PROBABLY APPROXIMATELY CORRECT

Nature's Algorithms for Learning and Prospering in a Complex World

LESLIE VALIANT

Probably
Approximately
Correct

Probably Approximately Correct

Nature's Algorithms for Learning and Prospering in a Complex World

LESLIE VALIANT

BASIC BOOKS
A Member of the Perseus Books Group
New York

Published by Basic Books,
A Member of the Perseus Books Group

A CIP catalog record for this book is available from the Library of Congress.
ISBN: 978-0-465-03271-6
ISBN (e-book): 978-0-465-03790-2
10 9 8 7 6 5 4 3 2 1

Contents

Summary

Algorithms are the step-by-step instructions used in computing for achieving desired results, much like recipes in cooking. In both cases the recipe designer has a certain controlled environment in mind for realizing the recipe, and foresees how the desired outcome will be achieved. The algorithms I discuss in this book are special. Unlike most algorithms, they can be run in environments unknown to the designer, and they learn by interacting with the environment how to act effectively in it. After sufficient interaction they will have expertise not provided by the designer, but extracted from the environment. I call these algorithms ecorithms. The model of learning they follow, known as the probably approximately correct model, provides a quantitative framework in which designers can evaluate the expertise achieved and the cost of achieving it.

These ecorithms are not merely a feature of computers. I argue in this book that such learning mechanisms impose and determine the character of life on Earth. The course of evolution is shaped entirely by organisms interacting with and adapting to their environments. This biological inheritance, as well as further learning from the environment after conception and birth, have a determining influence on the course of an individual's life. The focus here will be the unified study of the mechanisms of evolution, learning, and intelligence using the methods of computer science.

The book has the following simple structure. Chapters 1, 2, and 4 set the scene for the natural phenomena to which the quantitative computational approach is to be applied. Chapter 3 is an introduction to computer science, particularly the quantitative study of algorithms and their complexity, and describes the background for the methodology used. Chapters 5, 6, and 7 contain the resulting theory for learning, evolution, and intelligence, respectively.

The final chapters make some informal and more speculative suggestions with regard to some consequences for humans and machines.

Mathematics

The language of mathematics will be used, but only a little, and will be explained where used.

Chapter One

Ecorithms

In 1947 John von Neumann, the famously gifted mathematician, was keynote speaker at the first annual meeting of the Association for Computing Machinery. In his address he said that future computers would get along with just a dozen instruction types, a number known to be adequate for expressing all of mathematics. He went on to say that one need not be surprised at this small number, since 1,000 words were known to be adequate for most situations in real life, and mathematics was only a small part of life, and a very simple part at that. The audience reacted with hilarity. This provoked von Neumann to respond: "If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is."[1]

Though counterintuitive, von Neumann's quip contains an obvious truth. Einstein's theory of general relativity is simple in the sense that one can write the essential content on one line as a single equation. Understanding its meaning, derivation, and consequences requires more extensive study and effort. However, this formal simplicity is striking and powerful. The power comes from the implied generality, that knowledge of one equation alone will allow one to make accurate predictions about a host of situations not even conceived when the equation was first written down.

Most aspects of life are not so simple. If you want to succeed in a job interview, or in making an investment, or in choosing a life partner, you can be quite sure that there is no equation that will guarantee you success. In these endeavors it will not be possible to limit the pieces of knowledge that might be relevant to any one definable source. And even if you had all the relevant knowledge, there may be no surefire way of combining it to yield the best decision.

This book is predicated on taking this distinction seriously. Those aspects of knowledge for which there is a good predictive theory, typically a mathematical or scientific one, will be called *theoryful*. The rest will be called *theoryless*. I use the term theory here in the same sense as it is used in science, to denote a "good, effective, and useful theory" rather than the negative sense of "only a theory." Predicting the orbit of a planet based on Newton's laws is theoryful, since the predictor uses an explicit model that can accurately predict everything about orbits. A card player is equally theoryful in predicting an opponent's hand, if this is done using a principled calculation of probabilities, as is a chemist who uses the principles of chemistry to predict the outcome of mixing two chemicals.

In contrast, the vast majority of human behaviors look theoryless. Nevertheless, these behaviors are often highly effective. *These abundant theoryless but effective behaviors still lack a scientific account, and it is these that this book addresses.*

The notions of the theoryful and the theoryless as used here are relative, relative to the knowledge of the decision maker in question. While gravity and mechanics may be theoryful to a physicist, they will not be to a fish or a bird, which still have to cope with the physical world, but do so, we presume, without following a theory. Worms can burrow through the ground without apparently any understanding of the physical laws to which they are subject. Most humans manage their finances adequately in an economic world they don't fully understand. They can often muddle through even at times when experts stumble. Humans can also competently navigate social situations that are quite complex, without being able to articulate how.

