**Senior Author’s Note**

**In many ways, Nick Thompson is the idea candidate for the role of Posthumous Author. He has written, over fifty years on a constant theme, but the work has not been widely read. Most of the published work has appeared in obscure journals. In addition, he has piles of unpublished lecture notes delivered over the years to hundreds of students, whom he cruelly used as guinea pigs for the development of his ideas. Such a group of students might comprise a rump readership for a compilation of the work of such an obscure author, because, presumably, they wake up from time to time in the middle of the night wondering, “What was THAT all about?” Finally, and most important of all, Thompson has a talented and energetic co-author, Eric Charles, who is dedicated to the proposition that Thompson’s career, little more than a running gag at the moment, needs to be pulled together into a Final Statement.**

**So far as I can see, only one serious impediment stands in the way the writing of a successful posthumous compilation of Thompson’s “Thought”. Thompson is alive. This is an impediment in so many ways. The most obvious is that Thompson continues to think, and so Thompson’s Thought, although moving a lot slower than it used to, is still a moving target. Imagine the frustrations to be endured by a co-author, no matter how talented and energetic, trying to compile a career that isn’t finished. Such a co-author would be like a shipbuilder, asked not only to put to sea on a pile of lumber and build the ship during the voyage, but also to keep faith with a captain who keeps changing his mind about what kind of ship he wants to sail and where he wants to sail it.**

**Another problem is, of course, in whose VOICE should such a “posthumous” work be written? In a standard posthumous book, the editor/co-author uses the voice of the deceased author to fill in between the runes, but then uses his own voice to comment and critique what he has written there. But manners would seem to dictate that a posthumous author who is around to speak for himself should be allowed to.**

**Once obvious solution to these difficulties comes unbidden to a mind unfettered by social and legal niceties. We will say no more about that. We co-authors have agreed to put aside extreme solutions and try, for the moment, to capitalize on whatever slight benefits might be found to trying to co-author a posthumous book with a living author. For one thing, the labor might be shared. For another, a living author might better be able to answer such questions as, “What in God’s green earth did you mean when you wrote THAT!?” Finally, there’s a lot to be said for a *dialectic* between a posthumous author and his co-author. A reader’s enjoyment might just be enhanced by the opportunity to hear his doubts echoed by the co-author, and then, miracle of miracles, answered posthumously by the primary author.**

**Nick Thompson**

**Introduction**

During the last 60 years, thousands of scientific papers and hundreds of books have been devoted to evolutionary explanations of human and animal behavior. Many of these books have been enormously influential, not only with the public, but also with experts and students who read them as textbooks and collateral readings. Dawkins’s, *The Selfish Gene*, Dennett’s Darwin’s Dangerous Idea, ….. etc. have sold millions of copies between them. Yet, over that same period, skepticism about evolutionary theory in the United States and about its application to human behavior in particular has only increased. Polls across the country reveal that Darwin’s theory of evolution has less respect around the country than the theory of global warming. Why is this? Is it that Americans are dumb? Or is it that something about the manner in which this theory is often presented is giving rise to doubts.

We think that the theory of evolution has been seriously misrepresented in this literature and that the manner in which it has been applied to behavior has brought dishonor upon the whole. When it is properly understood, we will argue, most readers will see that while most of it is gloriously correct, it has been over extended. We think you will agree that, while only a fool would doubt the theory as a whole, its application to behavior has been careless and misguided.

**\* \* \***

Please take a moment to imagine the following scene:

***If you take a walk in the farm fields and pastures that surround the major cities in the Eastern states, you will probably encounter a “Killdeer”. A killdeer is a medium sized, plover-like bird that nests on the ground. Its young are cryptically colored and protect themselves from predators by scattering and squatting motionless in the grass. If you approach a mother killdeer when her young are about her, she will flutter away from you a short distance, dragging one wing as if it seriously hurt; if you follow her, she will do it again, repeatedly. When you have been led sufficiently far from her young, she will suddenly fly off, her wing miraculously healed. As she flies, she gives a plaintive but triumphant call, “Killdeer! KillDeer! Killdeer! You, like thousands of potential predators before you, have then been fooled by the "broken wing display," a pattern of behavior in which the parent bird feigns injury and draws you away from her young.***

We can explain the killdeer’s behavior in many ways that are not evolutionary explanations. "The killdeer performed this behavior because you approached her nest" - a situational explanation. "The killdeer performed this behavior because sensory cells in the eyes are connected in a certain way to the motor areas of the brain" - a physiological explanation. "The killdeer performed the behavior because genetic and environmental events in its individual past made it ready and able to produce this behavior at this time" - a developmental explanation. All of these are obviously not Darwinian ways of explaining the behavior of the killdeer. What then is a Darwinian explanation?

