Philosophy Compass 2/3 (2007): 564–591, 10.1111/j.1747-9991.2007.00081.x

The Psychology of Scientific Explanation

J. D. Trout*
Loyola University Chicago

Abstract

Philosophers agree that scientific explanations aim to produce *understanding*, and that good ones succeed in this aim. But few seriously consider what understanding is, or what the cues are when we have it. If it is a psychological state or process, describing its specific nature is the job of psychological theorizing. This article examines the role of understanding in scientific explanation. It warns that the seductive, phenomenological *sense* of understanding is often, but mistakenly, viewed as a cue of genuine understanding. The article closes with a discussion of several new paths of research that tie the psychology of scientific explanation to cognate notions of learning, testimony, and understanding.

1. Understanding and Scientific Explanation

Theories of scientific explanation can be quite technical, and the content of particular explanations quite arcane, but the cue for acceptable explanation remains cheerfully informal. A good explanation 'feels right': it offers a phenomenologically familiar *sense of understanding*.[1] But clearly this sense of understanding is not always a reliable guide to truth, nor a necessary or sufficient condition for good explanation. As one philosopher of explanation puts it: 'It is no explanation to provide a distorted representation of the world, and the "understanding" induced by such incorrect models is illusory at best' (Humphreys 103).[2] The 'sense of understanding' can mislead us.

Paradoxically, though, this very 'sense' is often invoked as a reason for accepting an explanation; it is often treated as justificatory. According to many thinkers, this type of error is common, influential, and costly. Dawes ('Message from Psychologists to Economists') argues that scientific and everyday explanation is dominated by flabby storytelling, in large part due to the sense of understanding it offers. Rogeberg contends that theories of economic behavior – particularly rational addiction theories – are driven by ad hoc stories that convey a sense of understanding but offer little else. And Wright argues that much of the controversy about the 'hard problem' of consciousness stems from the expectation that an adequate account should supply a sense of understanding about how consciousness could arise from a physical system. Although the occurrence of this sense or feeling of understanding is neither necessary nor sufficient for good explanation, it

drives judgments of the plausibility and, too often, the acceptability, of an explanation.

Characteristically, scientific explanations describe underlying causes that bring about an effect. Explanation has been understood in a number of ways, including as a process of unifying disparate phenomena (Kitcher), identifying the causal or statistical relevance of mechanisms (Hempel; Salmon), or a merely pragmatic virtue offering a narrative designed to account for an effect (van Fraassen). But these explanatory descriptions could be epistemically valuable only if we could, implicitly or explicitly, understand them. As a result, explanation must maintain some contact with our psychological powers of understanding. So some facts about the universe may be irreducibly mysterious to humans, even though they may have true explanations for a possible being with differently organized cognitive powers. We should not expect that all metaphysical differences in the universe are cognitively discriminable by us, any more than they are perceptually discriminable by us.

What factors contribute to making a scientific explanation good? One contributing factor may be its explanatory potential. Explanatory potential requires that the propositions bear the right sort of (typically logical) relation to the explanandum. But that factor alone is obviously too weak a constraint on good explanation; two or more badly false but competing explanations may have explanatory *potential*. So while explanatory potential may be a contributing factor to a good explanation, it is hard to count this as a significant contributing factor if it regularly coincides with false explanations, and may indicate a positive contribution only when the explanation is in fact true.

Of course, whether a factor counts as contributing to a good explanation will depend on what account of explanation you favor. For example, scientific realists suppose that the epistemic reliability of a methodology may provide reason for thinking that a theory, or one of its explanations, is not just accurate at the observable level (and so empirically adequate), but also approximately true. By contrast, an explanation arrived at using a merely instrumentally reliable methodology (one that allows us to make accurate predictions at the observable level) may be a reason for thinking that the explanation is merely empirically adequate, and so 'saves the appearances', but not for thinking the explanation is approximately true.

These features may secure explanations that are true, simple,[3] or unified, but they don't necessarily make the mechanisms transparent to introspection. For that, scientific explanation must convey understanding. But what conception of understanding is important to theories of explanation? Traditionally, understanding is conceived as the result of a process of reconstruction: it happens when we manage to piece together disparate bits of information. Reconstructive accounts agree that understanding is central to explanation, whether as an intellectual goal or as a means of unifying practice.

Peter Achinstein asserts a 'fundamental relationship between explanation and understanding' (16). David Lewis requires of an explanation that 'the recipient understands and believes what he is told' (185). Wesley Salmon proposes that we achieve scientific understanding in two ways: by 'fitting phenomena into a comprehensive scientific world-picture' ('Importance of Scientific Understanding' 77), and by detailing and thereby exposing the 'inner mechanisms' of a process (77). Michael Friedman claims that the relation of phenomena that 'gives understanding of the explained phenomenon' is 'the central problem of scientific explanation' (189). Peter Railton associates the scientific understanding of quantum mechanics with 'the ability to fill out arbitrarily bits of ideal explanatory text' (170–1), which produces a mass of theoretical detail about the object of explanation. Philip Kitcher relates understanding and explanation so closely that elucidating this connection 'should show us *how* scientific explanation advances our understanding' ('Explanatory Unification' 168). James Woodward claims that a theory of explanation should 'identify the structural features of such explanation which function so as to produce understanding in the ordinary user' ('Theory of Singular Causal Explanation' 249; also see *Making Things Happen*).

Accounts of explanation, no matter how different, agree that explanation is intimately related to understanding. The most potent practical feature of explanation, and the understanding it induces, is its guidance of action. We use our achievement of understanding as a signal that we can stop pursuing any further explanation and commence acting on our explanatory belief. Often, a bad explanation will lead us astray; it conveys a sense of understanding that is not a valid cue of accurate understanding. How do we know that our understanding is accurate? In short, we tap into the phenomenology associated with offering the explanation, the sense of understanding conveyed by detailing it.

2. The Traditional Epistemic Analysis of Explanatory Understanding

According to the research findings on judgment and decision-making, the sense of understanding is a common, but routinely unreliable, index of intellectual achievement. To a large extent it can be traced to three causes: hindsight bias, overconfidence bias, and a mistaken attachment to the idea that transparency is routinely achievable. These biases interfere with the truth-tracking role of explanation.

The best-known accounts of explanation are objectivist: they treat the goodness or otherwise of an explanation as wholly dependent on how it relates to external objects and wholly independent of the psychology of the particular explainer. But even the most trenchantly objectivist philosophers of science are tempted by the allure of internal access. Hempel treated explanations as formal arguments, but weakened the formal criteria to address counterexamples. His Deductive-Nomological (D-N) model of explanation

attempts to capture the subjective component in the psychological experience of expectability:

> [G]iven particular circumstances and the laws in question, the occurrence of the phenomenon *was to be expected*; and it is in this sense that the explanation enables us to *understand why* the phenomenon occurred. (337)

Of course, the crucial phrase 'understand why' must be glossed in a nomic way here.[4]

Wesley Salmon also makes understanding central to explanation, even in the context of his treatment of explanation as 'causal-mechanical' and 'ontic'. 'To understand the world and what goes on in it, we must expose its inner workings. To the extent that causal mechanisms operate, they explain how the world works' (Salmon, *Scientific Explanation* 133). Presumably 'understand' here is cast in terms of internal access. That is, what makes the knowledge appropriately expository is that significant parts of its complexity can be represented, and what affords understanding is that *we* can represent it. This latter feature is made possible by subjective access.

In his later work, Salmon became more permissive about the requirements of explanatory understanding, and located the intellectual value of scientific explanations in their power to achieve a number of different goals, all of which

> enhance our understanding of the world. Our understanding is enhanced (1) when we obtain knowledge of the hidden mechanisms, causal or other, that produce the phenomena we seek to explain, (2) when our knowledge of the world is so organized that we can comprehend what we know under a smaller number of assumptions than previously, and (3) when we supply missing bits of descriptive knowledge that answer why-questions and remove us from the particular sorts of intellectual predicaments. Which of these is *the* function of scientific explanation? None *uniquely* qualifies . . . ('Four Decades of Explanation' 134–5)

In his wonderfully accessible 'The Importance of Scientific Understanding', Salmon distinguished between what he calls 'scientific' and 'psychological' senses of understanding. The scientific sense of understanding

> involves the development of a world-picture, including knowledge of the basic mechanisms according to which it operates, that is based on objective evidence – one that we have good reason to suppose actually represents, more or less accurately, the way the world is. (90)

The objectivity of explanation is undermined, however, if the 'good reason to suppose' condition is interpreted in terms of internalist markers.

