CRI
CHRISTIAN RESEARCH INSTITUTE
PO Box 8500, Charlotte, NC 28271

Feature Article: **JAF1396**

WHO'S AFRAID OF HAL? WHY COMPUTERS WILL NOT BECOME CONSCIOUS AND TAKE OVER THE WORLD

by Charles Edward White

This article first appeared in the CHRISTIAN RESEARCH JOURNAL, volume **39**, number **06** (**2016**). For further information or to subscribe to the CHRISTIAN RESEARCH JOURNAL, go to: http://www.equip.org/christian-research-journal/.

Back in 1968, Stanley Kubrick's film *2001: A Space Odyssey* played on the fear that computers soon would become conscious, independent, and dangerous to humanity. In the story, the computer, called HAL, controls a spaceship with a human crew. When two crewmen decide to override HAL and retake control of the spacecraft, HAL murders one of them and attempts to kill the others. Three issues ago in this journal, James Hoskins reported that such a fear of computers is not just the stuff of sci-fi nightmares but also is shared by Stephen Hawking, Elon Musk, and Bill Gates.[1]

They need not worry.

While Kubrick, Hawking, Musk, Gates, and the rest of us should be afraid of what some people armed with supercomputers and artificial intelligence can do to us, we have no need to fear what the *computers themselves* can do. A groundbreaking mathematical discovery in 1930, and its implications for computer science, can put our minds at rest.

The story of this astonishing discovery starts in the eighteenth century, when David Hume challenged Galileo's idea that mathematics is the language in which God writes the laws of nature. Reacting to Hume, Kant proposed that even if we cannot be sure that mathematics works in the external world, we can know that it does work inside our minds by the laws of reason. This idea set mathematicians the task of shoring up the foundations, of proving that all mathematics is securely founded on reason. Beginning by establishing that one plus one equals two, Gottlob Frege thought that he had demonstrated that arithmetic and algebra were reasonable and logically consistent.

He was just getting ready to publish his *magnum opus, The Foundations of Arithmetic,* when he got a letter from Bertrand Russell. Russell showed him that the set theory on which Frege had based his whole work was logically inconsistent. "Arithmetic is tottering," shuddered Frege.[2]

Wondering what other branches of mathematics were tottering, many nineteenth-century mathematicians began searching for previously unrecognized holes in their arguments. Their quest came to an end in 1930 when Kurt Gödel announced his earth-shattering "incompleteness theorem." This idea is one of the intellectual milestones of the twentieth century: thinkers rank it with the discoveries made by Newton, Einstein, and Heisenberg. Gödel proved that no system of mathematics will be able to prove itself completely. Every system will be "incomplete" because there always will be true mathematical statements within the system that the system itself cannot prove. How Gödel proved this startling result is beyond the scope of this article, but that it is true is universally recognized.[3]

A trivial example of the incompleteness theorem involves a mathematical system, called S1, consisting of odd and even numbers and the addition operation. Within that system, it is impossible to prove that no three odd numbers add to twenty. By going outside S1, however, it is easy to show that three odd numbers can never make twenty:

1. Let X, Y, and Z be any whole numbers.
2. Then (2X + 1), (2Y + 1), and (2Z + 1) are odd numbers.
3. Assume (2X + 1) + (2Y + 1) + (2Z + 1) = 20
4. Then 2(X + Y + Z) + 3 = 20
5. Then 2(X + Y + Z) = 17
6. Then (X + Y + Z) = 8.5
7. But by number theory, the set of whole numbers is closed under addition (one can never get fractions by adding whole numbers), so statement 6 is false.
8. Therefore statement 3 must be false, and no three odd numbers add to twenty.

