EXCERPTED FROM

STEPHEN WOLFRAM
A NEW KIND OF SCIENCE

SECTION 10.12

Human Thinking

So even if one allows rather general structure, the evidence is that in the end there is no way to set up any simple formula that will describe the outcome of evolution for a system like rule 30.

And even if one settles for complicated formulas, just finding the least complicated one in a particular case rapidly becomes extremely difficult. Indeed, for formulas of the type shown on page 618 the difficulty can already perhaps double at each step. And for the more general formulas shown on the previous page it may increase by a factor that is itself almost exponential at each step.

So what this means is that just like for every other method of analysis that we have considered, we have little choice but to conclude that traditional mathematics and mathematical formulas cannot in the end realistically be expected to tell us very much about patterns generated by systems like rule 30.

Human Thinking

When we are presented with new data one thing we can always do is just apply our general powers of human thinking to it. And certainly this allows us with rather modest effort to do quite well in handling all sorts of data that we choose to interact with in everyday life. But what about data generated by the kinds of systems that I have discussed in this book? How does general human thinking do with this?

There are definitely some limitations, since after all, if general human thinking could easily find simple descriptions of, for example, all the various pictures in this book, then we would never have considered any of them complex.

One might in the past have assumed that if a simple description existed of some piece of data, then with appropriate thinking and intelligence it would usually not be too difficult to find it. But what the results in this book establish is that in fact this is far from true. For in the course of this book we have seen a great many systems whose underlying rules are extremely simple, yet whose overall behavior is sufficiently complex that even by thinking quite hard we cannot recognize its simple origins.

Usually a small amount of thinking allows us to identify at least some regularities. But typically these regularities are ones that can also be found quite easily by many of the standard methods of perception and analysis discussed earlier in this chapter.

So what then does human thinking in the end have to contribute? The most obvious way in which it stands out from other methods of perception and analysis is in its large-scale use of memory.

For all the other methods that we have discussed effectively operate by taking each new piece of data and separately applying some fixed procedure to it. But in human thinking we routinely make use of the huge amount of memory that we have built up from being exposed to billions of previous pieces of data.

And sometimes the results can be quite impressive. For it is quite common to find that even though no other method has much to say about a particular piece of data, we can immediately come up with a description for it by remembering some similar piece of data that we have encountered before.

And thus, for example, having myself seen thousands of pictures produced by cellular automata, I can recognize immediately from memory almost any pattern generated by any of the elementary rules—even though none of the other methods of perception and analysis can get very far whenever such patterns are at all complex.

But insofar as there is sophistication in what can be done with human memory, does this sophistication come merely from the experiences that are stored in memory, or somehow from the actual mechanism of memory itself?

The idea of storing large amounts of data and retrieving it according to various criteria is certainly quite familiar from databases in practical computing. But there is at least one important difference between the way typical databases operate, and the way human memory operates. For in a standard database one tends to be able to find only data that meets some precise specification, such as containing an exact match to a particular string of text. Yet with human memory we routinely seem to be able to retrieve data on the basis of much more general notions of similarity.

In general, if one wants to find a piece of data that has a certain property—either exact or approximate—then one way to do this is just to scan all the pieces of data that one has stored, and test each of them in turn. But even if one does all sorts of parallel processing this approach presumably in the end becomes quite impractical.

So what can one then do? In the case of exact matches there are a couple of approaches that are widely used in practice.

Probably the most familiar is what is done in typical dictionaries: all the entries are arranged in alphabetical order, so that when one looks something up one does not need to scan every single entry but instead one can quickly home in on just the entry one wants.

Practical database systems almost universally use a slightly more efficient scheme known as hashing. The basic idea is to have some definite procedure that takes any word or other piece of data and derives from it a so-called hash code which is used to determine where the data will be stored. And the point is that if one is looking for a particular piece of data, one can then apply this same procedure to that data, get the hash code for the data, and immediately determine where the data would have been stored.