How then could we further elaborate Salmon's 'global understanding'? Because scientists disagree about what the correct scientific world-picture is, it is not clear what's involved in fitting phenomena into that picture. What Salmon might intend here is the attempted integration of the best-tested and most influential views of the recent history of science: quantum mechanics, atomic theory, the germ theory of disease, Darwinian evolution,

and so on. Taken together, these theories form a truly impressive monument to the scope, detail, and accuracy of modern science; perhaps they constitute, as Salmon puts it, a scientific world picture. But if explanatory scientific understanding requires seeing 'how we can fit them [phenomena] into the general scheme of things, that is, into the *scientific world-picture*' (87), then most people are incapable of explanatory scientific understanding, including most scientists. Indeed, when scientists piece together phenomena, they do so by focusing on the detailed findings of their (usually) narrow specialization. In contemporary science, global unification arises spontaneously from coordinated piecemeal efforts, not from a meta-level at which the philosopher or reflective scientist assembles remote domains (Miller). In fact, in light of the arcaneness of contemporary theoretical knowledge, only the rarest of individuals have the talent, knowledge, and stamina to connect remote domains – say, evolutionary theory and radioactive decay.[5] For technical reasons, they are used together all of the time, but not because scientists are attempting to unify the domains.

Does Salmon's talk of global explanation *entail* this internalist conception of understanding? Of course not. Natural definitions seldom have such entailments. But these accounts of explanation are the work of philosophers, not psychologists. And though philosophers of science tend to offer naturalistic accounts of target interests, philosophers performing standard conceptual analyses will rely on the only 'data' available to casual inspection – the content of one's internal states. Accordingly, we should expect distinctly philosophical accounts of understanding to begin, and perhaps end, with an internalist account of subjective sense.

But this convenient account of justification should not be mistaken for an accurate description of scientific practice. When a scientist (or anyone else, for that matter) ultimately accepts an explanation, it is more likely that the scientist, without any such detailed internal representation of a global (and so diverse) subject matter, simply defers to outside experts or appeals to textbook knowledge.[6] But when philosophers piece together a global explanation, they look to the subjective sense of fit as a cue for an acceptable explanation.

Michael Friedman's account of explanation, too, is tied to the goal of understanding, and it stipulates special criteria for increased understanding:

> I claim that this is the crucial property of scientific theories we are looking for; this is the essence of scientific explanation – science increases our understanding of the world by reducing the total number of independent phenomena that we have to accept as ultimate or given. A world with fewer independent phenomena is, other things equal, more comprehensible than one with more. (195)

The connection that Friedman draws between understanding and a number of independent phenomena concerns cognitive efficiency and tractability. If phenomenon P is reduced to Q, then Q is more basic, and so more likely itself to be irreducible. If what makes a phenomenon irreducible is that it can't be explained in terms of anything else more basic, then it is

inexplicable. A theory (or explanation) that has fewer inexplicable commitments would seem to be rationally preferable to one that has more. As a psychological hypothesis, it may be easier to comprehend a theory that has just one, rather than two or more, irreducible mysteries.

Friedman proposes that the kind of understanding provided by scientific knowledge is global rather than local. Accordingly, accounts of explanation that identify the objects of understanding as individual phenomena can be expected to fail: 'Scientific explanations do not confer intelligibility on individual phenomena by showing them to be somehow natural, necessary, familiar, or inevitable' (197). It is for this reason that scientific understanding reduces the number of ultimate (and thus inexplicable) commitments.[7]

Philip Kitcher advances a novel and detailed version of the unification approach, adding crucial categories of learning to this account. His approach is crafted to honor the work-a-day details of scientific practice. On Kitcher's account, understanding does not emerge spontaneously from the mere reduction of independent theoretical commitments; comprehensibility depends upon the possession of background knowledge and the training required for the assimilation of superficially different phenomena to common categories:

> Understanding the phenomena is not simply a matter of reducing the 'fundamental incomprehensibilities' but of seeing connections, common patterns, in what initially appeared to be different situations. Here the switch in conception from premise-conclusion pairs to derivations proves vital. *Science advances understanding of nature by showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again, and, in demonstrating this, it teaches us how to reduce the number of types of facts we have to accept as ultimate (or brute).* [fn deleted] ('Explanatory Unification and Causal Structure of the World' 432)

Kitcher captures this objective, psychological and pedagogical feature of explanation by asserting that knowledge of a theory

> involves the internalization of the argument patterns associated with it, and that, in consequence, an adequate philosophical reconstruction of a scientific theory requires us to identify a set of argument patterns as one component of the theory. This is especially obvious when the theory under reconstruction is not associated with any 'grand equations' and when reconstructions of it along traditional lines produce a trivialization that is remote from the practice of science. (438)

The ability to identify argument patterns is an effect of scientific training. Unlike other descriptions of unification, for Kitcher the role of explanation issues from dirtying our hands with local details of scientific practice, rather than from philosophical rumination or the scientist's supposed aesthetic preference for grand coherence. Explanation plays an epistemic role in science not chiefly because global scientific worldviews set an explicit, top-down research agenda, but because working scientists must piece together local findings within the very small handful of models and theories they use and evaluate.

Once reconstituted as an openly causal account, this treatment of explanation has sufficient detail and descriptive accuracy to accommodate the history of science and current practice. Any such argument patterns – if they are to represent the esoteric practice of science – must be pedagogically valuable and highly enthymematic. Also, there is no reason that the internalization of argument patterns cannot be understood in a more straightforward metaphysical way, in terms of the causal features of internal representations of the explainer and the relevant parts of the world being represented.

These declarations associate explanation with understanding without saying what understanding is. Certainly none of them entail that a sense of understanding is a key criterion for a good explanation. Nor do any of them account for understanding in terms of a potential explanation. But most of them do draw upon our ordinary, ungarnished notion of understanding, in terms of 'detailing' and 'identifying' the inner mechanisms and structural features of interest. Terms like 'detailing', 'identifying', and 'fitting' imply that explaining a phenomenon is a meta-cognitive process, not an implicit or otherwise inarticulate process or skill. Accordingly, explanation is portrayed as a kind of detective work, in which the investigator consciously pieces together otherwise disparate facts into a coherent picture. But the explanatory concerns of working scientists are local rather than global: they are more likely to focus on the details of their experiments, and niceties of their methodologies, than the larger issues, such as a theory's fit with its neighbors. It is philosophers who are more prone to address the overarching issues, such as nomic subsumption, establishing fit, and describing mechanisms. It is also philosophers rather than scientists who are interested in the *concept* of explanation. Ronald Giere's characterization of philosophical work on explanation nicely captures the current situation: '[M]ost philosophical writing on 'scientific explanation' is not really about explanation *within* science, but about the use of scientific knowledge in the explanation of events in everyday life' (104–5).[8]

Current theories of explanation may leave us embarrassed, but we should be at least as embarrassed by our 'feels-right' diagnostic standard for the acceptance of an explanation. Sadly, humans probably have more counterfeit than genuine experiences of 'understanding'. Cognitive research has documented an overconfidence effect in scientists and laypeople alike (Fischhoff, Slovic, and Lichtenstein 1977): the mere 'feeling of knowing' (see Koriat, 'How Do We Know that We Know?'; 'Dissociating Knowing') is highly unreliable evidence of knowing. This overconfidence is not the result of individual differences in personality or of clinical delusions of grandeur; it is the normal consequence of a normal individual's routine thinking.

3. The Phenomenology of Scientific Explanation

If philosophers have examined the normative function of explanation, it is scientists who have ably captured the nature of the sense of understanding.