Here is an example of one kind of Darwinian explanation one often encounters often in the popular and sometimes, even, in the scientific literature.

The killdeer performed this behavior at this time, because in the history of the interbreeding population of killdeers of which this killdeer is a member, those killdeers that performed this behavior under such circumstances had more offspring than those killdeers that did not, with the result that most or all killdeers now alive perform this behavior.

What characterizes these sorts of Darwinian explanation? Perhaps their most striking feature is that they are ***stories***, historical narratives, that explain the way creatures are today by reference to a particular course of past events. Asked to explain why this killdeer is the way it is today, a Darwinian story explains by referring to a particular history of killdeers. Moreover, these stories involve a rather odd shift in level of organization. Having asked a question about *a* killdeer, we are given an answer in terms of a history of populations of killdeers. Moreover, these stories are quantitative, in that they explain by reference to a history of changes in the relative frequencies of different traits. Gathering evidence concerning these stories requires the counting of something, over a long period. Discovery of one killdeer now, or even during the history to which the explanation refers, that does (or does not) perform the behavior would not cause us to doubt the validity of this particular Darwinian story.

In fact, were you to ask for a Darwinian explanation of *any* particular trait of a particular organism, you might receive an explanation with all of these same odd characteristics. Thus, we can give the generic form of a complete, well-formed Darwinian story:

Trait T in species S evolved over the history of the species because, in each generation, those individuals bearing the more adapted traits had more offspring (than those individuals bearing less adapted traits), and the resulting offspring tended also to have the adapted traits, and Trait T was an adapted trait.

If this is how we understand a Darwinian explanations of behavior, we should not be surprised that some people doubt them. Ostensibly, gathering evidence for a Darwinian Story requires going back in time and collecting fossils until one finds a population that contains killdeers (or proto-killdeers) which do not perform the behavior. One would then need to observe the gradual elimination of those birds in favor of killdeers which do perform the behavior. But how are we to determine the behavior of long dead killdeers? Even if we could determine their behavior from fossils, how are we to discover a large enough random sample of fossils, widely enough distributed in time and space, to observe the transition from predominance of non-‘broken-wing-feigning’ birds to a predominance of ‘broken-wing-feigning’ birds? Given all these difficulties, it is no wonder so many people are pessimistic about the possibility of ever validating Darwinian explanations, in general, and Darwinian explanations of behavior, in particular. Moreover, perhaps you agree with us that an explanation, to be useful, must at least have the *possibility* of being validated. If so, then you would not be unwise to view the entire project of Darwinian explanation with some skepticism.

But you would also be unwise to reject Darwinian explanation because of the weaknesses uncovered by our description of Darwinian Stories. Why? Because Darwinian Stories trivialize the theory. True, Darwinian Stories are often consistent with the theory; true, they often raise challenging questions and lead to interesting field work. True, they are therefore often a useful tool in scientific research and in public presentations of that research. However, they fail entirely to capture the grandiosity of Darwin’s project. Darwin brought under one explanatory roof the vast array of biological phenomena. There was no field of biology before Darwin provided a point of view from which it could be unified. “Nothing in biology makes sense except in the light of evolution,” said Dobzshanky, and Dobzhansky knew his biology. If you wanted to play with Darwinism, you might be satisfied with Darwinian Stories; but if you want to truly understand it, if you want to claim to have rejected it OR accepted it, you must consider the full range of phenomena that Darwin’s Theory of the Origin of Species was called upon to explain.

**Descriptions and Explanations**

The goals of this book are 1) to determine what it is exactly about behavior that Darwinian Theory tries to explain with the theory of evolution, and how the theory explains it, and 2) demonstrate how clarity on that matter helps us better understand and evaluate the implications of evolutionary theory. So-stated the project seems enormous. It would put us into competition with the most eminent of evolutionary biologists, with the likes of Ernst Mayr who polished off this 50 career as an evolutionary biologist with a enormous work of 900 close-typed pages. Therefore, we will narrow our focus to the evolutionary explanation of behavior. As any reader of the popular press knows, the evolutionary explanation of behavior has, itself, become a vast industry. The last quarter century has spawned an entire new field, Evolutionary Psychology, with at least a dozen research centers, and hundreds of researchers, dedicated to the evolutionary explanation of human behavior mostly, but based on explanations drawn from the study of animal behavior generally.

But before we can we can begin even that more limited project we must have some agreement on what we mean by “description” and “explanation.” Well, if not agreement, at least you must be clear what we, the authors, mean when we use those terms.