But to make this work, does one need a complex hashing procedure that is carefully tuned to the particular kind of data one is dealing with? It turns out that one does not. And in fact, all that is really necessary is that the hashing procedure generate enough randomness that even though there may be regularities in the original data, the hash codes that are produced still end up being distributed roughly uniformly across all possibilities.

And as one might expect from the results in this book, it is easy to achieve this even with extremely simple programs—either based on numbers, as in most practical database systems, or based on systems like cellular automata.

So what this means is that regardless of what kind of data one is storing, it takes only a very simple program to set up a hashing scheme that lets one retrieve pieces of data very efficiently. And I suspect that at least some aspects of this kind of mechanism are involved in the operation of human memory.

But what about the fact that we routinely retrieve from our memory not just data that matches exactly, but also data that is merely similar? Ordinary hashing would not let us do this. For a hashing procedure will normally put different pieces of data at quite different locations—even if the pieces of data happen in some sense to be similar.

So is it possible to set up forms of hashing that will in fact keep similar pieces of data together? In a sense what one needs is a hashing procedure in which the hash codes that are generated depend only on features of the data that really make a difference, and not on others.

One practical example where this is done is a simple procedure often used for looking up names by sound rather than spelling. In its typical form this procedure works by dropping all vowels and grouping together letters like "d" and "t" that sound similar, with the result that at least in some approximation the only features that are kept are ones that make a difference in the way a word sounds.

So how can one achieve this in general?

In many respects one of the primary goals of all forms of perception and analysis is precisely to pick out those features of data that are considered relevant, and to discard all others.

And so, as we discussed earlier in this chapter, the human visual system, for example, appears to be based on having nerve cells that respond only to certain specific features of images. And this means that if one looks only at the output from these nerve cells, then one gets a representation of visual images in which two images that differ only in certain kinds of details will be assigned the same representation.

So if it is a representation like this that is used as the basis for storing data in memory, the result is that one will readily be able to retrieve not only data that matches exactly, but also data that is merely similar enough to have the same representation.

In actual brains it is fairly clear that input received by all the various sensory systems is first processed by assemblies of nerve cells that in effect extract certain specific features. And it seems likely that especially in lower organisms it is often representations formed quite directly from such features that are what is stored in memory.

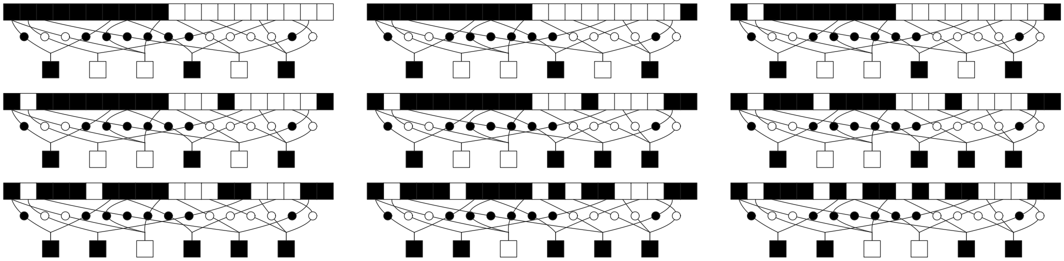
But at least in humans there is presumably more going on. For it is quite common that we can immediately recognize that we have encountered some particular object before even if it is superficially presented in a quite different way. And what this suggests is that quite different patterns of raw data from our sensory systems can at least in some cases still lead to essentially the same representation in memory.

So how might this be achieved? One possibility is that our brains might be set up to extract certain specific high-level features—such as, say, topological structure in three-dimensional space—that happen to successfully characterize particular kinds of objects that we traditionally deal with.

But my strong suspicion is that in fact there is some much simpler and more general mechanism at work, that operates essentially just at the level of arbitrary data elements, without any direct reference to the origin or meaning of these data elements.