Peirce identifies the distinctive cognitive experience of explanatory understanding by isolating the moment of final acceptance; the good explanation 'is turned back and forth like a key in a lock' ('Neglected Argument' 100). This description alone should supply little solace to those holding that 'understanding' an explanation is epistemically reliable. After all, alchemists surely felt the key turn, but once inside we find only false descriptions of causal mechanisms. Yet some writers (de Regt; de Regt and Dieks) insist that rejecting the 'aha!' sense of understanding amounts to abandoning the very idea of understanding, and perhaps explanation, altogether. Others argue that understanding consists in the solution to certain 'insight problems' (Hardcastle), such as the Dunker-candle problem.[9] Such analyses of understanding should not prejudge the nature of understanding appropriate to explanation. Rather, our analyses should reflect that the required notion of understanding falls out of the actual practice of explanation in science.

Some of our favorite stories in the history of science, such as Kekule's famous 'Eureka' episode, feature dramatic journeys of insight to compelling explanations.[10] In these scenarios, the explanation 'felt right'. I will argue that the psychological sense of understanding is just a kind of confidence, abetted by hindsight, of intellectual satisfaction that a question has been adequately answered.[11] Thus the sense of satisfaction can be described simply as a heightened confidence that one enjoys an accurate description of the underlying causal factors. Confidence is, notoriously, not an indicator of truth.

The track record of this subjective sense of understanding is not entirely treacherous. This sense sometimes seems to be associated with very secure explanations as well. Darwin claimed that the explanation of the distribution and origin of species in terms of natural section was 'so satisfactory' (476). Avogadro drew a similar inference to the best explanation for the existence of molecules. Diverse domains and measurement procedures led to the unification of molecular and atomic phenomena. Perrin commented on the confidence with which the molecular hypothesis could be held, and did so in distinctly psychological terms:

> Our wonder is aroused at the very remarkable agreement found between values derived from the consideration of such widely different phenomena. Seeing that not only is the same magnitude obtained by each method when the conditions under which it is applied are varied as much as possible, but that the numbers thus established also agree among themselves, without discrepancy, for all methods employed, the real existence of the molecule is given a probability bordering on certainty. (215–16)

Perrin describes the potency of Avogadro's hypothesis in psychological terms. The hypothesis 'arouses wonder'; the agreement is 'remarkable'. This subjective sense of understanding is a psychological impression that the explanatory mechanisms are transparent and coherent, or that the explanation seems plausible, and so should be confidently accepted. This is often the structure of inference to the best explanation, but it needn't be. IBE need not rely on a sense of understanding or a feeling of coherence when

identifying the best alternative. At the moment, there is neither a satisfying formal account of explanation nor agreement about the important informal criteria for good explanation, producing what one review casts as 'an embarrassment for the philosophy of science' (Newton-Smith 132).

Sometimes explanations stop when we have reached the limits of our understanding. What makes explanations stop? Most explanations stop well before they have reached irreducible mechanisms, even when more basic mechanisms could be cited. Indeed, many explanations are not even reductive in focus or purpose.

Frank Keil has emphasized that explanations have indefinite depth. But 'indefinite' does not mean 'arbitrary'. We are not free to say when to stop explaining a phenomenon. Garfinkel argues that explanations are like water: they seek their own level. In this case, it means that explanations stop when the causal depth of the mechanisms reach the level of processing required to answer the original explanatory question. One strain of lore on explanation pursues explanations that are 'rock bottom' (Watkins) – that spell out all of the metaphysical constituents of the processes covered in the explanation. This claim consolidates a widely held conviction that anything less than a complete reduction is unsatisfying; it allows too mysterious a connection between higher- and lower-level processes. But, as many have pointed out (Fodor), such reductive explanations come at a cost: you miss important generalizations that are taxonomic. Explanations are, after all, expressed in language, and so are prone to linguistic effects. By mentioning micro-causes, hyperconcrete specifications (Garfinkel) of causes evoke Gricean implicatures that their reductive status is explanatorily relevant: If you mention a feature, it must be relevant to the phenomenon you are explaining. It is in this way that hyperconcrete explanations misrepresent the sensitivity of the system to changes.

Not all accounts of the sense of understanding venture evolutionary origins. Some simply describe deep causes or unify disparate domains. Consider our explanatory beliefs. Preston and Epley show that we value beliefs much more that we apply widely than beliefs that are themselves explained by a variety of facts. This suggests a number of virtues people attribute to our most cherished explanations; they should be robust, explaining a wide range of otherwise inexplicable phenomena. They should be explained in virtue of nothing else. It is this virtue of basic-ness that gravity is so powerful an explanatory concept for Newton, and an 'unmoved mover' is for so many religious thinkers.

Although these may be psychological facts about humans that make certain kinds of beliefs more valued that others, they are not *merely* psychological facts; they indicate something concerning the nonpsychological subject matter they are about. Deep causal insights are related to illuminating unification. Successful unification of disparate findings tells you something about the structure of the world. And the identification of deep rather than shallow causes is preferred for a reason: Deep causes generate the very kinds of evidence that is unifying.

Psychological theories of explanation, too, appeal to the important role of a sense of understanding, in both everyday and scientific explanation. What makes an explanation 'feel right'? What is the right combination of coherence, sense-making, anxiety relief, closure, and completeness? Like some global, unifying accounts of explanation in the philosophy of science, a prominent psychological account focuses on the unified conceptual framework it provides: '[I]n everyday use an explanation is an account that provides a conceptual framework for a phenomenon (e.g., fact, law, theory) that leads to a feeling of understanding in the reader-hearer' (Brewer, Chinn, and Samarapungavan 120). And scientific explanations are no different in this respect; they should 'provide a feeling of understanding' (121), in addition to whatever objective virtues the explanation might possess. The operative assumption may be that this sense of understanding is an epistemic virtue; the fact that an explanation conveys this sense of understanding offers a reason for thinking it is also a true, or at least a good, explanation. However, in order to be an epistemic virtue, it should be positively correlated with accurate causal descriptions, or the goodness of an explanation. But it isn't. Instead, this sense is caused mainly by nonepistemic forces, such as the demonstrated psychological tendency of overconfidence, or the desire to reduce, as William James put it, 'that peculiar feeling of inward unrest known as indecision' ('Will' ch. 26, 1136).[12] Formulating a unified, consistent story is one way to eradicate that peculiar feeling of inward unrest. An explanation pleasantly discharges that feeling of intellectual unease.[13] But unity and consistency with background knowledge is a poor substitute for accuracy and truth, as the litany of false but unifying and consistent theories in the history of science should warn us. A good story, and so a good explanation, is coherent. In fact, one theory of explanation, 'the story model', is based on the nature of stories (see Hastie and Pennington). Hastie and Pennington performed jury research in order to determine the potent causes of juror decisions. It appears that the model that underlies juror decision-making is not Bayes' Theorem but the narrative skill of the storyteller. In their view, evidence is convincing to a juror not because it shifts his or her conditional probability of guilt but because it fits into a convincing story of how the crime happened. Coherence appears to consist of three features: completeness, plausibility, and consistency. One might suppose that a plausibility requirement addresses the worry about false theories. However, judgments of plausibility are themselves theory-dependent, and so if your theory is badly false you cannot expect your explanations to fare much better.[14]

4. Where Does the Sense of Understanding Come From? The Role of Overconfidence and Hindsight Biases

The sense of understanding may be unreliable yet weakly correlated with genuine understanding. In this case, it might be a decent cue for acceptance even if not optimal. After all, there is no solid evidence to indicate that the

sense of understanding and genuine understanding are literally uncorrelated. An explanation may be related to genuine understanding in a number of ways. It may contribute to our ability to account for a wider range of phenomena. Generating understanding could be constitutive of being an explanation, in which case it is at least a necessary condition for being an explanation. Generating understanding could just be a symptom of being an explanation. Moreover, the relation between explanation and genuine understanding may aid in comparing competing explanation. In this case, generating *greater* understanding could be constitutive of being a *better* explanation. At the same time, generating greater understanding could merely be a symptom of being a better explanation, so that greater understanding would be a cue of a better explanation. But the validity of this cue must be provisional. It is difficult to assess the positions surveyed in section 2 in terms of the relation between explanation and genuine understanding, because those philosophical accounts tend to be orthogonal to that psychological task.

