However, for a game utterly beyond the Turing Limit, see the "undetermined" game featured in (Gale and Stewart, 1953): this is a game where a winning strategy cannot be devised by ordinary computation (in fact, there is no mathematical function which is a winning strategy!). It seems to us that infinite games, perhaps especially uncomputable infinite games, provide promising frameworks for mind-machine competition. The first step, which we are in the process of taking, is to take a computable infinite game and cast it in terms allowing for mind-mind competition. [We are starting with McNaught. The task of declaring a winner in *finite* time is rather challenging. See the approach indicated in (McNaughton, 1993).] Other frameworks might involve competition centering around the *creation* of infinite (and other types of) games.

On the "Big" Questions Driving the Workshop

We end by turning to questions in the 6 bullets from the original call for submissions (we have separated questions when more than one is given under a bullet):

- *Ontological*:

  O1 Are there thinking machines?

  O2 Is Deep Blue one of them?

- *Epistemological* **E**: What are the sufficient/necessary conditions for "sensing" intelligence?

- *Foundational*:

  F1 What does Kasparov versus Deep Blue mean to AI?

  F2 Is Deep Blue "AI"?

- *Historical* **H**: What are the important milestones in the development of chess-playing programs?

- *Technological*:

  T1 What software technology underlies the best chess playing programs?

  T2 What is the future of this technology?

- *Cultural* **C**: Why the negative emotional reaction to the notion of AI by some philosophers and cognitive scientists?

In order to answer these questions, let's distinguish between thinking$_a$ and thinking$_p$. Thinking$_a$ is "access thinking," which merely involves the processing of information in certain impressive ways. Thinking$_p$ is quite another thing: it is "phenomenal thinking," i.e., thinking that crucially involves subjective or phenomenal awareness: if one thinks$_p$ about that trip to Europe as a kid (e.g.), one remembers what it was like to be (say) in Paris on a sunny day with your older brother – whatever: any such example will do. The distinction between these two senses of thinking has its roots in a recent distinction made by Ned Block between A-consciousness and P-consciousness (Block, 1995). Adapting the first of these notions, we can hazard the following definition.

Thinking$_a$ An agent S thinks$_a$ iff it has internal states the representations of which are

1. inferentially promiscuous, i.e., poised to be used as a premise in reasoning;
2. poised for (rational) control of action; and
3. poised for rational control of speech.

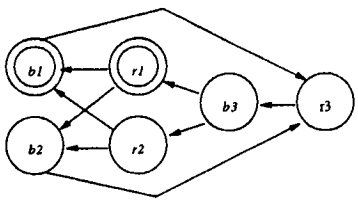

Figure 3: A Simple Game of "McNaught"

Here is how Block characterizes the notion of P-consciousness:

> 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. [(Block, 1995), p. 230]

Accordingly, we can say that an agent S thinks$_p$ iff it has P-conscious states.

Now we can synoptically present our answers to the big questions [many of which are discussed in (Bringsjord, 1992)]:

O1 There are certainly thinking$_a$ machines!

O2 Deep Blue is one of them. (So is BRUTUS$_1$. There are no thinking$_p$ machines, and if the machines in question are computers, thinking$_p$ machines won't ever arrive.)

E The Turing Test (and the debate game, S^{3}G, and possibly the infinite games we pointed to above) forms a sufficient condition for intelligence$_a$ (= thinking$_a$). I.e., if x passes TT (excels in S^{3}G), then x is intelligent$_a$ (= thinks$_a$). There are no empirical tests for thinking$_p$ (Bringsjord, 1995).

F1 It means that we are heading toward an age where the boundaries between human persons and intelligent$_a$ machines will blur. It's a milestone, a *big* one. It indicates that people had better buckle their seatbelts for an age in which, *behaviorally*, AIs can truly walk among us.

F2 Deep Blue is AI$_a$. Deep Blue is not AI$_p$. Deep Debate, if successful, might lay a better claim to AI$_p$ — but we still wouldn't have any way to know for sure.

H We defer to others.

T1 We defer to others.

T2 The future is incredibly bright. We currently have the technology to create ever more sophisticated thinking$_a$ machines. And it may be that such machines can do 80% of the work done presently by humans.

E Hey, this question is backwards. It should be: "Why the emotional attachment to Strong AI seen in many philosophers, cognitive scientists, and AIniks?"

References

Ned Block. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2):227–247, 1995.

Selmer Bringsjord and David Ferrucci. *Artificial Intelligence and Literary Creativity: The State of the Art in Story Generation.* Lawrence Erlbaum, Mahwah, NJ, 1997.

Selmer Bringsjord. *What Robots Can and Can't Be.* Kluwer, Dordrecht, The Netherlands, 1992.

Selmer Bringsjord. Could, how could we tell if, and why should–androids have inner lives. In Ken Ford, Clark Glymour, and Pat Hayes, editors, *Android Epistemology*, pages 93–122. MIT Press, Cambridge, MA, 1995.

Selmer Bringsjord. Philosophy and super computation. In Jim Moor and Terry Bynum, editors, *The Digital Phoenix: How Computers are Changing Philosophy*, pages n–m. Basil Blackwell, Cambridge, UK, 1997.

Selmer Bringsjord. Strong AI is simply silly. *AI Magazine*, 18(1):9–10, 1997.

Selmer Brinsjord and Michael Zenzen. *Super-Minds: A Defense of Uncomputable Cognition.* Kluwer, Dordrecht, The Netherlands, 1997.

D. Gale and F.M. Stewart. Infinite games with perfect information. In *Annals of Math Studies 28–Contributions to the Theory of Games*, pages 245–266. Princeton University Press, Princeton, NJ, 1953.

Robert McNaughton. Infinite games played on finite graphs. *Annals of Pure and Applied Logic*, 65:149–184, 1993.

M.S. Peck. *People of the Lie.* Simon and Shuster, New York, NY, 1983.

Stuart Russell and Peter Norvig. *Artficial Intelligence: A Modern Approach.* Prentice-Hall, Saddle River, NJ, 1995.

T. Trabasso. Review of "Knowledge and Memory: The Real Story". *Minds and Machines*, 6(1):399–403, 1996.

A.M. Turing. Computing machinery and intelligence. In A.R. Andersen, editor, *Minds and Machines*, pages 4–30. Prentice-Hall, Englewood Cliffs, NJ, 1964.

R.S. Wyer. *Knowledge and Memory: The Real Story.* Lawrence Erlbaum, Hillsdale, NJ, 1995.