And one can imagine quite a few ways that such a mechanism could potentially be set up with nerve cells.

One step in a particularly simple scheme is illustrated in the picture below. The basic idea is to have a sequence of layers of nerve cells—much as one knows exist in the brain—with each cell in each successive layer responding only if the inputs it gets from some fixed random set of cells in the layer above form some definite pattern.



One step in a very simple model of the way hash codes for arbitrary data might be generated by layers of nerve cells in the brain. The response of a single layer of idealized nerve cells to a sequence of progressively different inputs is shown. Each nerve cell fires and yields black output only if the inputs it gets from certain fixed positions match a particular template. The sequence of outputs from all the nerve cells can be used as a hash code, whose value tends to be the same for inputs that differ only by small changes.

In a sense this is a straightforward generalization of the scheme for visual perception that we discussed earlier in this chapter. But the point is that with such a setup detailed changes in the input to the first layer of cells only rarely end up having an effect on output from the last layer of cells.

It is not difficult to find systems in which different inputs often yield the same output. In fact, this is the essence of the very general phenomenon of attractors that we discussed in Chapter 6—and it is seen in the vast majority of cellular automata, and in fact in almost any kind of system that follows definite rules.

But what is somewhat special about the setup above is that inputs which yield the same output tend to be ones that might reasonably be considered similar, while inputs that yield different outputs tend to be significantly different.

And thus, for example, a change in a single input cell typically will not have a high probability of affecting the output, while a change in a large fraction of the input cells will.

So quite independent of precisely which features of the original data correspond to which input cells, this basic mechanism provides a simple way to get a representation—and thus a hash code—that will tend to be the same for pieces of data that somehow have enough features that are similar.

So how would such a representation in the end be used? In a scheme like the one above the output cells would presumably be connected to cells that actually perform actions of some kind—perhaps causing muscles to move, or perhaps just providing inputs to further nerve cells.

But so where in all of this would the actual content of our memory reside? Almost certainly at some level it is encoded in the details of connections between nerve cells.

But how then might such details get set up?

There is evidence that permanent changes can be produced in individual nerve cells as a result of the behavior of nerve cells around them. And as data gets received by the brain such changes presumably do occur at least in some cells. But if one looks, say, at nerve cells involved in the early stages of the visual system, then once the brain has matured past some point these never seem to change their properties much. And quite probably the same is true of many nerve cells involved in the general process of doing the analog of producing hash codes.

The reason for such a lack of change could conceivably be simply that at the relevant level the overall properties of the stream of data corresponding to typical experience remain fairly constant. But it might also be that if one expects to retrieve elements of memory reliably then there is no choice but to set things up so that the hashing procedure one uses always stays essentially the same.

And if there is a fixed such scheme, then this implies that while certain similarities between pieces of data will immediately be recognized, others will not.

So how does this compare to what we know of actual human memory? There are many kinds of similarities that we recognize quite effortlessly. But there are also ones that we do not. And thus, for example, given a somewhat complicated visual image—say of a face or a cellular automaton pattern—we can often not even immediately recognize similarity to the same image turned upside-down.

So are such limitations in the end intrinsic to the underlying mechanism of human memory, or do they somehow merely reflect characteristics of the memory that we happen to build up from our typical actual experience of the world?

My guess is that it is to some extent a mixture. But insofar as more important limitations tend to be the result of quite low-level aspects of our memory system it seems likely that even if these aspects could in principle be changed it would in practice be essentially impossible to do so. For the low levels of our memory system are exposed to an immense stream of data. And so to cause any substantial change one would presumably have to insert a comparable amount of data with the special properties one wants. But for a human interacting with anything like a normal environment this would in practice be absolutely impossible.

So in the end I strongly suspect that the basic rules by which human memory operates can almost always be viewed as being essentially fixed—and, I believe, fairly simple.

But what about the whole process of human thinking? What does it ultimately involve? My strong suspicion is that the use of memory is what in fact underlies almost every major aspect of human thinking.