But there are two sources of evidence, some of it historical, that indicate there is little practical wisdom in relying on the sense of understanding as a cue of a good explanation, or even a potential one. To the extent that the subjective sense of understanding is not produced by a reliable relation to the truth, where does the subjective sense of understanding come from? At least in a significant number of cases, the 'understanding-inducing' nature of explanations comes from two well-documented psychological biases – hindsight and overconfidence. These two biases that are difficult to correct, and they survive different experimental methods, test items, and classes of people.[15]

OVERCONFIDENCE BIAS

We are systematically prone to believing that we are right when we are not. The literature demonstrating overconfidence is large, and counter-examples to the effect have been difficult to produce. Fischhoff, Slovic, and Lichtenstein asked subjects to (a) indicate the most frequent cause of death in the U.S., and (b) estimate their confidence that their choice was correct (in terms of 'odds'). When subjects rated the odds of their answer's correctness at 100 : 1, they were correct only 73% of the time. Remarkably, even when they were so certain they were right that they rated the odds between 10,000 : 1 and 1,000,000 : 1, they were correct only between 85% and 90% of the time.

The overconfidence effect is systematic and directional. It is highly replicable, and survives changes in task and setting. And the effect is in the direction of over- rather than underconfidence. Given the systematic, stubborn nature of the overconfidence effect, we might expect that expert training is not the key to deliverance, and indeed physicists, economists, and demographers all suffer from this bias, even when reasoning about the

content of their special discipline (Henrion and Fischoff). It is little consolation, then, that good explanation has been associated with the 'feeling' or 'sense' of understanding. The same sense of understanding, of intellectual conviction, accompanied the subjects' incorrect answers. And there is no reason to think that answers to *explanatory* why-questions are different in a way that allows them to evade that fact.

Overconfidence, then, may be a truly disastrous component in explanatory reasoning. Our judgments of accuracy are systematically correct only when we have a good theory or model of the process we are making judgments about. The tendency toward overconfidence would not be so damaging if our judgments were 'calibrated', a term used to designate the extent to which our confidence matches our accuracy. But in science we seldom have such accurate information available to us.[16]

Why do people find it so difficult to learn from the lessons of overconfidence? First, it is difficult to learn from individual cases when we typically don't get (or don't retain) systematic feedback about the quality of our judgments. To mention just one restriction on feedback, we can't compare the long-term outcomes of our actual decisions against alternative decisions we didn't implement. Second, the overconfidence is general and persistent, and this attachment to subjective evaluation of complex incoming evidence has specific consequences for learning. Perhaps this is best seen in the 'interview effect', an especially striking example of our overconfidence in our powers of subjective evaluation. When 'experts' such as hiring and admissions officers, are able to review applicants in unstructured interviews, they are outperformed by statistical prediction rules that take no account of the interviews. In fact, unstructured interviews actually degrade the reliability of human prediction (Bloom and Brundage; DeVaul et al.; Oskamp; Milstein et al.). That is, people degrade the reliability of their predictions by availing themselves of unstructured interviews; the interview information used is irrelevant to (and so dilutes) accurate prediction about future performance. Although the interview effect is one of the most robust findings in psychology, highly educated people ignore its obvious practical implications.

Our conceits are abetted by a classic and systematic frailty of interpreting probabilistic information. In the face of a half-century of experiments showing that human judgment is inferior to well-tested statistical prediction rules in such contexts, many disciplinary experts and others still base judgments on subjective impressions and unmonitored evaluation of the evidence (Bishop and Trout, *Epistemology and the Psychology of Human Judgment*). Resistance to these findings runs very deep, and typically comes in the form of a self-serving bias we might call Peirce's Problem.[17] Peirce ('Doctrine of Chances' 281–2) raised what is now the classic worry about frequentist interpretations of probability: How can a probability claim (say, the claim that 99 out of 100 cards are red) be relevant to a judgment about a particular case (whether the next card will be red)? After all, the next card will be red or not, and the other 99 cards can't change that fact. Those who

resist the findings of inferior performance are typically quite willing to admit that *in the long run*, a simple statistical prediction rule will be right more often than human experts. Yet their (over)confidence in subjective powers of reflection leads them to deny that we should believe we are unreliable *in some particular case*. Whatever the long-run performance of humans, so the sentiment goes, in this case I am the right person to make the subjective judgment. And the effect of this lack of discipline is made worse because it is difficult to secure all of the information necessary for clear feedback: We can't compare the long-term outcomes of our actual decisions against the decisions we would have made if we hadn't interviewed the candidates but simply used a rule.

On the basis of this evidence, humans appear naturally disposed to exaggerate the powers of our subjective faculties. No matter what this disposition's origin, it won't do to reply to my claims of unreliability that this result seems implausible, for that is exactly the point. For example, the theory predicts that people will find the sad lesson of the interview effect implausible. Despite a moral that should be humbling, we continue to look for, and when it is present we use, that special, subjective sense of illumination as a cue of correctness. By now it should be clear that this cue, by itself, is unreliable, and often isn't present when we have explanatory knowledge. So clearly if explanation imparts understanding, understanding must consist in something other than a subjective sense of meta-cognitive control.

HINDSIGHT BIAS

The overconfidence bias is not the only supporter of explanatory indolence,[18] for the *hindsight bias* feeds the lazy beast as well: Hindsight bias is our tendency to find past events more predictable than they in fact are, and thus induce a sense of certainty in the explainer and a sense of inevitability about the event. Traditionally, the phenomenon of hindsight bias is established by asking subjects to estimate the likelihood of various outcomes of an upcoming event, and then retesting them after the event, asking them to recall how likely they had found each of the possible outcomes the first time around.

Prior to President Nixon's trip to China and the Soviet Union in 1972, Fischhoff and Beyth asked subjects how likely they found a variety of possible outcomes (for example, whether Nixon would meet Mao, that the Soviet Union and US would establish a joint space program, etc.). Two weeks to six months after the trip, the subjects filled out the same questionnaire. In addition, they were asked to recall the probabilities they had initially assigned to the same events or, if they couldn't recall, to say what probability they would have assigned immediately before Nixon's trip. They were also asked whether each of the listed outcomes had, in fact, occurred.

The results were a striking demonstration of the distorting influence of hindsight. For those outcomes that subjects thought had occurred, they

remembered their estimates as more accurate than they in fact were. For those outcomes thought not to have occurred, they recalled their estimates as having been lower than they in fact were. The effect seems to strengthen with the passage of time. After three to six months, 84 percent of the subjects displayed hindsight biases. In a nutshell: Fischhoff and Beyth showed that after we know outcomes about which we earlier made predictions, *we believe the actual outcomes were more predictable that they actually were*.