A skeptic might suggest we modify the system by adding number theory, fractions, and multiplication, making system S2. So now the system S2 can show that no three odd numbers can combine to make twenty. But there still is at least one true statement that is unprovable by that system. S2 cannot tell us the sum of the series 1-

1/3+1/5-1/7+1/9-…. (The elegant answer is π/4.) Now if we upgrade S2 to S3 by adding transcendental numbers, such as π, there will still be at least one true statement that is unprovable in S3. What Gödel proved "indisputably" is that this process can go on forever.[4]

Another way of understanding Gödel's discovery involves the barber's paradox. In the town of Seville, there is only one barber, Figaro, and this man shaves every man who does not shave himself. Does Figaro have a beard? Simply from hearing the story, it is impossible to tell. If he has a beard, then he does not shave himself. But if he does not shave himself, then the barber shaves him; but since he is the barber, he does shave himself, and so he does not have a beard. The only way to find out is to go to Seville and look at Figaro, who either has a beard or does not have a beard. This "Seville" system has two statements, but from it alone, it is impossible to determine the truth of the conclusion "Figaro has a beard." Thus the "Seville system" is incomplete. What Gödel showed is that every mathematical system is a "Seville system." One has to go outside the system (by going to look at Figaro) to prove all the truths it contains.

About ten years after Gödel's discovery, Alan Turing, the father of artificial intelligence, applied this thinking to computers. He realized that working a mathematical system is what a computer does. Thus, if all mathematical systems are incomplete, then the machines that implement those systems must be incomplete as well. He showed that even a "Turing Machine," an imaginary computer of infinite speed and capacity, that ran forever could never prove that no three odd numbers add to twenty, if it were programed to use only odd and even numbers and addition (S1). Since there is an infinite quantity of odd and even numbers, both positive and negative, the computer would never run out of integer triplets to evaluate. It would find that none of them equaled twenty, but it could never exhaust the infinite number of combinations.[5]

A real life example of Turing's findings comes from the world of computer chess. (See Figure 01.)

99

(Figure 01)

It's White's move. Should the pawn capture the rook? White should not take the rook. Black has a strong material advantage, but unless White breeches the pawn wall, there is no way for Black to move any of the pieces into a position to threaten White's king. White should move the king around behind this impregnable defense until a fifty-move draw occurs. This solution is obvious to any but a novice human player, but it escaped Deep Thought, the best chess-playing computer of its day. Despite being able to beat several grandmasters, the computer took the rook, and suffered the inevitable loss. The reason a human gets the draw and the computer loses is that the computer never "sees" the pawn barrier.[6]

Turing's adaptation of Gödel's finding prompted Oxford philosopher John Lucas to realize that if computers could never solve some problems whose solutions are obvious to humans, then there must be an essential difference between human minds and machines. He wrote, "Gödel's theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines."[7] The difference between a human mind and a computer is not just quantitative but qualitative. It is not simply one of degree but of kind. Minds and computers must therefore be different in their ontology, not just in their power. What Lucas means by "Mechanism" is also called "physicalism": energy, matter, time, and space are all there is. Physicalism comes in two flavors: reductive and nonreductive. The difference is that reductive physicalism

believes that the phenomenon of human consciousness is an illusion, while nonreductive physicalism thinks that consciousness is a real feature of the brain's physical activity.[8] Lucas understands that Gödel's incompleteness theorem precludes both types of physicalism.

Another Oxford thinker, mathematical physicist Roger Penrose, builds on the thought of Gödel, Turing, and Lucas to argue not only that minds and machines are fundamentally different but also that certain aspects of human thought can never be rivaled by computers. Computers perform "computations," and minds engage in "conscious thinking." One way to tell if there is a difference is to run a "Turing test." A person communicates using a keyboard with either a computer or a person who is hidden from view. If the real person can figure out what the other conversation partner is, then there is a difference between computation and conscious thought. If the real person cannot tell whether the conversation partner is another real person or a computer, then conscious thought must be reducible to computation, and some form of physicalism must be right. If the computer can fool the person, reductive physicalists say the computer has achieved consciousness, and nonreductive physicalists say it has simulated it.[9] Either way, the computer wins.