To anticipate slightly, this distinction will be crucial because one of the main criticisms of the theory of evolution is that it involves circular reasoning, meaning that its crucial terms are inter-defined in a manner that makes it impossible for the theory to actually *explain* anything. As we will later discuss, it is true that some common ways that professionals and laymen think about evolution do indeed suffer from a circularity problem. However, we believe that the theory of evolution, as Darwin conceived of it, *does not* suffer from a circularity problem. Alas, once these issues are clarified, we will find that the problem Darwin’s theory *does* suffer from is that it is wrong.

Yes…Wrong!

But not wrong in the way that vehement critics of evolution seem to think. Darwinian Theory is wrong in a much more limited sense – empirical evidence shows that a comprehensive explanation for adaptation will require the inclusion of other explanatory principles, to complement the explanatory power of natural selection.

The development of these ideas will take time, and we must get the preliminaries out of the way before we can even begin. Before we can talk about what Darwinian Theory explains and how it explains it, we must be clear about what explanations are, and how the process of explanation works.

**How Explanations Differ from Descriptions**

One way in which explanations are often understood in everyday conversation is through a contrast with description:--A description is understood as a simple statement of a fact, whereas an explanation is an interpretation. A description simply says what happened, whereas an explanation says why it happened. The description is, by this distinction, an objective statement; the description makes one person’s direct observations available to others, while explanation is speculative. As such, many people feel that the danger of confusing description and explanation is the danger of mistaking objective fact for subjective bias. However, this intuitive method of distinction has been criticized, because it is often difficult to distinguish descriptions from explanations in terms of objectivity. Peoples’ attempts to describe events are notorious for smuggling in interpretations, and their explanations are notorious for looping-back to make assertions about what was or was not observed. Does this intertwining of explanation and description occur because some people are sloppy and imprecise? Or is it because explanations and descriptions more similar than they originally seemed?

To illustrate the dilemma, consider the following thought experiment:

An observer sits in front of you, a few feet away. You hold a flat object, such as a book, at chest height in front of you with your left hand, so that the flat side is displayed to the observer. You hold some small object, such as an eraser, at head height in your right hand, so that if it were released it would fall between you and the book. Next, you release the object. Shortly thereafter it hits the ground. How should the observer *describe* what happens?

Most people would consider "The eraser fell behind the book" as a completely objective, i.e. not-speculative, description of what transpired; but really such a description ascribes a lot more to the situation than what the observer actually saw. The eraser, after all, was not in the observer's view continuously. The observer saw it drop as far as the book. Then it reappeared below the book and continue to the floor. What the eraser did while behind the book is, *for the observer*, a matter for speculation. The observer assumes that the eraser behaved as all "falling" objects behave; that is, it simply continued to accelerate in a downward direction. While this assumption is consistent with the observed facts, the observed facts are also consistent with an infinite number of other ways in which the eraser might have moved. The eraser might have slowed down and then sped up behind the book; the eraser may have done a dipsy-doodle behind the book, etc. Thus it would seem, upon further reflection, that “The eraser fell behind the book” *is* interpretive. What at first seemed like pure description appears to be, at least in part, an *explanation* of the path of the eraser

A observer might try to escape this difficulty by piling on details in an attempt at clarification. S/he might now say something like, "Oh, all right! The eraser followed a straight downward course until it passed out of sight behind the book. Then it reappeared moments later, moving more rapidly, and continued its downward path to the ground." Such a statement is an improvement, in that it ascribes no particular behavior to the eraser while behind the book, but it too has its problems. First, it assumes that the eraser that disappears behind the book is the *same* eraser that reappears below the book. “Identity” is in itself a concept used to explain the similarity of observations made at different times. At a magic show, for example, we might not be surprised to learn that the eraser that hit the ground was not the one that was dropped.

The second problem the observer’s attempt at a more accurate description is that it is not a very satisfying description. It is linguistically awkward, and does not convey to the listener an accurate image of what occurred. Most readers, upon encountering such a verbose description would ask, "Do you mean that the eraser fell behind the book?" Thus, even in this extremely simplified example, the functions of description and explanation are not readily distinguished in terms of an objective-subjective distinction. “The eraser fell” seems to be describing some aspects of what happened, while postulating others.