In each of these examples the entity manages to *cope* somehow, without having the tenets of a theory or a scientific law to follow. Almost any biological or human behavior may be viewed as some such coping. Many instances of effective coping have aspects both of the mundane and also of the grand and mysterious. In each case the behavior is highly effective, yet if we try to spell out exactly how the behavior operates, or why it is successful, we are often stumped. How can such behavior be effective in a world that is too complex to offer a clear scientific theory to be followed as a guide? Even more puzzling, *how can a capability for such effective coping be acquired in the first place?*

Science books generally restrict their subject matter to the theoryful. However, I am impressed with how effectively life forms "cope" with the theoryless in this complex world. Surely these many forms of coping have some commonality. Perhaps behind them all is a single basic phenomenon that is itself subject to scientific laws.

This book is based on two central tenets. The first is that the coping mechanisms with which life abounds are all the result of learning from the environment. The second is that this learning is done by concrete mechanisms that can be understood by the methods of computer science.

On the surface, any connection between coping and computation may seem jarring. Computers have traditionally been most effective when they follow a predictive science, such as the physics of fluid flow. However, computers also have their softer side. Contrary to common perception, computer science has always been more about humans than about machines. The many things that computers can do, such as search the Web, correct our spelling, solve mathematical equations, play chess, or translate from one language to another, all emulate capabilities that humans possess and have some interest in exercising. Depending on the task, the performance of present-day computers will be better or worse than humans. But in regarding computers merely as our slaves for getting things done, we may be missing the point. The overlap between what computers and humans do every day is already vast and diverse. Even without any extrapolation into the future, we have to ask what computers already teach us about ourselves.

The variety of applications of computation to domains of human interest is a totally unexpected discovery of the last century. There is no trace of anyone a hundred years ago having anticipated it. It is a truly awesome phenomenon. Each of us can identify our own different way of being impacted by the range of applications that computers now offer. A few years ago I was interested in the capabilities of a certain model of the brain. In a short, hermit-like span of a few weeks I ran a simulation of this model on my laptop and wrote up a paper based on the calculations performed by my laptop. I used a word processor on the same laptop to write and edit the article. I then emailed it off to a journal again from that laptop. This may sound unremarkable to the present-day reader, but a few generations ago, who would have thought that one device could perform such a variety of tasks? Indeed, while for most ideas some long and complex history can be traced, the modern notion of computation emerged remarkably suddenly, and in a most complete form, in a single paper published by Alan Turing in 1936.[2]

Science prior to that time made no mention of abstract machines. Turing's theory did. He defined the mathematical notion of computation that our all-pervasive information technology now follows. But in offering his work, he made it clear that his goal went beyond understanding only machines: "We may compare a man in the process of computing a real number to a machine which is only capable of a finite number of conditions." With these words he was declaring that he was aiming to formalize the process of computation where a human mechanically follows some rules. He was seeking to capture the limits of what could be regarded as mechanical intellectual work, where no appeal to other capabilities such as intuition or creativity was being made.

Turing succeeded so well that the word computation is now used in exactly the sense in which he defined it. We forget that a "computer" in the 1930s referred to a human being who made a living doing routine calculations. Speculations that philosophers or psychologists entertained in earlier times as to the nature of mechanical mental capabilities equally dim in the memory. Turing had discovered a precise and fundamental law that both living and inert things must obey, but which only humans had been observed to exhibit up to that time. His notion is now being realized in billions of pieces of technology that have transformed our lives. But if we are blinded by this technological success, we may miss the more important point that Turing's concept may enable us to understand human activity itself.

This may seem paradoxical. Humans clearly existed before Turing, but Turing's notion of computation was not noticed before his time. So how can his theory be so fundamental to humans if little trace of it had even been suspected before?

My answer to this is that even in the pre-Turing era, in fact since the beginning of life, the dominating force on Earth within all its life forms *was* computation. But the computations were of a very special kind. These computations were weak in almost every respect when compared with the capabilities of our laptops. They were exceedingly good, however, at one enterprise: adaptation. These are the computations that I call ecorithms—algorithms that derive their power by learning from whatever environment they inhabit, so as to be able to behave effectively in it. To understand these we need to understand computations in the Turing sense. But we also need to refine his definitions to capture the more particular phenomena of learning, adaptation, and evolution.

Understanding learning has been one of my personal research goals for several decades. The natural phenomenon of young children learning is extraordinary. A spectacular facet of this learning is that, beyond remembering individual experiences, children will also generalize from those experiences, and very quickly. After seeing a few examples of apples or chairs, they know how to categorize new examples. Different children see different examples, yet their notions become similar. When asked to categorize examples they have not seen before, their rate of agreement will be remarkably high, at least within any one culture. Young children can sort apples from balls even when both are round and red.