Learning the outcomes of scientific theory testing – the preamble to any explanation – places scientists in a position similar to the hindsight subjects. Explanation is retrospective. When constructing an explanation, we draw one line through many events, a line that could have been drawn in different ways, even if we know the point of termination. As we draw this line, we are not very accurate judges of how much we are affected by information about an outcome, be it a mass extinction, an explosion, or infection. But once the line is drawn, we conceptualize the event as inevitable, and thus people tend to say that the event was fairly predictable all along. Thus, the hindsight bias is also known as the 'I-knew-it-all-along effect'. In particular, people tend to overestimate how probable they thought the event was before it occurred.[19]

A familiar situation: you tell a colleague over lunch about an experimental finding in human behavior, and getting the response (dressed up with a knowing chuckle) 'You didn't need to run an experiment to know that'. Your colleague, insensitive to the effect that the reported outcome has just had on him or her, claims to have 'known it all along'. But this 'knowledge' is probably illusory. In most cases, experiments get run because the underlying psychological processes and behaviors are complex: you don't gain knowledge of them by mere reflection or casual observation. Yet after the fact it *seems* that you could have.[20]

The hindsight bias lies behind our tendency to confuse predictability with statistical contingency. We think that prediction lifts the epistemic burden from explanation. And while predictive accuracy can be an important index of scientific integrity, prediction is both epistemically over-rated and often difficult to secure. For instance, knowing the factors contributing to an effect in a complex system does not necessarily allow you to predict the outcome. Plane crashes are especially good examples of the limited fare supplied by prediction. Upon analysis of cockpit transcripts, a handful of important factors are identified that contributed to the crash. But as Robyn Dawes puts the point:

> [T]hese factors would not allow us to predict future crashes very well at all. Airplane crews are often fatigued; bad weather occurs frequently; miscommunication is not that unusual, nor are temporary breakdowns of radio communication or panic at the last minute. (37)

Dawes's point clarifies the doubly-distracting nature of the hindsight bias. After the fact, it causes us to acquiesce in the belief that we have an understanding of an effect (because 'we knew how it would turn out all

along'). After that, we regard our search as complete. And, the hindsight bias overestimates the instrumental importance of predictability. In the plane-crash case Dawes considers, we have all of the predictive details we might want, but none of them explains the crash very well. These effects of fatigue, wind, and miscommunication are too high-frequency to satisfy the condition that a good prediction should have a low-probability of being true on any other theory. This unjustified conceit is what makes the hindsight bias so damaging to the search for accurate explanations.

The history of science is a rich source of examples. In a classic moment of hindsight and overconfidence, Ptolemy claimed that

> it is idle to seek for causes for the motion of objects towards the centre, once it has been so clearly established from the actual phenomena that the earth occupies the middle place in the universe, and that all heavy objects are carried towards the earth. (43/H22)

Here hindsight bias acts as a stopping rule: Ptolemy's gloss is that further inquiry is idle, a commitment that assumes an earthly position different from the one we now believe it to have. Ptolemy's conservative appeal to coherence with (in this case, false) background beliefs, together with his dismissive treatment of alternatives, displays his influence by hindsight.

Our history is littered with inaccurate explanations we confidently thought were true: the explanation for mental illness in terms of demonic possession, the humoral theory of illness, and so on. The sense of understanding would be epistemically idle phenomenology were it not so poisonous a combination of seduction and unreliability. It actually does harm, sometimes making us squeamish about accepting true claims that we don't personally understand, and more often operating in the opposite direction, causing us to overconfidently accept false claims because they have a kind of anecdotal or theoretical charm. And it is this brew that produces the sense of understanding.

5. The Psychology of Explanatory Storytelling

Perhaps our explanatory errors are simply an artifact of our cognitive limitations. Stories impose the coherent structure and the control over detail so prized in explanation. Whether allegory, parable, or gossip, people find a good story irresistible. If the story has vivid detail, causal connections to issues that matter to us, and the promise of resolving mysteries, the explanation is difficult for us to resist. People bought the explanation of the start of WWI – that Gavrilo Princip shot the Archduke Ferdinand. And, it was a common 19th-century story that erratic behavior results from demonic possession. Consider a third example, the sense of understanding induced by evolutionary explanations of human behavior. According to this story, in his struggle to have nature select his genes, a man will attempt to increase the representation of their characteristics in the gene pool. For this purpose, an enhanced disposition toward violence can be advantageous, particularly

in the Pleistocene environment in which these dispositions evolved. So men are quick to kill sexual rivals, denying them access to that pool. They dominate women who threaten that advantage, killing the vessels that carry rivals' offspring. Indeed, we are told that perhaps humans have special-purpose modules that make this behavior automatic.

Evolutionary stories are stories nonetheless. Experiments show that people don't understand the process of natural selection very well (Lombrozo, 'Why Adaptationist Explanations are so Seductive'; Shtulman). Most notably, they don't understand population-level properties of natural selection. At the population level, natural selection involves differential reproduction and variation. We can easily imagine – even picture vividly – an individual's search for prey, or their ultimate triumph in the struggle to mate. But it is much harder to imagine the actual dispersion of individuals' traits within and between species.

Consider rampant adaptationism. Adaptationism is the view that that most of our traits are optimally adaptive. We commonly observe this form of evolutionary explanation for the presence of a trait: Males are relatively aggressive, and so aggression must be adaptive. And now, humans have a sense of understanding, and so it must be adaptive to have a sense of understanding. Why do people favor adaptationist explanations so much, given that they don't seem to have the causal knowledge that the explanations tacitly assume? One possibility lies in the differential persuasive power of some merely potential explanations.

The adaptationist explanation is given a chance – or is seen as potentially warranted – if we can provide a coherent sketch of the processes that underlie the event we want to explain. This is so even if we don't really understand how the process works. For example, people may treat an adaptationist explanation as potentially warranted if they know of a biological process able to produce adaptation (such as natural selection). If they know this 'process fact', then, together with knowledge of the causal fact that the function caused the adaptation to occur, people will be moved by the adaptationist explanation (Lombrozo, 'Why Adaptationist Explanations are so Seductive'). These explanations can still be poor, but they are often accepted as true because they have a variety of psychologically attractive features (e.g., they may be familiar, computationally tractable, etc.), however orthogonal these features may be to the truth.

I have chosen the case of evolutionary psychology for its powerful irony. The very mechanisms that make the explanations of evolutionary social psychology so seductive might also be invoked to explain the evolutionary basis of the sense of understanding. These attractions are not news to psychologists. Deborah Keleman explains the overall appeal of adaptationist explanations in terms of the specific attraction of intentional explanations, which provoke 'promiscuous teleology':

> Promiscuous Teleology (PT) makes two claims: First, it argues that the tendency to view objects as designed for purposes developed as part of our ability to view

> intentional agents as having purposes. Second, PT suggests that because of the way our minds are designed, intention-based teleological explanations come easily to us. One consequence of this is that we might overzealously apply teleological reasoning to inappropriate domains unless we have learned not to. Before acquiring alternative explanations then, children (and scientifically naïve adults) might, by default, construe almost any sort of living and non-living entity as intentionally caused for a purpose. (287)

If we can expose the undue enthusiasm for evolutionary accounts of the sense of understanding, then perhaps we can begin to assess the validity of the sense of understanding on its own merits. After all, it might be adaptive for us to want explanations, and even to venture them. Some psychologists have even made this argument, and have suggested that we assess in evolutionary terms the individual drive to wonder, and to attempt explanation.

Alison Gopnik, for example, treats the gratifying sense of explanation as the impetus behind search and discovery. The phenomenology of good and bad explanations may be indistinguishable, but it keeps us looking. According to Gopnik, this sense of understanding has an evolutionary function similar in effect to orgasm; there may be many motivations to reproduce, but with the sensation of orgasm, who needs further motivation? As it turns out, though, we are not very good at tracking the ratio of good explanatory feelings to accurate outcomes. So, as a cognitive matter, using the sense of understanding as a cue to good explanation is a poor strategy for finding good explanations. On the other hand, as a matter of motivation, the drive to understand may be the only impetus to keep searching, and so is crucial to ultimately finding good explanations. There may be a characteristic feeling that accompanies an explanation, but that is not the aim of explanation.

Not surprisingly, as a piece of evolutionary science, the reproductive story of the orgasm is controversial (see Lloyd). The evolutionary analogy to the sense of understanding, now once removed, is so much adaptationist storytelling, without any independent evidence supporting it. The idea seems to be that the sense of understanding has survived, and so having it must be adaptive. And it wouldn't be adaptive, presumably, unless it tracked the truth. Yet the appendix has hung around for a long time. It is easy to find bad adaptationist arguments in the popular literature, but in philosophy those have already been canvassed. In fact, some psychologists study empirically why good people believe bad adaptationist arguments.[21] We will turn to this topic in a moment.