Working through thought-experiments like the one above leads us to conclude that all descriptions, particularly satisfying ones, are inevitably explanatory and that all explanations are descriptive. And yet, you cannot explain something until you have something to explain – so all explanations must be based on prior descriptions. The only reasonable conclusion, if you take both of these claims at face value, is that all explanations are based on prior explanations! The distinction between description and explanation concerns their position in an argument, not their objectivity or subjectivity in some enduring sense. Whether a statement is explanatory or descriptive depends upon the understandings that exist between the speaker and his or her audience at the time the statement is made. *Descriptions are explanations that the speaker and the audience take to be true for the purpose of seeking further explanations*.[[1]](#footnote-1) Thus, if the question is "Why did the eraser fall?”, then “the eraser fell" is taken for granted, and becomes a description for the purposes of a subsequent explanation, such as, "Because I let go of it." On the other hand, if the question is "Why is the eraser, which used to be in my hand, now on the floor?”, then "The eraser fell" is explanatory – it is not taken for granted in the question.[[2]](#footnote-2)

**Circular and Not-quite Circular Reasoning**

If our analysis is correct, then the distinction between explanation and description takes on an entirely new importance in science. When we distinguish between explanation and description, we are not distinguishing between objective-empirical-observation on the one hand and subjective-theoretical-interpretation on the other. Instead we are distinguishing between that which we take for granted in a scientific argument and that which is in dispute, between what we are arguing *from* and what we are arguing *about*. Thus the risk of confusing explanation and description is not the risk of introducing bias; rather it is the risk of circular reasoning, the risk of recycling a description of a phenomenon as its own explanation.

**Circular reasoning.**

Circular reason takes a variety of forms. In its most blatant form, it is called the “naming fallacy.” Venerable psychologists’ joke:

The worried young man leaned across the desk and appealed to his therapist, “Doc, I'm 35 years old and I am still unmarried. How come?" The wise doctor scratched his head. After a moment of deep thought he replied, "Because, son, you are a bachelor." The patient sank back into his chair with awe and relief. "Ah," he said. "Now I see."

The relief the young man experiences from this "explanation" may well be real. It *is* nice to know that there is a name for one's problems. That there is a name implies, for one thing, that there are many other people like oneself, and such information seems always to be comforting. The young man thinks, "This is not a unique problem, I am just a bachelor," and goes about his business with a happier heart.

However, such relief is the philosophical equivalent of a placebo, and it may be short-lived. Knowing that he is a bachelor tells the young man nothing about his predicament that he did not already know. He knew that he was unmarried, and that is *all* that it means to say one is a bachelor. Moreover, he has learned nothing that might help him find a solution to the problem. Compare, for instance, the consequences of the Doctor having said, "Has it occurred to you that you might be gay?" Such an explanation might lead the young man to seek partners from a new category of people and thus might, if correct, lead to a solution to his problem.

**Not-quite circular reasoning.**

You may want to accuse us of having introduced a straw man into the discussion. Nobody, you might say, would ever take seriously an explanation of the form, “un-married, because a bachelor.” You would be correct, in that such an obvious example of circular argumentation would likely be rejected. However, it can be very difficult to evaluate whether explanatory claims are circular when the explanation is filled with jargon, and when descriptions become longer and more involved. Evolutionary explanations often contain much jargon, invoked to describe complexly described phenomenon – and hence provide the perfect context in which circular reasoning can lurk. But for now, let us look at a more mundane example of circular reasoning:

An old friend called me up on my Birthday to wish me well. I asked him how he was and he replied “Better!”

“Better from what?" I asked.

He explained that for many years he had suffered from midwinter periods of unhappiness and that finally he had gone to see a psychiatrist about it. The Good Doctor had informed him that he suffered from Seasonal Affective Disorder. That, and some pills designed to elevate his mood, seemed to satisfy him as an explanation.

But, of course, Seasonal Affective Disorder means *only* that your mood (your “affect”) changes over the course of the seasons – and the most typical change in mood is a drop in happiness during the cold months. When you cut past the jargon and the complexity of the description, this friend readily had accepted a circular explanation: “You are unhappy in midwinter, because you are the type of person that is unhappy midwinter!”

Actually, to be honest, the “Seasonal Affective Disorder” explanation is not quite as bad as the “Bachelor" explanation. It is not wholly empty. It is actually an example of a more subtle explanatory fault called a “Molièrization”. The name comes from the name of the French playwright, Molière. Molière wrote a play called *The Imaginary Invalid*. It contained a character who explained the sleep-inducing effects of opium by reference to the drug's "dormitive virtue." Since "dormitive virtue" means "something that causes sleep”, the explanation is equivalent to saying that "opium causes sleep, because it contains something in it that causes sleep." This is often presented as *the* classic example of circular garbage. However, it is not as viciously circular as it initially appears. In fact, the physicians in Molière’s play were not called upon to explain why opium produces sleep, but rather why a particular person fell asleep shortly after taking opium. An explanation in terms of the dormitive virtue of opium does, in fact, rule out some alternative explanations for the patient’s current sleep. For example, the doctor’s explanation asserts that the opium is the cause of the sleep, rather than the late hour or a sleeping pill taken previously..