This ability to generalize looks miraculous. Of course, it cannot really be a miracle. It is a highly reproducible natural phenomenon. Ripe apples fall from the tree to the ground predictably enough that one can base a universal law of gravitation on this phenomenon. Children generalizing successfully from their specific experiences manifest a similarly predictable phenomenon, which therefore also begs for a scientific explanation. I seek to explain this in terms of concrete computational processes.

The phenomenon of generalization has been widely discussed by philosophers for millennia. It has been called the problem of induction. I have found that as a scientist I have some advantages over philosophers: It is sufficient to aim to capture the fundamental part of a specific reproducible phenomenon. I need not explain all of the many senses in which the words induction or generalization have been used. Scientific discovery—for example, Johannes Kepler discovering his laws of planetary orbits—may have some commonality with the phenomenon of generalization exhibited by children learning words, but it may be a secondary and harder to reproduce by-product of a more basic and fundamental capability. Turing did not attempt to capture all the connotations that the word computing may have had in his day. He sought only to uncover a phenomenon associated with that word that had fundamental reality independent of any word usage.

What kind of explanation of induction do we need? Does it need to be mathematical? There is no better answer to this than what is implicit in the work of Turing himself. I have already referred to his successful mathematical formulation of computation. But he is also famous for the notion that is now known as the Turing Test, which he offered as a test for recognizing whether a machine can be considered to think. A simplified definition is as follows. A machine passes the Turing Test if a person, conversing with it via remote electronic interactions, cannot distinguish it from a person. The Turing Test is an important notion, and researchers in artificial intelligence have not succeeded in either building machines that can pass the test or in showing it to be irrelevant. However, it is an informal notion. Unlike Turing's mathematical definition of computation, it does not tell us how exactly to proceed in order to emulate thinking. As a result, it has not led to progress in artificial intelligence remotely comparable to the success of general computation.

Hence the right goal must be to find a *mathematical* definition of learning of a nature similar to Turing's notion of computation, rather than an *informal* notion like the Turing Test. After all, where would we be if Turing had given for computation only an informal definition? Let us think about that. What would have been an informal notion of the "mechanically computable" that would have sounded plausible in Turing's time? How about this: "A task is mechanically computable if and only if it can be computed by a person of average intelligence while at the same time doing a mundane but exacting task, such as eating spaghetti." Few could have disputed the *reasonableness* of such a definition. But I doubt such a definition in 1936 could have spawned the twenty-first century we see around us.

At the heart of my thesis here is a mathematical definition of learning. It is called the PAC or the probably approximately correct model of learning, and its main features are the following:[3] The learning process is carried out by a concrete computation that takes a limited number of steps. Organisms cannot spend so long computing that they have no time for anything else or die before they finish. Also, the computation requires only a similarly limited number of interactions with the world during learning. Learning should enable organisms to categorize new information with at most a small error rate. Also, the definition has to acknowledge that induction is not logically fail-safe: If the world suddenly changes, then one should not expect or require good generalization into the future.

The biology of living organisms can be described in terms of complex circuits or networks that act within and between cells. Our biology is based on proteins and the interactions among them. Our DNA contains more than 20,000 genes that describe various proteins. Additionally, the DNA encodes descriptions of the regulation mechanism, a specification of how much new protein of each kind is to be produced, or expressed. This overall regulation mechanism is absolutely fundamental to our biology, and is called the protein expression network. It is of enormous complexity. Even though many of its details remain to be discovered, we can ask: *How have these well-functioning, highly intricate networks with so many interlocking parts come into being?* I believe that all these circuits are the result of some learning process instigated by the interactions between a biological entity and its environment.

Life's interactions can be viewed in terms of either a single organism's lifetime or the longer spans during which genes and species evolve. In either case the information gained by the entity from the interaction is processed in some mechanical way by what I call an ecorithm. The primary purpose of the ecorithm is to change the circuits so that they will behave better in the environment in the future and produce a better outcome for the owner.

Human biochemistry is an important enough topic. However, our neural circuits, comprising some tens of billions of neurons, may be viewed as being involved in our personal experiences even more intimately. Our psychological behavior is controlled by these circuits. How do these circuits arise in evolution, and how are they updated during life? By the same arguments they too must be the result of information obtained from interactions, by ourselves or our ancestors, and incorporated in our genes or brain by some adaptive mechanism.

If biological circuits are fundamentally shaped by learning processes, there seems little chance of understanding them, or their manifestations in our psychology, unless we recognize their origins in learning. We may not yet know in detail the actual ecorithms used in biology on Earth. However, the fact that our behaviors have their origins in such learning algorithms already has implications.