Perhaps we can have it both ways: The phenomenology of explanation motivates us to learn about the environment, but we are satisfied with shallow explanatory mechanisms (Rosenblit and Keil). Rather than describing how causal mechanisms work, we assert when causal claims are *potentially* warranted. Most people's understanding of evolutionary theory is quite shallow. The intellectual satisfaction that people get from an evolutionary social explanation does not depend on – and so doesn't increase from –

knowing anything else about evolution, like the role of variation or differential reproduction (Lombrozo, 'Why Adaptationist Explanations are so Seductive'). And although shallower explanations may be preferable when having more reductive detail would only be misleading[22] (functional explanations are a case in point; see Miller), they may also convey that little more than casual familiarity with a subject affords a deep grasp of a topic.

If we can get by with superficial knowledge, what makes some particular account a just-so story? As Robyn Dawes puts it, the general problem 'is that there is a many-many relationship between antecedents and consequences in the course of human life. As we retrospect, in contrast, we can create many-one relationships' ('Message from Psychologists to Economists' 37). Explanation creates many-one relationships. Dawes's claim suggests an interesting account of people's acceptance of evolutionary theory despite incomplete understanding of it: If people evaluate explanations while neglecting relevant causal mechanisms, they won't be in a position to invoke causal mechanisms when evaluating confirming or falsifying evidence. As a result, they will be unable to track the truth of such explanations reliably when the underlying theory is inadequately developed. This is precisely what happens when lay adaptationists aren't in a position to evaluate the relevance of evidence from variation. Here what is in question is not the core claims of evolutionary theory, but the speculative assertions of evolutionary social psychology.

6. Understanding and Learning

Sometimes understanding comes packaged with the 'Aha' experience, and we have already raised the question of whether that experience is a reliable cue of genuine understanding. But sometimes understanding comes without the phenomenological fanfare. In both cases, understanding is the consequence of a learning process. There are, in fact, many modes of learning. We can learn by feedback with explicit instruction or not (see, for example, Pearl; Spirtes, Glymour, and Scheines). We can learn by analogy, representing to ourselves or 'internalizing' a model of a process, and then generating scenarios about the performance of the system under a range of circumstances. The idea of a model is thus a kind of metaphor by which one can transfer understanding to a new subject.[23]

If understanding is a consequence of learning, then understanding can be as implicit or explicit as learning itself. We constantly learn without awareness of learning, and cognitive psychology has produced a spate of research establishing the importance of implicit learning in a variety of cognitive domains. Psychologists, too, advance and array of learning procedures that rest on understanding, but only some of them involve awareness. The others are largely implicit. When we learn by analogy, the dimensions of similarity we are responsive to may be unconscious. We learn skills through imitation and repetition, we often categorize objects by their unattended dimensions,

we learn synonyms by associative learning (by pairing semantic affiliates), and we use cognitive maps or mental models when learning complex cognitive tasks. These kinds of learning depend on a pertinent kind of understanding, but not one that need involve an 'Aha' experience.

Consider the research on implicit learning of particular voices. You begin by exposing individuals over several days to a series of words uttered by a number of people. You then ask them to distinguish between words, half of which were, and half of which were not, among the original test set. In this case, reaction times are faster, and error rates lower, for words that were uttered by the same people as those in the original test set. Thus, the same words, uttered by voices different from those in the original test set, required more perceptual and cognitive effort to identify. The redundancy gains achieved by memory for concrete vocal detail were the product of implicit learning. Throughout the training sessions, the participants were not aware that they were learning vocal detail of the particular word-tokens; they thought they were simply learning word-types, as the instructions asked them to do (see Nygaard, Sommers, and Pisoni; Goldinger; Church and Schacter). In effect, subjects developed a tacit model of each talker's vocal characteristics, and their cognitive mechanisms engaged that model in matching the input to a prior lexical entry. Because their model was accurate, they enjoyed increased speed and accuracy. Implicit learning has been demonstrated in a variety of perceptual and cognitive tasks, such as lexical decision, picture naming, object decision, word association, category instance generation, and answering general knowledge questions (Baddeley). The lesson here is the same as that for the implicit learning of particular voices: neither awareness of the dimensions attended to, nor 'a sense of understanding', is required for learning. This sensitivity to relevant dimensions – a recognitional capacity – is certainly one form of understanding. And, even if it is true that humans are unusually good at internalizing some classes of statistical contingencies, these contingencies can be acquired without any accompanying 'sense of understanding'.

If we are trying to sort a collection of instances into natural classes or 'kinds' – a taxonomic task required by any science – we must rely on feedback about negative instances in order to improve our discrimination. We must see some non-gazelles in order to sort a group of antelope-ish objects into gazelles and oryxes. But there is one major exception to this generalization, and it provides the basis of the pedagogical account of explanation I favor. We can learn without exposure to negative instances when we have a theoretical model. To quote a famous study on medical diagnostic classification, a model can give us

> access to the major causal influences, possesses accurate measuring instruments to assess them, and uses a well-corroborated theory to make the transition from the theory to fact (that is, when the expert has access to a specific model). (Dawes, Faust, and Meehl 1670)

In short, for a combination of accuracy and efficiency, there is no explanatory substitute for an accurate model, or a good theory.

7. *The Role of Psychology in Good Explanation: Small Steps toward Understanding*

Explanation evidently involves understanding, and the next few decades may see the incremental description of the many cognitive components of understanding. Naturalistic philosophers of science use this psychological knowledge to articulate the nature of explanation. Others believe that the concept of scientific explanation must be tied to an ordinary concept of explanation, and of understanding, using concepts that may have no standing in a scientific psychology. If understanding is a kind of recognitional capacity, or some other outcome of a learning process, then our uncanny ability to learn contingencies without supervision, without awareness of insight or the phenomenological sense of understanding, stands as a clear counter-example to the classic 'Aha' view of understanding. After all, a true explanation teaches us something, or, if you like, we learn from it. But no matter what the correct account of understanding, there is no received consensus, and many plausible alternatives to the classic 'Aha' view.

As we have seen in efforts to explain the attractions of an evolutionary account of social behavior, one approach to developing a psychological account of scientific explanation examines actual cases of scientific explanation. In most natural systems, there are many variables, each contributing only a small amount to the explanation of an outcome. And because there are so many contributing variables, we couldn't possibly hold them all in our minds to register the light. Consider the explanation for the process of speciation, either in a particular population or in natural populations generally. Normally, the number of factors contributing to genetic drift, and ultimately reproductive isolation, is enormous: original location of population, geography, climatic changes, food, seasonal changes, floods, geological events, number of offspring, etc. This is true of Ensatina salamanders in the Central Valley of California, and the cichlids of Lake Victoria and Lake Tanganyika. On any causal model of this speciation process, each cause may contribute some percentage to the speciation, and so to the explanation of it. It is hard to see how you get a sense of understanding from so diffuse and numerous an array of factors, when there are so many you would have to write them down; memory could not hold them all at once.

And there are many other routine cases in which the variables to produce the flash of insight are just not there – and shouldn't be, given the modest nature of each cause's potency: explanations for health and disease, for war and peace, the success of educational reforms, the causes of property crimes, and the behavior of complex physical systems like plasmas and fluid, as well as for the many proprietary components of disciplinary domains like meteorology and statistical mechanics.

Philosophers and social scientists have also begun to examine the structure and function of deferred knowledge or testimony in scientific practices like explanation. And this, too, is an obstacle to the 'Aha' view of explanation and understanding. Scientists in domains remote from one another often appeal to the arcane findings of another discipline when accounting for their own data. As initiates to the remote field, they may not understand the theoretical details of these arcane results. This relation of epistemic dependence (Hardwig, 'Epistemic Dependence'; 'Role of Trust in Knowledge') governs the most familiar and routine cases of the 'division of cognitive labor' (Kitcher, 'Division of Cognitive Labor'; 'How Do We Know that We Know?'), and 'mercenary reliance' (Trout, 'Theory Conjunction'; *Measuring the Intentional World*), in scientific investigation. Yet, the explanations that implicate causal knowledge that is arcane to a researcher is unlikely to produce the 'Aha' experience so prized by those treating it as a necessary condition of good scientific explanation. In fact, it is a rare individual who can traffic in both paleontology and magnetism, or psycholinguistics and the engineering implementations of signal processing. The scientific explanations of results in, respectively, paleomagnetic dating of rock strata and the measurement of lexical access, integrate these remote fields, and yet the gaps in relevant expertise will typically defy the relevant 'Aha' requirement on understanding.