Thus, Dormitive Virtue explanations are not always totally worthless. Their value depends upon the relation between the question that is asked and the answer that is given. Therefore, to decide whether an explanation is perniciously circular, we have to examine carefully what is understood by the questioner. Or, to make it impersonal, we have to examine closely the description taken for granted by the question.

**Explanations as the Invocation of Models**

If the distinction between descriptions and explanations is merely one of logical priority, then a proper understanding of description and explanation is to be sought not in the distinction between them, but in the properties they share. An analysis of the falling-eraser example suggests that a common property of explanations and descriptions is to classify and to organize information about a phenomenon under observation.

Consider, once again, the statement "The eraser fell behind the book." The statement implicitly classifies certain witnessed events together, while separating other events. The path of the eraser above the book and the path below the book are classified together and summarized by the term "fell." All the separate blurred, fleeting images of the descending eraser are classified together as "the eraser." The observer does not mention the title of the book behind which the eraser fell, nor does he tell us how he feels about his eraser-book experience, nor does he tell us the air temperature of the room in which the event occurred. Implicitly these details are classified apart, classified as non-relevant.

The "engine" that does this classification is an analogy to another process or event that serves as a model.[[3]](#footnote-3) Thus, to understand explanations, we must understand models. We use the term in the everyday sense, such as a *model* train one might give a child, but we will add some technical language. To explain the soon-to-come jargon, we need a familiar example that we can use as an analogy to help us understand less familiar models (i.e., we need a “model” model). The model train will serve nicely:

Imagine a child playing with a toy train under the Christmas tree – a Lionel electric train steam locomotive, which is a model of the real thing. Because it is a *model* of a real train, the child can expect it to do many things in the same way that a real steam locomotive on real tracks does them. For instance, the train tracks have switches. The child can guide the train over the switch again and again, exploring how the switch redirects the train from one track to another. Playing with this feature, she can learn much about the boundary of the switch: when a train is on *this* track vs. when it is on the other. For that purpose, a toy locomotive is every bit as good as a real steam locomotive. However, for some purposes, of course, it is not any good at all. She can learn very little about steam as a propellant by looking at the toy engine, which of course is not propelled by steam at all.

The child is not using an abstraction of a locomotive; she is using a particular locomotive. So, bear in mind that a model is not (on our understanding of it) an abstraction; it is a particular. If I was standing in a train yard, and asked my companion for an explanation of how the yardmen keep the trains going where they are supposed to go, she might explain by reference to the switches on toy trains. When using this model, we assert, she imagines a very particular thing standing in for the objects and events she is trying to explain, e.g., the *particular* toy train set she played with on her 4th birthday, or the *particular* toy train set she paid a quarter to control at the fairgrounds. She chose to invoke the model train as an explanation, because the model and the thing-to-be-explained are clearly similar in some respects, and she thinks her understanding of the model-thing will carry over to unobserved aspects of the thing-to-be-explained. At the same time, she is well aware that some aspects of the model-thing will *not* carry over to the thing-to-be-explained.

Let us now compare our model model to the previous example, the case of the eraser. The model invoked by our observer is an instance of the familiar experience of watching an object fall, with an *un*obstructed view. Yesterday our observer saw an apple fall from a tree to the ground; today he sees the eraser descend behind the book. By offering “the eraser fell” as an explanation for the latter event, he uses his past apple-experience as a model for the present eraser-experience; he asserts, on the basis of observed similarities, that the *unseen* behavior of the eraser is like the seen behavior of the apple.

Thus, the effect of applying a model is twofold: First, it organizes known information about the phenomenon to be explained; second, it generates hypothetical predictions related to the phenomenon to be explained. Another way to say this is that an explanation has both basic and surplus implications. *Basic implications* are those things that the observer already knows to be true about the phenomenon, and which he has chosen the model to accommodate. In the case of the eraser, the basic implications of the model include the observer’s description of the visible details of the eraser demonstration: the release of the eraser, its path down to the notebook, the path of the eraser below the notebook, and its impact upon the surface below. *Surplus implications* are those for which the observer has no evidence, but that are implied by the application of the model. When one applies the falling-apple model to the eraser, these implications include the hypothesized identity of the above-book eraser and the below-book eraser, as well as the hypothesized path behind the notebook. These implications are “surplus” because they are above and beyond what was actually observed. It is the very presence of this surplus implication that distinguishes a description from an explanation.