Capabilities like generalization, analogy and intuition immediately seem very closely related to the ability to retrieve data from memory on the basis of similarity. But what about capabilities like logical reasoning? Do these perhaps correspond to a higher-level type of human thinking?

In the past it was often thought that logic might be an appropriate idealization for all of human thinking. And largely as a result of this, practical computer systems have always treated logic as something quite fundamental. But it is my strong suspicion that in fact logic is very far from fundamental, particularly in human thinking.

For among other things, whereas in the process of thinking we routinely manage to retrieve remarkable connections almost instantaneously from memory, we tend to be able to carry out logical reasoning only by laboriously going from one step to the next. And my strong suspicion is that when we do this we are in effect again just using memory, and retrieving patterns of logical argument that we have learned from experience.

In modern times computer languages have often been thought of as providing precise ways to represent processes that might otherwise be carried out by human thinking. But it turns out that almost all of the major languages in use today are based on setting up procedures that are in essence direct analogs of step-by-step logical arguments.

As it happens, however, one notable exception is *Mathematica*. And indeed, in designing *Mathematica*, I specifically tried to imitate the way that humans seem to think about many kinds of computations. And the structure that I ended up coming up with for *Mathematica* can be viewed as being not unlike a precise idealization of the operation of human memory.

For at the core of *Mathematica* is the notion of storing collections of rules in which each rule specifies how to transform all pieces of data that are similar enough to match a single *Mathematica* pattern. And the success of *Mathematica* provides considerable evidence for the power of this kind of approach.

But ultimately—like other computer languages—*Mathematica* tends to be concerned mostly with setting up fairly short specifications for quite definite computations. Yet in everyday human thinking we seem instead to use vast amounts of stored data to perform tasks whose definitions and objectives are often quite vague.

There has in the past been a great tendency to assume that given all its apparent complexity, human thinking must somehow be an altogether fundamentally complex process, not amenable at any level to simple explanation or meaningful theory.

But from the discoveries in this book we now know that highly complex behavior can in fact arise even from very simple basic rules. And from this it immediately becomes conceivable that there could in reality be quite simple mechanisms that underlie human thinking.

Certainly there are many complicated details to the construction of the brain, and no doubt there are specific aspects of human thinking that depend on some of these details. But I strongly suspect that there is a definite core to the phenomenon of human thinking that is largely independent of such details—and that will in the end turn out to be based on rules that are rather simple.

So how will we be able to tell if this is in fact the case? Detailed direct studies of the brain and its operation may give some clues. But my guess is that the only way that really convincing evidence will be obtained is if actual technological systems are constructed that can successfully be seen to emulate human thinking.

And indeed as of now our experience with practical computing provides rather little encouragement that this will ever be possible. There are certainly some tasks—such as playing chess or doing algebra—that at one time were considered indicative of human-like thinking, but which are now routinely done by computer. Yet when it comes to seemingly much more mundane and everyday types of thinking the computers and programs that exist at present tend to be almost farcically inadequate.

So why have we not done better? No doubt part of the answer has to do with various practicalities of computers and storage systems. But a more important part, I suspect, has to do with issues of methodology.

For it has almost always been assumed that to emulate in any generality a process as sophisticated as human thinking would necessarily require an extremely complicated system. So what has mostly been done is to try to construct systems that perform only rather specific tasks.

But then in order to be sure that the appropriate tasks will actually be performed the systems tend to be set up—as in traditional engineering—so that their behavior can readily be foreseen, typically by standard mathematical or logical methods. And what this almost invariably means is that their behavior is forced to be fairly simple. Indeed, even when the systems are set up with some ability to learn they usually tend to act—much like the robots of classical fiction—with far too much simplicity and predictability to correspond to realistic typical human thinking.

So on the basis of traditional intuition, one might then assume that the way to solve this problem must be to use systems with more complicated underlying rules, perhaps more closely based on details of human psychology or neurophysiology. But from the discoveries in this book we know that this is not the case, and that in fact very simple rules are quite sufficient to produce highly complex behavior.

