

Scientifically Informative Discovery of (Gödel's) Model-Based Mathematical Discovery

Selmer Bringsjord

The Minds & Machines Laboratory

Dept. of Philosophy, Psychology & Cognitive Science

Department of Computer Science

Rensselaer Polytechnic Institute (RPI)

Troy NY 12180 USA

`selmer@rpi.edu` • `www.rpi.edu/~brings`

October 3, 1998

Framing My Goal: Deep Blue and Bad Questions

Many used to ask:

Q1 Could a computer ever beat the best human chess-player?

With Kasparov brooding and Deep Blue and his silicon cousins improving every week, many are *now* asking:

Q2 Before long, will a computer beat all human grandmasters *time and time again in normal tournament play*?

To Q2 all of us should unhesitatingly answer in the affirmative, given the dizzying ascension of raw computing power on this planet. But to put it baldly, these are really pretty bad questions, at least from the standpoint of scientific discovery.¹ The fact of the matter is that Deep Blue, though the wondrous product of three decades of research and development, tells us precious little about how Kasparov (or any other human) plays chess. Even those who know nothing of the niceties of search algorithms intuitively grasp the mathematical fact that perfect chess can be played in utterly mechanical fashion if one can look enough moves ahead; and the attempt to look sufficiently ahead on the strength of standard search algorithms in order to prevail against the likes of Kasparov is Deep Blue's *modus operandi*.² Given all this, despite Deep Blue's victory, a question remains:

Q3 Could a computer ever play chess *as Kasparov plays*?

If we could manage an answer to Q3, we would be guaranteed to have discovered along the way how the best human chess-players discover the powerful moves and strategies they do. (They seem to reason more on the basis of holistic patterns than mechanical search.) But though I know a thing or two about search algorithms, I know little about chess psychology; I'm not seeking an answer to Q3. My objective is an answer to an analogous question regarding Gödel's first and second incompleteness theorems (= Gödel I and II), as well as an answer to a generalized form of this question. (For ease of exposition, in the remainder of the abstract I refer only to Gödel I.)

¹I must confess up front that I'm as guilty as anyone in AI when it comes to putting heart and soul into an attempt to engineer systems that *simulate* (as opposed to *replicate*), by hook or by crook, sophisticated elements of human cognition — where that simulation fails to aid discovery of the nature of those elements. I refer to my sustained attempt over the past decade, undertaken with Dave Ferrucci and others, to engineer artificial agents able to generate creative stories. The best fruit of this attempt is the agent Brutus₁, detailed in [4]. This system is designed to produce stories that encourage humans to ascribe consciousness and creativity to the system, despite the fact that the system, like Deep Blue, is mindlessly syntactic. For a general treatment of the difference between a system that *appears* to be conscious, and one that is *in fact* conscious, see my [2].

²I have discussed these issues as they relate to the attempt to build creative agents (see the previous note) in [1].

My Goal: Discovering The Nature of (Gödel’s) Model-Based Discovery

Gödel I has given rise to a rather well-known philosophical question, viz.,

Q4 Does the theorem imply that people have an ability that can never be matched by machines?

Rather a lot of ink has been devoted to Q4 of late (some of it flowing from my own pen)³; less attention has been paid to a second, related question:

Q5 Could a machine ever prove Gödel I?

A “Yes” answer to Q5 has been produced by, among others, Art Quaife [7]. The mechanized proof of Gödel I that Quaife engineered was carried out by OTTER, a purely syntactic resolution-based theorem prover particularly well-suited to reasoning in first-order extensional logic. The trick that allows OTTER to prove a meta-mathematical theorem such as Gödel I is Quaife’s encoding of this theorem in the modal system KT4, but despite the encoding the proof remains utterly syntactic — “Deep Blueish,” if you will.

The OTTER-based proof of Gödel I is nothing at all like what Gödel himself did, at least as reported by Gödel himself and others (e.g., Hao Wang). But to avoid controversies that would inevitably arise if we undertook to read Gödel’s mind, my strategy is to focus on the model-based reasoning used to explain and prove Gödel I today (for the sake of convenience I will use my own versions of Gödel I, used in my classes and forthcoming in my book *Gödel and the Mind*). So my goal is to answer the following two questions (where the second is a generalization of the first) as a way to concretize an attempt to discover how a particular class of scientific discoveries (viz., mathematical ones) come about.

Q5’ Could a machine ever prove Gödel I in the model-based manner human logicians and mathematicians prove it?

Q6 Could a machine ever prove things in the model-based manner human logicians and mathematicians prove them?

Plan; Demonstrations

The following is the step-by-step plan I will take to reach my goal. This enumeration coincides with the flow of both the complete written paper and my presentation at MBR 98, after the context has been set.

1. Provide solutions to two elementary, stylized logic problems in two ways, the first way using OTTER, the second using model-based reasoning as it can be specified in Barwise and Etchemendy’s Hyperproof system. (The two problems are listed in the Appendix.) In my talk, demonstration using OTTER and Hyperproof will be supplied.

³Roger Penrose [5] [6], for example, has famously argued that Q4 should be answered in the affirmative. The bulk of my own writing on Q4 can be found in [3].

2. Explain Gödel I, using a certain Gödelian puzzle (see the Appendix) as a starting point. A rudimentary understanding of symbolic logic is more than sufficient to understand my explanation.
3. Explain and demonstrate Quaipe's mechanized proof of Gödel I in OTTER.
4. Provide a model-based proof-sketch of Gödel I, drawing upon analogies and diagrams. Consider if and how a machine could produce such model-based reasoning (see Q5' and Q6), anchoring this discussion to the current state of the art in diagrammatic reasoning in AI and cognitive science.
5. Based on a comparison of syntactic versus model-based reasoning as seen in points 3 and 4, offer a list of possible discoveries about how human mathematicians make model-based discoveries.