Every explanation further sorts surplus meaning into two additional categories. Just as we don't expect the model train-set to display *all* aspects of the real locomotive, the observer doesn't expect the model event to be exactly like the event to be explained. *Intended surplus**implications* are those aspects of the model that the explainer *intends* to apply to unobserved aspects of the current situation. In this case, the explainer asserts that the fall of the apple when continuously in view is the same as the fall of the eraser behind the book – in terms of a continuous acceleration, a straight line of descent, and retaining the same identity the whole way down. *Unintended surplus implications* are those aspects of the model the explainer does not intend to apply to the current situation. In this case, the explainer is not asserting, for example, that the eraser will be red, or that it will bruise when it lands.

**Examining the Model-model**

We proposed that explanations invoke models we are more familiar with to account for aspects of situations we are less familiar with. We then invoked the model train to help us understand the case of explanations in general. If we are being consistent, then the well-understood model-train model should provide insight into the less-well-understood generic case.

The distinctions that we have laid out between types of implications gives us a formal way to talk about the fact that all models, even the very best, most perfectly chosen models, are expected to differ from the situation they are invoked to explain. This tells us one thing about models in general that might not have been obvious: *Differences between the model and the thing being modeled do not necessarily make a model bad.* Further examining our understanding of the model-train model leads us to an even more specific conclusion: *Differences between the model and the thing being modeled are bad* only *if the differences are crucial to the discussion at hand*; that is, only if the difference is part of the *intended* surplus meaning. Recall: If I invoke the Lionel train when asked to explain switches, it is a good model; but if I invoke the Lionel train when asked to explain about steam as a propellant, it is not useful. This is because the Lionel train is not intended to be a model of steam propulsion.

Explanations often guide further investigation, and this is the crucial role of explanations in science. Sometimes, when researchers disagree on a very fundamental level about the phenomenon under investigation, they might argue over the basic meaning of a model. However, the normal state of affairs is that a group of scientists agree on the basic implications of a model, and then they get busy investigating the intended surplus implications. They look to try to confirm, or refute, what an explainer asserted about yet-to-be-observed aspects of a system. For models that have been investigated, it is often useful to further distinguish between *verified*and *unverified*surplus meaning. Scientists, as a general rule, hope that all intended implications of their models will one day be verified.

This breakdown of the content of an explanatory model is illustrated in Figure 1.1.

Model Situation

(Offered in Explanation)

Surplus Meaning

(asserted similarities)

Basic Meaning

(*a prior* similarities)

Indented

Surplus Meaning

Unintended

Surplus Meaning

Verified

Intended Meaning

Unverified

Intended Meaning

*Figure 1.1*. The effect of offering a model of a phenomenon is to organize information about that phenomenon into various categories: first, to separate basic from surplus meaning, i.e., to separate what is known about the phenomenon from what is implied by the explanation; second to separate surplus meaning into that which is intended and that which is *un*intended; and finally, when the explanation is tested, to separate intended surplus meaning that is *verified* from that which is *unverified*.

When you start breaking down the meanings of models, it is clear that any one of numerous particulars could serve as a model for our eraser experience. The model might be a different eraser, or a set of keys, or a coin. In fact, when we say "the eraser fell" we mean it behaved like any of an infinite number of unsupported objects, all of which differ substantially from an eraser… except in so much as they descend continuously, accelerating at a certain rate, and retaining their identity during their descent. However, the fact that any one of a class of particulars could have been used does not mean that a single particular was not chosen. In fact, it is the relation between the particular model being invoked, and the class of alternative model-objects that gives us intuitive insight into the intended and unintended meanings. If our explainer says that the eraser “fell like an apple off a tree,” then, typically, *our* experience of a falling apple can substitute adequately for a more detailed description of the explainer’s past experience.

**Levels of Explanation**

So far we have argued for two points about explanations: First that they *take descriptions for granted*, and second that they *use models* to organize known and hypothetical information about the phenomena they explain. We have also argued that there is no essential difference between the contents of a description and the contents of an explanation, and so we could just as easily have asserted that explanations *take other explanations for granted*. Together, these two ideas suggest that explanations may be stacked on top of each other, or nested into interconnected sets. Such a structure is called a *theory*. The simplest theories consist of a nested set of explanations, each member of the set being taken as a description for the purposes of another explanation, and each taking other explanations as descriptions to be explained. This structure is shown in Figure 1.2 below.

which explains

(i.e., takes as a

description)

is the model for

**Phenomenon 1**

*Model 1*

**Phenomenon 2**

*Model 2*

**Phenomenon 3**

*Model 3*

**Phenomenon 4**

which explains

is the model for

which explains

is the model for

*Figure 1.2*. A *theory* represented as a hierarchy of explanations in which each explanation makes use of a model to explain the explanations below it.