Nevertheless, if one maintains the goal of performing specific well-defined tasks, there may still be a problem. For insofar as the behavior that one gets is complex, it will usually be difficult to direct it to specific tasks—an issue rather familiar from dealing with actual humans. So what this means is that most likely it will at some level be much easier to reproduce general human-like thinking than to set up some special version of human-like thinking only for specific tasks.

And it is in the end my strong suspicion that most of the core processes needed for general human-like thinking will be able to be implemented with rather simple rules.

But a crucial point is that on their own such processes will most likely not be sufficient to create a system that one would readily recognize as exhibiting human-like thinking. For in order to be able to relate in a meaningful way to actual humans, the system would almost certainly have to have built up a human-like base of experience.

No doubt as a practical matter this could to some extent be done just by large-scale recording of experiences of actual humans. But it seems not unlikely that to get a sufficiently accurate experience base, the system would itself have to interact with the world in very much the same way as an actual human—and so would have to have elements that emulate many elaborate details of human biological and other structure.

Once one has an explicit system that successfully emulates human thinking, however, one can imagine progressively removing some of this complexity, and seeing just which features of human thinking end up being preserved.

So what about human language, for example? Is this purely learned from the details of human experience? Or are there features of it that reflect more fundamental aspects of human thinking?

When one learns a language—at least as a young child—one implicitly tends to deduce simple grammatical rules that are in effect specific generalizations of examples one has encountered. And I suspect that in doing this the types of generalizations that one makes are essentially those that correspond to the types of similarities that one readily recognizes in retrieving data from memory.

Actual human languages normally have many exceptions to any simple grammatical rules. And it seems that with sufficient effort we can in fact learn languages with almost any structure. But the fact that most modern computer languages are specifically set up to follow simple grammatical rules seems to make their structures particularly easy for us to learn—perhaps because they fit in well with low-level processes of human thinking.

But to what extent is the notion of a language even ultimately necessary in a system that does human-like thinking? Certainly in actual humans, languages seem to be crucial for communication. But one might imagine that if the underlying details of different individuals from some class of systems were sufficiently identical then communication could instead be achieved just by directly transferring low-level patterns of activity. My guess, however, is that as soon as the experiences of different individuals become different, this will not work, and that therefore some form of general intermediate representation or language will be required.

But does one really need a language that has the kind of sequential grammatical structure of ordinary human language? Graphical user interfaces for computer systems certainly often use somewhat different schemes. And in simple situations these can work well. But my uniform experience has been that if one wants to specify processes of any significant complexity in a fashion that can reasonably be understood then the only realistic way to do this is to use a language—like *Mathematica*—that has essentially an ordinary sequential grammatical structure.

Quite why this is I am not certain. Perhaps it is merely a consequence of our familiarity with traditional human languages. Or perhaps it is a consequence of our apparent ability to pay attention only to one thing at a time. But I would not be surprised if in the end it is a reflection of fairly fundamental features of human thinking.

And indeed our difficulty in thinking about many of the patterns produced by systems in this book may be not unrelated. For while ordinary human language has little trouble describing repetitive and even nested patterns, it seems to be able to do very little with more complex patterns—which is in a sense why this book, for example, depends so heavily on visual presentation.

At the outset, one might have imagined that human thinking must involve fundamentally special processes, utterly different from all other processes that we have discussed. But just as it has become clear over the past few centuries that the basic physical constituents of human beings are not particularly special, so also—especially after the discoveries in this book—I am quite certain that in the end there will turn out to be nothing particularly special about the basic processes that are involved in human thinking.

And indeed, my strong suspicion is that despite the apparent sophistication of human thinking most of the important processes that underlie it are actually very simple—much like the processes that seem to be involved in all the other kinds of perception and analysis that we have discussed in this chapter.