Earlier I listed as two central tenets that the behaviors that need explanation all arose from learning, and that this learning can be understood as a computational process. These tenets are not offered here as mere unproved assumptions, but as the consequences of the assumption that life has a mechanistic explanation.

The argument that these tenets actually follow from the formulation of ecorithms goes as follows: I start with the mechanistic assumption that biological forms came into existence as a result of concrete mechanisms operating in some environments. These mechanisms have been of two kinds, those that operate in individuals interacting with their environment, and those that operate via genetic changes over many generations.

I then make two observations. First, ecorithms are defined broadly enough that they encompass any mechanistic process. This follows from the work of Turing and his contemporaries that established the principle, known as the Church-Turing Hypothesis, that all processes that can be regarded as mechanistic can be captured by a single notion of computation or algorithm. Second, ecorithms are also construed broadly enough to encompass any process of interaction with an environment. From these two observations one can conclude that the coping mechanisms of nature have no sources of influence on them that are not fully accounted for by ecorithms, simply because we have defined ecorithms broadly enough to account for all such influences.

To put this in a different way, the news reported here is that there is a burgeoning science of learning algorithms. Once the existence of such a science is accepted, its centrality to the study of life is more or less self-evident.

Of course, the reader should be cautious when confronted with purported logical arguments such as the one I just gave. Indeed, later chapters will address the general pitfalls of reasoning about theoryless subject matter. It is appropriate, therefore, to attempt to corroborate my proposition. Is there somewhere we can turn for a sanity check? The answer is machine learning, a method for finding patterns in data that are usefully predictive of future events but which do not necessarily provide an explanatory theory.

Machine learning is already a widely used technology with diverse applications. For example, companies such as Amazon and Netflix make recommendations to shoppers based on the predictions of learning algorithms trained on past data. Of course, there is no theory of which books or movies you will like. You may even completely change your tastes at any time. Nevertheless, using machine learning algorithms, it is possible to do a useful job in making such recommendations. Financial institutions likewise use machine learning algorithms, in their case, for example, for detecting whether individual credit card purchase attempts are likely to be fraudulent. These algorithms pick up various kinds of relevant information, such as the geographical pattern of your previous purchases, to make some decisions based on data collected from many past transactions. The development of the learning algorithms used may well be theoryful. But this again does not mean that fraud itself is theoryful. New kinds of fraud are being invented all the time. The algorithms merely find patterns in past credit card purchases that are useful enough to give financial institutions a statistical edge in coping with this area of the theoryless.

Much of everyday human decision making appears to be of a similar nature—it is based on a competent ability to predict from past observations without any good articulation of how the prediction is made or any claim of fundamental understanding of the phenomenon in question. The predictions need not be perfect or the best possible. They need merely to be useful enough. The fact that these algorithms are already in widespread use, and produce useful results in areas most would regard as theoryless, is good evidence that we are on the right track.

However, the idea of an ecorithm goes well beyond the idea of machine learning in its current, general usage. Within the study of ecorithms several additional notions beyond the learning algorithms themselves are included. First, there is the notion that it is important to specify what we expect a learning algorithm to be able to do before we can declare it to be successful. Second, using such a specification, we can then discuss problems that are not learnable—some environments will be so complex that it is impossible for any entity to cope. Third, there is the question of how broad a functionality one wants to have beyond generalization in the machine learning sense. To have intelligent behavior, for example, one needs at least a reasoning capability on top of learning. Finally, biological evolution must fit somehow into the study of coping mechanisms, but it is not clear exactly how, since traditional views of evolution do not exactly fit the machine learning paradigm. In studying ecorithms, we want to embrace all of these issues, and more.

The problem of dealing with the theoryless is ever present in our lives. Every day we are forced to put our trust in the judgment of experts who operate outside the bounds of any strict science. Your doctor and car mechanic are paid to make judgments, based on their own experience and that of their teachers. We presume that their expertise is the result of learning from a substantial amount of real-world experience and, for that reason, is effective in coping with this complex world. Their expertise can be evaluated by how well their diagnoses and predictions work out. In some areas we can evaluate performance, at least after the fact.

We are also exposed every day to commentators and pundits whose diagnoses and predictions are infrequently checked for ultimate accuracy. We hear about what will happen in politics, the stock market, or the economy, but these predictions often seem hardly better than random guessing.