In the case of our eraser, "the eraser fell" is descriptive for the purposes of an explanation making use of the theory of gravity, and the theory of gravity is itself descriptive for the purposes of Einstein's theory of relativity. Furthermore, "The eraser fell" is explanatory for the purposes of the movements of the eraser, which are in turn explanatory for the purposes of the eraser-like blur above and below the clipboard. Each of these explanations assumes and answers a question. The question, “Why did I see a grey blur over there?” is answered by, “Because there was an eraser descending behind the book!” The question, “Why did the eraser descend behind the book?” is answered by, “Because the eraser fell!” The question, “Why did the eraser fall?” is answered by, “Because of the attraction of gravity!” This hierarchy is shown in Figure 1.3.

which explains

(i.e., takes as a

description)

is the model for

**General Relativity**

*Deformed Surface*

**Gravity (i.e., movement towards the center of large objects)**

*Attraction*

**Falling Objects**

*Apple off a Tree*

**What the eraser did**

which explains

is the model for

which explains

is the model for

*Figure 1.3*. At the top of this explanatory tree, Einstein’s General Relativity offered a deformed surface (bent space-time) as a model to explain gravity – that objects moving towards the center of mass of other objects. Gravity – that objects move towards the center of mass of other objects, is in turn an explanation that uses attraction as a model to explain falling objects. The apple going from the tree branch to the ground is in turn a model for falling objects, which explains what we saw the eraser do.

**Evaluating Explanations**

To telegraph a bit more explicitly: One of the main criticisms of Darwin’s theory of evolution is that it cannot be proven in the way we ordinarily think about proof. Darwinian explanations imply historical changes in the relative frequency of different types, and there is often no way of verifying the occurrence of such historical changes. Beyond that, even many prominent supporters of the theory of evolution have argued cogently that we cannot rationally evaluate Darwinian explanations *of behavior* (e.g. Steven J. Gould). But in the course of our examination of explanations in general, we have learned some things that should help moderate this pessimism. In the first place, we have learned that explanations do not always tell the whole truth; an explanation can be useful even if some of its implications are unproven or false. Indeed every scientific model will have implications suspected to be false from the start (the unintended surplus meaning), as well as implications we hope are true, but have no evidence for (the intended, but unverified, surplus meaning). Second, while we might want to think that unverified models are somehow defective, they are actually better in some ways than fully verified models: The intended-but-unknown aspects of our models constitute the driving force of science, motivating the search for evidence. These two points are reassuring. We cannot dismiss a Darwinian explanation of the killdeer’s behavior simply because it implies more than we currently know about the killdeer, or even because it implies more than we could ever know about the killdeer.

But we also want to think that science is about finding the truth about nature. So, having opened the door for falsehood-telling in scientific explanations, we have either to relinquish the notion of science as a truth-telling profession *or* we have to work out some rules for distinguishing when our explanatory story telling is useful and creative, and when it is just dangerous and sloppy. Opting for the latter option, we propose that evaluating an explanation requires two crucial steps.

**Step one: Specify the explanation.**

We cannot really evaluate an explanation unless we have specified it; that is, unless we know the answer to three questions about it.

***1. What does the explanation take as a description?*** What state of affairs does the explanation take for granted? Note that, since descriptions are inevitably also explanatory, understanding what an explanation takes as a description often entails understanding the theory – the nested structure of explanations – of which that explanation is part.

***2. What model does the explanation use?***Understanding the model means thoroughly understanding the process to which a metaphor is being made.

***3. What surplus meanings are generated by the model?*** What does the model imply about the phenomenon of interest that we do not already know? Understanding the surplus implications means sorting them into those which are unintended, and will therefore be unproductive in terms of future investigation, versus those which are intended, and which it would therefore be productive to investigate.

**Step two: Evaluate the explanation.**

Once we understand an explanation, we can begin to evaluate it. There are also three steps to an evaluation.

***1. Is the explanation distinct from the description on which it is based?*** An attempted explanation that can be reduced to the form, "X occurred because X occurred" is trivially true: It provides us with no surplus meaning, and therefore is not an explanation at all.