Appendix

0.1 The Mystery of the Missing Jam

The following puzzle is from [8].

“How about making us some nice tarts?” the King of Hearts asked the Queen of Hearts one cool summer day.

“What’s the sense of making tarts without jam?” said the Queen furiously. “The jam is the best part!”

“Then use jam,” said the King.

“I can’t!” shouted the Queen. “My jam has been stolen!”

“Really!” said the King. “This is quite serious! Who stole it?”

“How do you expect me to know who stole it? If I knew, I would have had it back long ago and the miscreant’s head in the bargain!”

Well, the King had his soldiers scout around for the missing jam, and it was found in the house of the March Hare, the Mad Hatter, and the Dormouse. All three were promptly arrested and tried.

“Now, now!” exclaimed the King at the trial. “I want to get to the bottom of this! I don’t like people coming into my kitchen and stealing my jam!”

“Did you by any chance steal the jam?” the King asked the March Hare.

“I never stole the jam!” pleaded the March Hare.

“What about you?” the King roared to the Hatter, who was trembling like a leaf.

“Are you by any chance the culprit?” The Hatter was unable to utter a word; he just stood there gasping and sipping his tea.

“If he has nothing to say, that only proves his guilt,” said the Queen, “so off with his head immediately!”

“No, no!” pleaded the Hatter. “One of us stole it, but it wasn’t me!”

“And what about you?” continued the King to the Dormouse.

“What do you have to say about all of this? Did the March Hare and the Hatter both tell the truth?”

“At least one of them did,” replied the Dormouse, who then fell asleep for the rest of the trial.

As subsequent investigation revealed, the March Hare and the Dormouse were not both speaking the truth. Who stole the jam?

0.2 The Dreadsbury Mansion Mystery

Someone Who lives in Dreadsbury Mansion killed Aunt Agatha. Agatha, the butler, and Charles live in Dreadsbury Mansion, and are the only people who live therein. A killer always hates his victim, and is never richer than his victim. Charles hates no one that Aunt Agatha hates. Agatha hates everyone except the butler. The butler hates everyone not richer than Aunt Agatha. The butler hates everyone Agatha hates. No one hates everyone. Agatha is not the butler.

Now, given the above clues, there is a bit of a disagreement between three (incompetent?) Norwegian detectives: Inspector Bjorn is sure that Charles didn't do it. Is he right? Inspector Reidar is sure that it was a suicide. Is he right? Inspector Olaf is sure that the butler, despite conventional wisdom, is innocent. Is he right?

0.3 A Gödelian Puzzle

Suppose there is a machine \mathcal{M}_0 which prints out various expressions built from the following five symbols:

$$\sim P M ()$$

By an **expression** we mean any finite non-empty string built from these five symbols. (So $PPPPPPMM(($ is an expression, as is $\sim P(P)$.) An expression is called **printable** if the machine can print it. We assume that \mathcal{M}_0 is programmed so that any expression it can print will be printed sooner or later.

The **mirror** of an expression ϕ is the expression $\phi(\phi)$ — e.g., the mirror of $P \sim$ is $P \sim (P \sim)$. A **sentence** is an expression having one of the following four forms:

1. $P(\phi)$
2. $PM(\phi)$
3. $\sim P(\phi)$
4. $\sim PM(\phi)$

P stands for “printable;” M stands for “the mirror of” and \sim stands (as it often does in logic) for “not.” Hence we define $P(\phi)$ to be **true** iff⁴ ϕ is printable. We define $PM(\phi)$ to be true if the **mirror** of ϕ is printable. We call $\sim P(\phi)$ true iff ϕ is not printable, and $\sim PM(\phi)$ is defined to be true iff the mirror of ϕ is not printable.

We are given that the machine \mathcal{M}_0 is accurate in that all sentences printed by the machine are true. So, for example, if the machine ever prints $P(\phi)$, then ϕ really is printable (i.e., ϕ will be printed by \mathcal{M}_0 sooner or later). Also, if $PM(\phi)$ is printable, so is $M(\phi)$.

Suppose ϕ is printable. Do we then know that $P(\phi)$ is printable? No; here's why. If ϕ is printable then $P(\phi)$ is certainly *true*, but we are not given that the machine is capable of printing *all* true sentences — only that the machine never prints any false ones.

Question: Is it *possible* that the machine *can* print all true sentences? Why?⁵

⁴Traditional abbreviation in logic for ‘if and only if.’

⁵This puzzle is adapted slightly from a puzzle given by Smullyan in his *Gödel's Incompleteness Theorems*.

References

- [1] Bringsjord, S. (1998) “Chess is Too Easy,” *Technology Review*.
- [2] Bringsjord, S. (forthcoming) “The Zombie Attack on the Computational Conception of Mind,” *Philosophy and Phenomenological Research*.
- [3] Bringsjord, S. (1992) “Chapter VIII: Gödel,” in his *What Robots Can and Can't Be* (Dordrecht, The Netherlands: Kluwer).
- [4] Bringsjord, S. & Ferrucci, D. (1999) *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus₁, A Storytelling Machine* (Hillsdale, NJ: Lawrence Erlbaum).
- [5] Penrose, R. (1969) *The Emperor's New Mind* (Oxford, UK: Oxford University Press).
- [6] Penrose, R. (1994) *Shadows of the Mind* (Oxford, UK: Oxford University Press).
- [7] Quaipe, A. (1992) *Automated Development of Fundamental Mathematical Theories* (Dordrecht, The Netherlands).
- [8] Smullyan, R. M. (1982) *Alice in Puzzleland* (NY, NY: Morrow).