In late 2008 Queen Elizabeth II asked a group of academics why the world financial crisis had not been predicted. She was not the only one asking this question. Was the crisis inherently unpredictable in some sense, or was the failure due to some gross negligence? After the crisis a substantial amount of public discussion pertained to this question. Is there a rational way of predicting rare events? Why do humans have so many intellectual frailties and behave as irrationally as they do? Why are humans subject so easily to deception and self-deception? Why do humans systematically delude themselves into thinking that they are good predictors of future events even if they are not?

Many reasons have been given for the difficulty of making predictions, and the mistakes that people are prone to make have been widely analyzed.[4] The following, for example, is an instructive argument. After any significant historical event numerous explanations of the causes are offered. These explanations can be so beguilingly plausible that we easily mistake them for actual causes that might have been detected before the events in question. We are then communally led into the belief that world events have identifiable causes and are generally predictable. Hence popular disappointment that the world financial crisis had not been better anticipated can be ascribed to widespread overexpectation and naïveté with regard to the possibility of making predictions.

This book departs from this approach and takes an opposing, more positive view. While making predictions may be inherently difficult, and we humans have our special failings, human predictive abilities are substantial and reason enough for some celebration. Humans, and biological systems generally, do have an impressive capability to make predictions. The ability of living organisms to survive each day in this dangerous world is surely evidence of an ability to predict the consequences of their actions and those of others, and to be prepared for whatever happens, and be rarely taken totally by surprise. In human terms, the fact that we can go through a typical day, one that may include many events and interactions with others, and be seldom surprised is testament surely of our good predictive talents. Of course, the domains in which we make these reliable predictions often relate only to everyday life—what other people will say or other drivers do. They are mundane, almost by definition. But even mundane predictions become mystifying once one tries to understand the process by which the predictions are being made, or tries to reproduce them in a computer.

From this viewpoint, the general disappointment that the world financial crisis had not been better predicted was not based entirely on naïve illusion. It was based on the well-justified high regard we have for our predictive abilities, and so it would be clearly to our advantage to identify why they failed. It may be that the world was changing in such a random fashion that the past did not even implicitly contain reliable information about the future. Or perhaps the past did indeed contain this information, but that it was somehow so complex that it was not practically feasible to dig it out. A third case is that prediction was indeed feasible, but the wrong algorithm or the wrong data had been used.

The study of ecorithms is concerned with delineating among these possibilities. Having the ability to make these distinctions among topics of everyday concern, such as predictions about the course of the economy, seems important. One may be able to do more than merely lament human frailties in this regard. Are there inherent reasons why reliable predictions are not possible regarding the course of a country's economy? Perhaps one can show that there are. It would then follow that there is no reason to listen to pundits other than for entertainment.

Computation allows one to construct concrete situations in which the world does reveal sufficient information for prediction in principle, but not in practice. Consider the area of encryption. If messages in the wireless connection of your home computer are encrypted, the intention is that if your neighbor listens in, he should not be able to get any information about what you are doing. Even if he listens in over a long period and does clever computations on the data he collects using a powerful computer, he should not be able to invade your privacy. This is another way of saying that the environment defined by your enciphered messages should be too complex for your neighbor, or anyone else, to make any sense of.

How can entities cope with what they do not fully understand? The simplest living organisms have had to face this problem from the beginnings of life. With limited mechanisms they had to survive in a complex world and to reproduce. Every evolving species has faced a similar problem, as do individual humans going through their daily lives. I shall argue that solutions to these problems have to be sought within the framework of learning algorithms, since this is the mechanism by which life extracts information from its environment. By the end of the book I hope to have persuaded the reader that when seeking to understand the fundamental character of life, learning algorithms are a good place to start.

Chapter Two

Prediction and Adaptation

Only adapt.
Adapted from E. M. Forster

"You never walk into a situation and believe that you know better than the natives. You have to listen and look around. Otherwise you can make some very serious mistakes."[1] This was a lesson that Kofi Annan, the former Secretary General of the United Nations learned, not on some far-flung diplomatic posting for the UN, but as a young man in St. Paul, Minnesota. He had arrived from Africa to study economics as an undergraduate. Inexperienced as he was with cold weather, when he first saw local students wearing ear muffs he thought they looked ridiculous. But after walking round the campus on a cold day, he went out to buy some for himself.

The logic of ecorithms has much in common with Annan's analysis. That logic emphasizes listening and looking around. It encourages caution in applying specialized expertise gained in one environment to another, and gives respectful deference to observed experience. It says that it is we who must seek to adapt.

Such an adaptive imperative is absent from most aphorisms. "Neither a borrower nor a lender be" urges one to act in a specific way rather than to adapt to one's environment. The pitfalls of following such nonadaptive advice are clear. While the advice may be good in some circumstances, perhaps those from which it was derived, in others it may not be.