Some people believe that the Turing test has settled the issue. Several times computers have been able to convince human partners that they, too, were human. Since people cannot tell whether their conversation partner is human or mechanical, computers must be able to "think" as well as people. One victory for the computer happened when it was programed to sprinkle a few typing mistakes into its answers. The people, assuming that only a real person would make mistakes, were fooled. However, this kind of subterfuge simply underscores how different people and computers are. The only reason the computer was able to fool the humans was that other human programmers, knowing how people think, were able to build misleading "mistakes" into the computer's output. This clever strategy was imagined by people, not thought up by the computers themselves. Instead of proving that computers think, it shows that people are the only possible source of clever ideas.[10]

The creation of new ideas or "outside the box" thinking is exactly the kind of thought envisioned by Gödel's work. Computers are designed for systematic thinking, and Gödel showed that such systematic thinking never can produce complete results. The incompleteness theorem shows that the answer to the question, "Can there ever be a general method for solving all mathematical problems?" is "No." Because no mathematical system can prove all of its truths, and all computers depend on mathematical systems, the answer to the question, "Can a computer ever think exactly like a person?" is also "No." Thus Constance Reid concludes her book on higher mathematics by saying, "Now it is established — with all the certainty of logical proof

— that machines will never, even in theory, replace mathematicians."[11] The reason computers will never replace mathematicians — or even ordinary people — is that people think in ways that computers never can duplicate.

Besides setting us free from the fear that computers like HAL will somehow come to consciousness and take over the world, the truths discovered by Gödel, Turing, Lucas, and Penrose also have apologetic implications. If human minds cannot be reduced to computers made of silicon and steel, then they also cannot be reduced to computers made of protoplasm and protein. Our minds are more than our physical brains. Since there is more to the human mind than the material of the physical brain, something immaterial must exist in the universe. The existence of the immaterial, the metaphysical, opens the door to spiritual reality. Once it is clear that something other than the physical world exists, can God be far behind?

Charles Edward White, PhD (church history, Boston University), is professor of Christian Thought and History at Spring Arbor University.

NOTES

1 James Hoskins, "Digital Souls," *Christian Research Journal* 39, 2 (2016): 34–39.
2 Morris Kline, *Mathematics: The Loss of Certainty* (New York: Oxford University Press, 1980), 46, 74–76; Stephen M. Barr, *Modern Physics and Ancient Faith* (Notre Dame, IN: University of Notre Dame Press, 2003), 279–80; and Roger Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness* (New York: Oxford University Press, 1994), 65.
3 Barr provides an accessible explanation; see his *Modern Physics and Ancient Faith*, 279–88.
4 Penrose, *Shadows of the Mind*, 65.
5 Barr, *Modern Physics and Ancient Faith*, 211; and Constance Reid, *Introduction to Higher Mathematics for the General Reader* (New York: Thomas Y. Crowell, 1962), 174–78.
6 J. Seymour and David Norwood, "A Game for Life," *New Scientist* 139 (September, no. 1889), 23–26, cited in Penrose, 46–47. Of course, the computer could be reprogrammed to cope with pawn barriers, but because of Turing's insight, there will be at least one ploy a person could imagine that the computer could not anticipate. The computer could then be reprogrammed to accommodate that strategy, and the person could find a new one, and so on until the computer had mastered all possible games of chess, a number estimated to be 10120. See "Shannon number" in Wikipedia.
7 John R. Lucas, "Minds, Machines, and Gödel," in *The Modeling of Mind*, ed. K. M. Sayer and J. M. Crosson (Notre Dame, IN: University of Notre Dame Press, 1963), 255.
8 Penrose, Shadows of the Mind, 12–13.
9 Ibid., 12–15.
10 See "Turing Test" in Wikipedia.
11 Reid, *Introduction to Higher Mathematics for the General Reader*, 180.