Betting your life on an algorithm

Daniel C. Dennett

Department of Philosophy, Tufts University, Medford, MA 02155

What minds can do, Penrose claims, is to see or judge that certain mathematical propositions are true by "insight" rather than mechanical proof. Penrose then argues that there could be no algorithm, or at any rate no practical algorithm, for insight. This ignores an independently plausible possibility: The algorithms that minds use for judging mathematical truth are not algorithms "for" insight – but they nevertheless work very well.

Consider a parallel argument. Chess is a finite game, so there is an algorithm "for" either checkmate or draw: the brute force algorithm that draws the entire decision tree for chess and works backwards from the last nodes. That algorithm surely is not practical. Probably there is no practical algorithm "for" checkmate. There are plenty of practical algorithms that *achieve checkmate with great reliability,* however. They are the chess-playing programs and although none is mathematically guaranteed to achieve checkmate against any opponent, you could safely bet your life that the best of them will always achieve checkmate against me (for instance). There are algorithms for playing legal chess – that is guaranteed mathematically. Checkmate is an unprovable bonus, but it is not a gift out of the blue. It is to be explained in terms of the relative cunning of these chance-taking algorithms. Aside from sheer speed, no other properties of a chess-playing computer – its material composition or genealogy, for instance – would be relevant to its power to achieve checkmate.

The following argument is therefore fallacious:

1. X is superbly capable of achieving Y (e.g., checkmate).
2. There is no practical algorithm for achieving Y. therefore
3. X's power to achieve Y is not explicable in terms of any algorithm.

Therefore, even if mathematicians are superb recognizers of mathematical truth, and even if there is no algorithm, practical or otherwise, "for" recognizing mathematical truth, it does not follow that the power of mathematicians to recognize mathematical truth is not entirely explicable in terms of their brains executing one or another garden-variety algorithm. Not an algorithm "for" intuiting mathematical truth – for the sake of the argument, I will grant to Penrose that there can be no such algorithm – but an algorithm for something else. What? Most plausibly it would be an algorithm – one of many – for *trying to stay alive,* an algorithm that, by an extraordinarily convoluted and indirect generation of byproducts, "happened" to be a superb (but not foolproof) recognizer of friends, enemies, food, shelter, harbingers of spring, good arguments – and mathematical truths.

Chess programs, like all heuristic algorithms, are designed to take chances, and therein lies their vulnerability in principle. What are the limits of vulnerable-in-principle probabilistic algorithms running on a parallel architecture such as the human brain? Penrose neglects to provide any argument to show what those limits are; hence he fails to cut off the most plausible rival interpretation of the mathematicians' prowess, on which his whole case depends. Notice that it is *not* a question of what the in-principle limits of algorithms are; those are simply irrelevant in a biological setting. To put it provocatively, an algorithm may "happen" to achieve this 999 times out of 1,000, in jig time. This prowess would fall outside its official limits (since you cannot prove, mathematically, that it will not run forever without an answer or else give a false answer), but it might nevertheless be prowess you could bet your life on. Mother Nature's creatures do it every day.

Sometimes Penrose suggests that what human mathematicians do is something that could not even be approximated by a heuristic, mistake-prone algorithm, since mathematicians (in principle? always?) settle into a consistent shared view. If they make a mistake, they can (will?) always correct it. Is this supposed to be an independently confirmable empirical premise? This could not be proven mathematically, of course, for such consistency proofs of oneself (or oneselves acting in concert) are ruled out by the very mathematical results Penrose relies on. He can perhaps fervently believe, and assert, that the joint or Ideal Mathematician is consistent and capable (in principle) of intuiting every mathematical truth (and no falsehoods), but he cannot hope to persuade those of us who find this an unlikely and unmotivated dogma by offering a mathematical proof, and there seems every empirical reason for simply disbelieving it. Penrose's envisaged revolution in physics

may happen, but not – so far as I can see – because it is needed to explain any fact or phenomenon of human mental powers.

ACKNOWLEDGMENT

This commentary is a revision of material contained in my review of Penrose's book (Dennett 1989, pp. 1055ff).