Annan's strategy has the strength that it accepts that there are many possible worlds and warns against assuming that they are all the same. On the other hand, it is not too specific in prescribing a course of action. I shall argue that some of the most important phenomena of biology and cognition arise from general adaptive strategies akin to Annan's, empty as they may appear to be of any specific expert knowledge. Although such strategies as listening and looking are not fine-tuned to any particular environment, they may nonetheless be effective in any environment that has certain weak regularities hidden among all the complexities. I shall suggest that not only are they effective, but, further, they are integral to any explanation of life and culture as we witness these on Earth.

The new word ecorithm that I use to encapsulate these ideas derives from the word *algorithm* and the prefix *eco-*. An algorithm is simply any well-defined procedure. It is derived from the Latinized transliteration Algoritmi of the name of the mathematician Al-Khwārizmī, who worked in the House of Wisdom in Baghdad in the ninth century and authored an influential book on algebra. I invoke the word algorithm intentionally. In the domain in which it is most widely used, namely computer science, the standards of explicitness—of what is considered well defined—are high. In the words of computer scientist Donald Knuth, "Science is what we understand well enough to explain to a computer. Art is everything else we do."[2] I want to discuss evolution, learning, and intelligence in terms of algorithms that are unambiguous and explicit enough that they can be "explained to," and hence simulated by, a computer. The prefix *eco-*, from the ancient Greek word *oikos* meaning household or home (and which evokes the word ecology), signals that we are interested in algorithms that operate in complicated environments, especially environments that are much more complex than the algorithm itself. There is no contradiction in this. While the algorithm has to perform well in a complex environment, about which it has little knowledge initially, it has a chance of doing so if it is allowed to interact extensively with the environment and learn from it.

Within the realm of computation I make the following distinction. Algorithms as traditionally studied in mathematics and computer science are designed to solve instances of particular problems, such as solving algebraic equations or searching for a word in a text. All the expertise they need for their success is encoded in their own description by their designer. For example, Euclid in his textbook *The Elements* describes an elegant algorithm for finding the greatest common divisor of two numbers. (The greatest common divisor of 30 and 42 is 6.) His algorithm is correct and efficient in a specifiable sense even for arbitrarily large numbers. Its exact behavior on all pairs of numbers is entirely predictable, and no doubt foreseen by Euclid.

Ecorithms are special algorithms. In contrast with those designed to solve specific mathematical problems, these operate in environments that are not fully known to the designer, and may have much arbitrariness. Nevertheless, ecorithms can perform well even in these environments. While their success is foreseeable, the actual course they take will vary according to the environment.

The requirements that such an algorithm must meet to offer a plausible explanation of a natural phenomenon, such as biological evolution, are quite onerous. In particular, the algorithm must achieve its goals after a limited number of interactions and with the expenditure of limited resources. The concept of ecorithms and the general model of learning in which they are embedded, which I call probably approximately correct (or PAC) learning, insist on such quantitative practicality. The phenomena that they seek to explain are some of the most familiar to human experience: learning, resilience, and adaptation. I argue that broader phenomena still, in particular evolution and intelligence, are also best understood in these terms.

Evolution in biology is the idea that life forms have changed over time, and that these changes have resulted in the organisms seen on Earth today. Although closely associated with Charles Darwin, the roots of the idea reach back to antiquity and the recognition of evident family resemblances among the various animal and plant species. In more recent history, Charles Darwin's grandfather, Erasmus Darwin, wrote a treatise, *Zoonomia; or, The Laws of Organic Life*, arguing for this idea in the 1790s. This view was widely debated and controversial. William Paley, in a highly influential book, *Natural Theology* (1802), argued that life, as complex as it is, could not have come into being without the help of a Designer. Numerous lines of evidence have become available in the two centuries since, through genetics and the fossil record, that persuade professional biologists that existing life forms on Earth are indeed related and have indeed evolved. This evidence contradicts Paley's conclusion, but it does not directly address his argument. A convincing direct counterargument to Paley's would need a specific evolution mechanism to be demonstrated capable of giving rise to the quantity and quality of the complexity now found in biology, within the time and resources believed to have been available.

The main contribution of Charles Darwin was, of course, exactly so motivated.[3] He posited the outlines of an evolution mechanism with two primary parts, namely variation and natural selection, that he argued was sufficient to explain biological evolution on Earth without a Designer. In its simplest form, the theory of natural selection asserts that each organism has some level of fitness in a given environment and that it is capable of producing a range of variants of itself as its progeny. It then attributes evolution to the phenomenon that among the variants, individuals that have characteristics that constitute greater fitness will have a higher probability of having descendents in later generations than those with less fitness.

