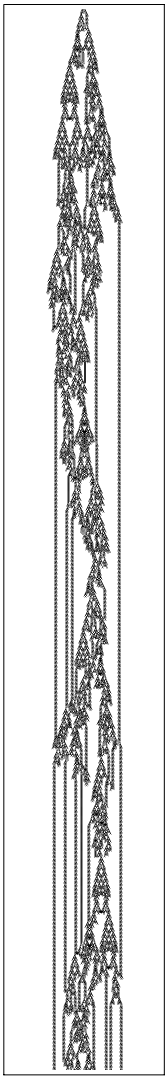
EXCERPTED FROM

STEPHEN WOLFRAM

A NEW KIND OF SCIENCE

SECTION 12.7

The Phenomenon of Free Will



A cellular automaton whose behavior seems to show an analog of free will. Even though its underlying laws are definite—and simple—the behavior is complicated enough that many aspects of it seem to follow no definite laws. (The rule used is the same as on page 740.)

this book even systems with very simple underlying rules can still perform computations that are as sophisticated as in any system.

And what this means is that to capture the essential features even of systems with very complex behavior it can be sufficient to use models that have an extremely simple basic structure. Given these models the only way to find out what they do will usually be just to run them. But the point is that if the structure of the models is simple enough, and fits in well enough with what can be implemented efficiently on a practical computer, then it will often still be perfectly possible to find out many consequences of the model.

And that, in a sense, is what much of this book has been about.

The Phenomenon of Free Will

Ever since antiquity it has been a great mystery how the universe can follow definite laws while we as humans still often manage to make decisions about how to act in ways that seem quite free of obvious laws.

But from the discoveries in this book it finally now seems possible to give an explanation for this. And the key, I believe, is the phenomenon of computational irreducibility.

For what this phenomenon implies is that even though a system may follow definite underlying laws its overall behavior can still have aspects that fundamentally cannot be described by reasonable laws.

For if the evolution of a system corresponds to an irreducible computation then this means that the only way to work out how the system will behave is essentially to perform this computation—with the result that there can fundamentally be no laws that allow one to work out the behavior more directly.

And it is this, I believe, that is the ultimate origin of the apparent freedom of human will. For even though all the components of our brains presumably follow definite laws, I strongly suspect that their overall behavior corresponds to an irreducible computation whose outcome can never in effect be found by reasonable laws.

And indeed one can already see very much the same kind of thing going on in a simple system like the cellular automaton on the left. For

even though the underlying laws for this system are perfectly definite, its overall behavior ends up being sufficiently complicated that many aspects of it seem to follow no obvious laws at all.

And indeed if one were to talk about how the cellular automaton seems to behave one might well say that it just decides to do this or that—thereby effectively attributing to it some sort of free will.

But can this possibly be reasonable? For if one looks at the individual cells in the cellular automaton one can plainly see that they just follow definite rules, with absolutely no freedom at all.

But at some level the same is probably true of the individual nerve cells in our brains. Yet somehow as a whole our brains still manage to behave with a certain apparent freedom.

Traditional science has made it very difficult to understand how this can possibly happen. For normally it has assumed that if one can only find the underlying rules for the components of a system then in a sense these tell one everything important about the system.

But what we have seen over and over again in this book is that this is not even close to correct, and that in fact there can be vastly more to the behavior of a system than one could ever foresee just by looking at its underlying rules. And fundamentally this is a consequence of the phenomenon of computational irreducibility.

For if a system is computationally irreducible this means that there is in effect a tangible separation between the underlying rules for the system and its overall behavior associated with the irreducible amount of computational work needed to go from one to the other.

And it is in this separation, I believe, that the basic origin of the apparent freedom we see in all sorts of systems lies—whether those systems are abstract cellular automata or actual living brains.

But so in the end what makes us think that there is freedom in what a system does? In practice the main criterion seems to be that we cannot readily make predictions about the behavior of the system.

For certainly if we could, then this would show us that the behavior must be determined in a definite way, and so cannot be free. But at least with our normal methods of perception and analysis one

typically needs rather simple behavior for us actually to be able to identify overall rules that let us make reasonable predictions about it.

Yet in fact even in living organisms such behavior is quite common. And for example particularly in lower animals there are all sorts of cases where very simple and predictable responses to stimuli are seen. But the point is that these are normally just considered to be unavoidable reflexes that leave no room for decisions or freedom.

Yet as soon as the behavior we see becomes more complex we quickly tend to imagine that it must be associated with some kind of underlying freedom. For at least with traditional intuition it has always seemed quite implausible that any real unpredictability could arise in a system that just follows definite underlying rules.

And so to explain the behavior that we as humans exhibit it has often been assumed that there must be something fundamentally more going on—and perhaps something unique to humans.

In the past the most common belief has been that there must be some form of external influence from fate—associated perhaps with the intervention of a supernatural being or perhaps with configurations of celestial bodies. And in more recent times sensitivity to initial conditions and quantum randomness have been proposed as more appropriate scientific explanations.

But much as in our discussion of randomness in Chapter 6 nothing like this is actually needed. For as we have seen many times in this book even systems with quite simple and definite underlying rules can produce behavior so complex that it seems free of obvious rules.

And the crucial point is that this happens just through the intrinsic evolution of the system—without the need for any additional input from outside or from any sort of explicit source of randomness.

And I believe that it is this kind of intrinsic process—that we now know occurs in a vast range of systems—that is primarily responsible for the apparent freedom in the operation of our brains.

But this is not to say that everything that goes on in our brains has an intrinsic origin. Indeed, as a practical matter what usually seems to happen is that we receive external input that leads to some train of thought which continues for a while, but then dies out until we get

more input. And often the actual form of this train of thought is influenced by memory we have developed from inputs in the past—making it not necessarily repeatable even with exactly the same input.

But it seems likely that the individual steps in each train of thought follow quite definite underlying rules. And the crucial point is then that I suspect that the computation performed by applying these rules is often sophisticated enough to be computationally irreducible—with the result that it must intrinsically produce behavior that seems to us free of obvious laws.

Undecidability and Intractability

Computational irreducibility is a very general phenomenon with many consequences. And among these consequences are various phenomena that have been widely studied in the abstract theory of computation.

In the past it has normally been assumed that these phenomena occur only in quite special systems, and not, for example, in typical systems with simple rules or of the kind that might be seen in nature. But what my discoveries about computational irreducibility now suggest is that such phenomena should in fact be very widespread, and should for example occur in many systems in nature and elsewhere.

In this chapter so far I have mostly been concerned with ongoing processes of computation, analogous to ongoing behavior of systems in nature and elsewhere. But as a theoretical matter one can ask what the final outcome of a computation will be, after perhaps an infinite number of steps. And if one does this then one encounters the phenomenon of undecidability that was identified in the 1930s.

The pictures on the next page show an example. In each case knowing the final outcome is equivalent to deciding what will eventually happen to the pattern generated by the cellular automaton evolution. Will it die out? Will it stabilize and become repetitive? Or will it somehow continue to grow forever?

One can try to find out by running the system for a certain number of steps and seeing what happens. And indeed in example (a) this approach works well: in only 36 steps one finds that the pattern