It is always possible to reply that these narratives may be intellectual endeavors of some sort, but without resources of familiarity to inspire an 'Aha' experience, they are not cases of understanding. But this reply seems suspiciously ad hoc. They certainly behave like other explanations, and they are treated as such by scientists. Even if there is a difference between the 'Aha' phenomenon and a more general sense of understanding that accompanies other cognitive phenomena like expertise, it would seem inappropriate to legislate from the armchair what counts as an explanation. Perhaps a suitably hedged description of conditions for genuine explanation will be useful. Genuine understanding exists, not when the sense of understanding dawns, but when the following objective conditions are met:

1. The explanandum-statement putatively understood is at least approximately true,
2. The agent has sufficient collateral theoretical knowledge or information (explanatorily) relevant to that explanandum, and
3. The explanandum-belief is produced by a reliable process, whether perceptual, cognitive, or social.

When present, these three conditions may or may not produce a characteristic feeling of intellectual satisfaction. Fortunately, genuine understanding does not require the 'Aha' experience, or any other distinctive, subjective feeling. What it requires is an appropriate causal relation to the evidence.

The resulting objectivist, ontic, account, in generic form, states that scientific understanding is the state produced, and only produced, by grasping

a true explanation.[24] Much depends on how 'grasping' gets worked out. One might, for example, treat grasping as a kind of knowing. In that case, if you can know without articulating your justification, then you can grasp without that articulation as well.

Genuine understanding has a healthy epistemic role to play in theory testing and development. But this naturalistic theory of scientific explanation should also account for explanation's failures. I have described a number of the psychological sources of failure that lie behind our tendency to be seduced by the sense of understanding.[25]

Other attempts to specify effective conditions for the acceptance of an explanation underestimate the theory-dependence of these judgments. It may be that people are more likely to accept an explanation if it seems applicable to other settings of prediction and manipulation (Lombrozo and Carey). But this is a deeply theory-dependent judgment, and if your theories are poor you are likely to identify poor explanations as useful. Surely this was part of the unifying promise of supernaturalism, of alchemy, or of the humoral theory of health. There seems no merely instrumental or pragmatic measure, such as usefulness, that tells us when an explanation should be accepted, even if there are more robust theoretical standards that might recommend acceptance (for example, that a theory is mature and has unified a diverse range of phenomena).

Moreover, the understanding conveyed by a good explanation may be a community achievement. Except for the simplest of events, explanatory understanding is not essentially an achievement of an individual. And any alternative account of explanation that requires the transmission of a sense of understanding must address this criticism. My positive account of scientific explanation asserts that, as a contingent matter of fact, the only feature of an explanation that can render explanation epistemic is its systematic tendency to produce increasingly accurate theories. In effect, only explanations capable of sustaining theoretical progress are good explanations. This pronouncement may not help us to decide now, rather than in retrospect, which explanation to take seriously. But this is not the job of an account of explanation. A theory of scientific explanation should not attempt to predict the future history of science, but set out what scientific explanation is, and what standards should be met by a proper explanation.

In order to accord explanation the epistemic role it seems to play in successful theory selection in contemporary science, we must abandon our attachment to the comforting idea that the 'sense of understanding' is a cue to at least a working version of the truth. But this will not be easy. Explanation is a backward-looking affair and 30 years of research on judgment shows both that people are not good at tracking how they are affected by knowledge of outcomes and that they are not good at admitting this limitation.

A distinctly philosophical analysis of explanatory understanding may include a role for the sense of understanding. But it is unlikely that this role

will be a justificatory one. In the series of cognitive steps that lead to understanding, phenomenology is a latecomer. In the absence of independent evidence of its reliable role – that its presence covaries importantly with progressive findings – and in light of the psychological and historical evidence that it is an unreliable cue, the sense of understanding is not a promising route to genuine understanding.

In fact, no one has the vaguest idea how this phenomenology is related to getting things right, so it is a field ripe for exploration. Scientific realists can assign a robust role to objective factors in explanation – such as statistical and causal relevance – and value the contribution of explanation to scientific progress. If your focus is not balanced by a positive account of the sense of understanding in a scientific theory of explanation, then it is easy to portray any criticism of the sense of understanding as the first step toward explanatory nihilism. But explanatory nihilism is surely premature. Disraeli once said '[t]here is no waste of time in life like that of making explanations'.[26] I disagree with Disraeli not just because there are bigger wastes of time, but because a positive account of the sense of understanding – one that is properly scientifically motivated – may help us to explain the phenomenon of theoretical progress itself. But even failing that, it can at least help us to clear the ground of philosophical idols. And one of these idols is the sentimental, Enlightenment attachment to the idea that getting something right is reliably attended by a subjective sense of intellectual authority.

Acknowledgment

This paper draws on themes and examples in Trout, 'Scientific Explanation'. I would like to thank Tania Lombrozo for conversations about psychological research on explanation, and for providing very useful comments on an earlier draft. I also want to thank Michael Strevens for helpful comments, as well as two anonymous referees. Finally, I want to thank a Central APA audience for questions and comments at a session on this topic in April 2006, and a Workshop audience at the University of Pittsburgh on Causation, Mechanisms, and Psychology, for the opportunity to present some of the ideas set out here.

Short Biography

J. D. Trout works at the borders of philosophy of science, epistemology, and psychology. He has published in *Philosophy of Science*, *Nous*, *Psychological Review*, *Current Directions in Psychological Science*, and the *British Journal for the Philosophy of Science*. His experimental research in speech perception has been published in *Language and Speech* and *Speech Communication*. He has authored or co-authored a number of books: *Measuring the Intentional World* (Oxford, 1998) argues that progress in the social and behavioral sciences warrants a measured argument for scientific realism in those domains;

Epistemology and the Psychology of Human Judgment (with Michael Bishop; Oxford, 2005), proposes that the Standard Analytic Epistemology practiced in the English-speaking world should be replaced by Ameliorative Psychology. Trout is currently working on a book (Viking/Penguin, forthcoming) on the contribution that psychological research on judgment can make to human welfare. He has held a National Science Foundation pre-doctoral fellowship, a Mellon post-doctoral fellowship, a National Endowment for the Humanities Summer Fellowship, and a National Science Foundation Scholars Award. Before arriving at Loyola University Chicago, Trout taught at Bryn Mawr College and Virginia Tech. He holds a B.A. in Philosophy and History from Bucknell University, and a Ph.D. in Philosophy from Cornell University.

Notes

* Philosophy Department and the Parmly Hearing Institute, Loyola University Chicago, 6525 North Sheridan Road, Chicago, IL 60626, USA. Email: jtrout@luc.edu.

[1] Although I will not defend the claim here, I believe that models play a central role in the most promising objectivist account of explanation. Models are durable repositories of explanatory information. They are not fleeting impressions in the scientist's mind, and they remain long after the often fanciful occasions of their production. Because we can actually represent their parts, they give us a permanent record against which to judge how our theories fare under test.

[2] Humphreys is clearly critical of the sense of understanding induced by distorted representations, but he is no 'understanding skeptic'; some representations induce genuine understanding. Some philosophers have found this distinction difficult to grasp (in particular, see de Regt; Grimm). In fact, I am not aware of any such skeptic, one who claims that understanding is not an epistemic good but instead merely a 'subjective feeling'. Like Humphreys's, other positions also distinguish between 'counterfeit understanding' and 'genuine understanding' (Trout, 'Scientific Explanation' 212, 214–15; Trout, 'Paying the Price' 208). The sense of understanding may provide no obvious or 'certain signs' of its reliability, but that doesn't mean it is useless, nor does it mean that those who make this observation are 'understanding skeptics'. Instead, it indicates that we have much to uncover about the relation between explanation and genuine understanding.