Among biologists there is broad consensus that Darwin's theory is essentially correct. Biochemical descriptions of the basis of life provide a concrete language in terms of which the actual evolutionary path taken by life on Earth may one day be spelled out in detail and explained. At present there are many gaps in our knowledge. The relationship between the DNA (the genotype) and the behavior and physiology of the organism (phenotype) to which it belongs is little understood. In spite of this, over the last 150 years Darwin's theory has become the central tenet of biology by virtue of substantial other evidence. Most recently, DNA sequencing has given incontrovertible experimental confirmation for the proposition that the varied life forms found on Earth are genetically related. Nothing that I will say here is intended or should be interpreted as casting doubt on this proposition. However, it remains the case that Darwin presented only an outline of a mechanism. It is not specific enough to be subject to a quantitative analysis or to a computer simulation. No one has yet shown that any version of variation and selection can account quantitatively for what we see on Earth. There is much that needs to be explained. Evolution has found solutions to many difficult problems that are of value to life on Earth. These include, among many others, locomotion, vision, flight, magnetic navigation, and echo location. Humans have managed to find artificial solutions to these physical challenges only after enormous effort.

The achievements of evolution are palpable and objectively impressive. The possibility remains that every version of variation and selection, as we currently understand these terms, would have needed a million times as long to yield existing life forms than is believed to have been available. Saying that evolution is a contest or even a struggle for life does not go far in explaining these facts. No theory is known that would explain how competition by itself leads to such spectacular achievements. Lotteries, singing competitions, and gladiatorial contests have not produced similar improvements or novelty. Evolution is a special kind of contest. How are we to go about understanding how this special contest, of whatever kind it is, has been able to produce the spectacular inventions that it has?

The term evolution evokes many images—indeed almost all facets of the history of life on Earth. I will restrict attention here to the one primary question of how complex mechanisms can arise at all within the limited time scale and resources in which they apparently have. The numerous other questions that are widely discussed by evolutionary theorists I regard as secondary to this one. The advantages offered by sex to evolution have been much debated, but evolution was far along when sex arrived on the scene. The intellectual challenge of understanding how peacocks could have acquired their elaborate plumage was much troubling to Darwin. But again, peacocks came along late in the game. In short, what I seek to address is a gap between the general formulation of natural selection as currently understood and any demonstration that any specific mechanism can account for the biological evidence we see around us. Every scientific theory has a gap that leaves some question unexplained. Evolution is by no means unique in that respect. Having a gap is no fatal flaw. However, the natural selection hypothesis as currently formulated has the gaping gap that it can make no quantitative predictions as far as the number of generations needed for the evolution of a behavior of a certain complexity. I believe that the time is ripe for working toward filling this gap. And I believe computer science is the tool for doing it.

This may be an unconventional claim, but I will argue that Darwin's theory lies at the very heart of computer science. Darwin's theory may even be viewed as the paradigmatic ecorithmic idea. One of computation's most fundamental characteristics is the separation between the physical realization of a mechanism and its manifest behavior. This is equally true of Darwin's theory. Although the fitness of a biological organism depends both on the biochemistry of the organism and on all the physical, chemical, and ecological factors present in its environment, the principle of natural selection makes no mention of biochemistry, physics, or ecology, and it incorporates no specific knowledge about the fitness of a particular species in a particular environment. We are driven to the almost paradoxical conclusion that organisms that perform at such a sophisticated level of expertise in physics, biochemistry, and ecology are the products of generic mechanisms that incorporate no such expertise. This striking contrast summarizes the basic challenge that ecorithms in general, and evolutionary algorithms in particular, need to overcome.

Given the central role that Darwin's theory now plays in biology, the following fact is more than a little disconcerting. From the first availability of digital computers many intelligent, curiosity-driven individuals have sought to simulate selection-based evolutionary algorithms in order to demonstrate their efficacy. These simulation experiments, carried out over more than half a century, have been disappointing, at least in my view, in creating mechanisms remotely reminiscent of those found in the living cell. In fact, these experiments are seldom quoted as corroborating evidence for evolution.

This failure cannot be ignored. It suggests that the natural selection hypothesis has to be refined somehow if it is to offer a more explanatory scientific theory. Further, the refinement will need to have a quantitative component that reflects the realities of the actual bounded numbers of generations, and bounded numbers of individuals per generation, that apparently have been sufficient to support evolution in this universe. That evolution could work in principle in some *infinite* limit is obvious and needs little discussion. But modern humans are believed to have existed for no more than about 10,000 generations and with modest population sizes for much of that history. Our predecessor species may have had not dissimilar statistics. Theories of evolution that assume unbounded resources for evolution, in generations or population sizes, or those that do not address this issue at all, cannot resolve the central scientific question of whether some instance of natural selection does fit the constraints that have ruled in this universe.