***2. Amongst the basic implications of the explanation, are there any that are false?*** We must decide whether any of the information taken for granted by the explanation is false. This amounts to evaluating the description on which the explanation is based. If the description is false, the explanation is aimed at explaining something that does not occur, and we need go no further. For instance, if there were no eraser, if in fact, the apparent eraser was an illusion of lights and mirrors, no explanation based on the premise of the existence of the eraser would be useful.

***3. Are the intended surplus implications of the explanation true?*** These are the serious empirical predictions of the explanation that will lead either to its confirmation or disconfirmation by new scientific research. These are what make scientific theories heuristic, what makes them useful. What if, because we held the theory that “the eraser fell,” we were to investigate and discover that the eraser in fact did a dipsy-doodle behind the book? Well, then our theory, though false, would have led to the discovery of a new phenomenon.

**Deficits in Explanations**

Notice that our evaluation may uncover three different kinds of defects in an explanation. A) Explanations can be defective in that they are viciously circular: the descriptions essentially presuppose everything offered in explanation. Such explanations are not wrong, strictly speaking, but they are useless. B) Explanations can be defective in that their descriptions are wrong: The question they ask is based on false premises. — These first two defects we might call “unforgivable” or “unproductive” defects, because explanations that contain them do not really help us to understand the world better or to discover new things about it. — C) Finally, explanations can be defective in that they generate implications that are later, *because of the explanation*, discovered to be false. — We might call this a “forgivable” or “productive” defect of an explanation.

Of course, we are not talking about forgiving or not-forgiving the explanation, but rather the person who has offered the explanation. As scientists, or really as anyone interested in explaining something, our time is wasted by people who offer us circular explanations and by people who offer us explanations for things that do not exist. On the other hand, we cannot fault someone who offers us a plausible explanation that just so happens to be found wrong based on further evidence. In fact, many well respected scientists base their careers on an uncanny ability to offer a series of high quality explanations for a phenomenon, and then to rigorously falsify their own proposed explanations.

**Afterword**

By now you may suspect that we have disappeared into one of those scholarly rabbit holes; the kind which academic authors frequently disappear down, but where wise readers never follow. In fact, nothing could be further from the truth. This classification of explanations, models, surplus implications, etc., accounts for every day explanations just as well as it accounts for scientific ones. Thus, if you think about the explanations you offer up every day for your own behavior or the behavior of others, you should discover that your own explanations fit the structure.

For instance, imagine that while walking down the main corridor of a mall you saw your old friend, Joan, who was shopping. You wave in a friendly way, but she does not wave back. Instead of veering to greet you, she turns away and heads into a store she did not seem to be walking towards originally.

Has she snubbed you? What did you do to deserve such an insult? – In asking the question, you apply a model of a previous occasion in which somebody snubbed you because of something careless you said or did. Common to both situations, the *basic implications* of the model, is the fact that somebody has unexpectedly failed to respond to a friendly overture. In the model situation, the snubber may have been a sibling, and she may have snubbed you because that morning you called her an “immature snot” when she hoarded the bathroom to decorate her face. But these features of the model are irrelevant to the present situation – while they are aspects of the model situation, they are *unintended*. Among the *intended* surplus implications of your explanation of Joan’s behavior is the role of an insult in its causation, and these implications suggest that an apology may be in order. But for the life of you, you cannot remember anything insulting you may have said to Joan. So, when you run into her later you decide to test your theory – i.e., you decide to try to verify some of the intended surplus meaning. Seeing Joan later that day, you step in front of her, and ask, “Are you mad at me or something?” She replies, “Oh, No! Why would you think that?” And after explaining the basis for your theory, she offers her own. She says, “I wasn’t snubbing you. I broke my glasses last night and … you know me… I can’t see more than ten feet ahead of me without my glasses.”

Notice that both your explanation for what happened, and Joan’s explanation, presuppose the same description: that Joan failed to respond to your greeting. But they have different surplus implications. Her theory implies that somewhere there is a pair of broken eyeglasses, and that without glasses Joan can’t see very far. If, a moment later, Joan reads a small sign from across the room, then you might have reason to doubt her purported theory.

1. Conversely, explanations are descriptions that the speaker and audience hold to be unverified under the present circumstances. [↑](#footnote-ref-1)
2. Note that this new question takes for granted the description that there was an eraser in my hand, that there is now an eraser on the floor, and that the two erasers are the same eraser. [↑](#footnote-ref-2)
3. This idea has been developed from the excellent work of Mary Brenda Hesse. Note also that several people have tried to drag us into arguments about whether we are talking about “analogy” or “metaphor.” For the purposes of this book, such arguments are entirely tangential. [↑](#footnote-ref-3)