I am not the first to point out that there is a tension between the long time apparently needed for evolution and the limited resources that evidence from the physical sciences suggests have been available. No one was more aware of this tension than Darwin himself. In an attempt to find corroborating evidence for the long time scale he believed was needed for evolution, he looked to geology. In the first edition of *On the Origin of Species* he included an estimate of 300 million years for the time needed for erosion to have created the Weald formation in southern England.[4] This estimate immediately came under fire from the scientific community. Darwin omitted it, and any other such estimate, from subsequent editions. William Thomson (later Lord Kelvin) and other authoritative physicists of his day derided Darwin's estimate as impossibly too high even for the age of the Earth itself. Their arguments were based on applying the principles of physics as then understood to the question of the rate at which the Earth had been losing heat. This indirect line of attack on his theory of evolution gave Darwin much reason for concern. He wrote, "Thomson's views on the recent age of the world have been for some time one of my sorest troubles."[5] Kelvin's final published estimate was as low as 24 million years.[6] Physicists now estimate the age of the Earth, thankfully, to be much higher, about 4.5 billion years (and 13.8 billion years for that of this universe). Nevertheless, we still do not have a quantitative explanation of how life could have reached its current state even within this more extended period that is now allotted by the physicists, whether on the Earth or in the broader universe.

The theory offered here, of treating Darwinian evolution as a computational learning mechanism and quantitatively analyzing its behavior, is the only approach I know that addresses these questions explicitly. Previous mathematical approaches to evolution, such as those of population genetics, analyze the effects of competition on relative population sizes. For example, the famous Hardy-Weinberg principle from the early twentieth century shows that if reproduction is sexual and members of a population have two copies of each gene, as in humans, then diversity in the gene pool will be conserved in the following sense. If two variants of a gene exist in the population in a certain ratio and they are equally beneficial, then their ratio of occurrence in the population will converge to a stable value, with both variants continuing to occur. Analyses of relative population sizes such as this, however, do not address how more complex forms can come into being from simpler ones—this is the most fundamental question and the one that opponents of evolution usually target. One is not performing a service to science if one pretends to have a solution when one does not.

Advances in biology over the last half century have made concrete what needs to be explained in ways that were not known to the earlier pioneers of population genetics such as the eminent statistician Ronald Fisher. We now know that biological organisms are governed by protein expression networks. To understand evolution we need to have an explanation of how such complex circuits can evolve from simpler ones and maintain themselves in changing environments. The protein expression networks on which our biology depends are known to have more than 20,000 genes, and the outputs they produce depend in a highly complicated way on the innumerably many possible input combinations. These circuits define how the concentration levels of the many proteins in our cells are controlled in terms of each other. We can seek to describe them mathematically. For example, the amount produced of our seventh protein may depend on the concentrations of three others—say, the third, twenty-first, and seventy-third. The dependence is something specific, perhaps $f_7 = 1.7x_3 + 3.4x_{21} + 0.5x_{73}$, or more likely something else. But in any case it is some particular dependency $f_7(x_1, \ldots, x_{20{,}000})$ on all the available proteins and possibly on some additional parameters, such as temperature. Whatever this dependency f_7 is, it will change during evolution if some other such dependency becomes more beneficial to the organism because of changing circumstances.

What an evolutionary theory must do is explain how these dependencies are updated during evolution. How long will it take to evolve to a new function f'_7 if the environment changes so that the new function f'_7 is better than the old f_7? Of course, this only accounts for evolution with a fixed set of proteins. A successful theory must also explain the evolution of new proteins. I believe that this will need a similar kind of analysis but for a different kind of circuit.

Over the last several decades it has emerged that there are computational laws that apply to the existence and efficiency of algorithms that are as striking as physical laws. These computational laws offer a powerful new viewpoint on our world that meets the challenge that the facts of biology lay down in regard to both evolution and learning. The laws that are most relevant to these phenomena are different from those that are the most useful for programmers of digital computers, and they need to be investigated separately. This will be our point of departure.

Nothing here is intended as the last word on any of the topics covered. The approach I propose needs extensive development both internally and in interaction with the experimental sciences it relates to. The idea that mathematical equations are useful for expressing the laws of physics, that laboratory experiments can uncover the facts of chemistry, and that statistical analyses in the social sciences yield clues about causation are all widely appreciated. But the notion that natural phenomena can be understood as computational processes or algorithms is much more recent. I have no doubt, however, that this algorithmic viewpoint is poised to take its place among the more familiar arsenal of weapons used for uncovering the secrets of nature. I hope to offer here a glimpse of how this algorithmic perspective will come to occupy a central position in science. First, however, we must turn to the questions of the nature and scope of computational processes in general.