[3] Lombrozo ('Simplicity and Probability') finds that people prefer simpler explanations. Philosophers have often wondered whether our tendency to value an explanation's simplicity is based on an ungrounded assumption that simple explanations are more likely to be true. This preference may not be groundless, but according to Lombrozo's findings, people may think that simpler explanations are more likely to be true than frequency information actually warrants.

[4] Against the covering-law model, Scriven emphasizes that we often embrace particular statements as explanations even when no statement about the phenomenon to be explained was derivable from it. Toulmin is another classic alternative to the formal treatment of the received view of explanation that dominated at that time.

[5] I will leave unaddressed the plausibility of a 'community internalism', according to which explanatory justification terminates in the transparent understanding of an appropriate person. The nature of deferred knowledge is a deep and interesting issue, however, and I discuss it in the final section of this article.

[6] Wittgensteinians and some pragmatists (for example) are suspicious of any explanatory account of behavior that adverts to ineliminable internal representations. But contemporary psychology treats this issue as settled in favor of the causal role of internal representations, and in empirical matters of the mind I cast my lot with psychology.

[7] Peter Lipton points out that the understanding associated with explanation is simply additional knowledge, not some kind of 'superknowledge' (207).

[8] Beyond these observations, I will not argue that the contemporary focus on global features of explanation invites us to treat subjective 'fit' or 'overall coherence' as a benchmark of successful

explanation. Because no theory of explanation explicitly defends this feature, this article does not treat any specific account as the target of my critique. Instead, I assume the above observations establish that the internalist notion is typically at play in many of the influential contemporary theories of explanation. Positions asserting the contrary must match the evidence presented here.

[9] The task of the Dunker candle problem is to fix a candle to the wall and light it using just matches, a candle, and a matchbox filled with thumbtacks. The solution, often slow to dawn on people, is to fix the matchbox to the wall with the tacks (requiring that you see the matchbox as a support rather than a container).

[10] One night, the 19th-century chemist August Kekule labored over the structure of compounds until he

> . . . fell into a reverie, and lo, the atoms were gamboling before my eyes. Whenever, hitherto, these diminutive beings had appeared to me, they had always been in motion; but up to that time, I had never been able to discern the nature of their motion. Now, however, I saw how, frequently, two smaller atoms united to form a pair; how a larger one embraced two smaller ones; how still larger ones kept hold of three or even four of the smaller; whilst the whole kept whirling in a giddy dance. I saw how the larger ones formed a chain. . . . I spent part of the night putting on paper at least sketches of these dream forms. (qtd. in Rothenberg 291)

[11] This sense of understanding is most likely at the bottom of many standards for good explanation, such as simplicity (consider Watson's claim about his and Crick's completed DNA model that '[A] structure this pretty just had to exist' [205]), or the rendering the unfamiliar to the familiar. These standards, however, are complicated. What counts as simple or familiar is theory-dependent, and not surprisingly, verdicts in particular cases are controversial. For example, the familiarity account may seem to be violated by quantum mechanics, in which it appears that the familiar is being explained in terms of the less familiar. But this appearance may be misleading. An idea that 'feels unfamiliar' to the outsider may convey a feeling to the scientist that the pieces of a theoretical puzzle have just fallen into place. Only a comprehensive overview can treat all of these issues. The reader is referred to Salmon ('Scientific Explanation' 14), where the reduction of the unfamiliar to the familiar is explicitly discussed.

[12] Some of the most ill-fated enterprises were mounted and sustained by thinkers with supreme confidence in their explanations. Consider the alchemist Paracelsus, who claimed to have found the 'Universal Medicine' in his *Archidoxis: Comprised in Ten Books*, on the grounds that 'By means thereof I have cured the leprosy, venereal disease, dropsy, the falling sickness, colic, scab, and similar afflictions; also lupus, cancer, noli-me-tangere, fistulas, and the whole race of internal diseases, more surely than one could believe' (35).

[13] Gopnik (121) likens the feeling of satisfaction conveyed by an explanation to orgasm.

[14] Thagard provides a deep and detailed account of coherence.

[15] This article focuses exclusively on the influence of psychological biases on explanation. Solomon discusses empirical findings concerning rationality in theory evaluation.

[16] Studies such as Arkes, Christensen, Lai, and Blumer demonstrate that a decision-maker can eliminate overconfidence when exposed to rigorous feedback. However, neither the working scientist nor anyone else is normally subjected to such calibrating efforts.

[17] See Bishop and Trout ('50 Years of Successful Predictive Modelling', *Epistemology and the Psychology of Human Judgment*).

[18] For a more sanguine interpretation of the hindsight bias, which concedes its occurrence but attempts to limit its significance, see Gigerenzer, Todd, and the ABC Group.

[19] Philosophers of science have long expressed caution that ad hoc hypotheses and explanations are bad. There may be something to the idea that this reflects a sensitivity to the problem of hindsight bias. But the hindsight bias probably gets nutrition from allied insensitivities, and our sensitivity to it is probably not basic. Ad hoc hypotheses/explanations are bad because they violate severity constraints on testing. These constraints are designed to insure that the probability of a hypothesis is high on *any* theory that could be tested. When this constraint is violated, it is often because people haven't recognized that each hypothesis tested is disjunctive, and the total probability of disjunctive hypotheses is additive.

[20] One conversational strategy clearly illustrates the pervasiveness of the hindsight bias. Tell your interlocutor about an experimental finding, but lie about the punch line. For example, tell them Darley and Latane found that the more people are present when someone needs help, the more

likely an individual is to provide help. This is false, but your interlocutor will happily provide an 'obvious' explanation for the counterfeit effect.

[21] Lombrozo and Carey; Lombrozo, 'Why Adaptationist Explanations are so Seductive'.

[22] See my introductory essay, 'Reductionism and the Unity of Science' in Boyd, Gasper, and Trout 387–92.

[23] Lombrozo ('Structure and Function of Explanations') provides a very careful overview of the psychological literature on explanation, with special attention to the relation between explanation and learning.

[24] An account friendly to this one can be found in Strevens.

[25] Also see Trout, 'Paying the Price for a Theory of Explanation'.

[26] Apparently he made this assertion in 1873, during a speech entitled 'University Education Bill'.

Works Cited

Achinstein, P. *The Nature of Explanation*. New York, NY: Oxford UP, 1983.

Arkes, H., C. Christensen, C. Lai, and C. Blumer. 'Two Methods of Reducing Overconfidence'. *Organizational Behavior and Human Decision Processes* 39 (1987): 133–44.

Baddeley, A. *Human Memory: Theory and Practice*. 2nd ed. Boston, MA: Allyn and Bacon, 1996.

Bishop, M. and J. D. Trout. *Epistemology and the Psychology of Human Judgment*. New York, NY: Oxford UP, 2005.

——. '50 Years of Successful Predictive Modeling Should be Enough'. *Philosophy of Science* 69 Suppl. (2002): S197–S208.

Bloom, R. F. and E. G. Brundage. 'Predictions of Success in Elementary School for Enlisted Personnel'. *Personnel Research and Test Development in the Naval Bureau of Personnel*. Ed. D. B. Stuit. Princeton, NJ: Princeton UP, 1947. 233–61.

Boyd, R., P. Gasper, and J. D. Trout, eds. *The Philosophy of Science*. Cambridge, MA: MIT Press, 1991.

Brewer, W., C. Chinn, and A. Samarapungavan. 'Explanation in Scientists and Children'. *Minds and Machines* 8 (1998): 119–36.

Church, B. and D. Schacter. 'Perceptual Specificity of Auditory Priming: Implicit Memory for Voice Intonation and Fundamental Frequency'. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20 (1994): 521–33.

Darley, J. and B. Latane. 'Bystander Intervention in Emergencies: Diffusion of Responsibility'. *Journal of Personality and Social Psychology* 8 (1968): 377–83.

Darwin, C. *The Origin of Species*. 6th ed. New York, NY: Collier, 1962 [1859].

Dawes, R. 'A Message from Psychologists to Economists: Mere Predictability Doesn't Matter Like It Should (Without a Good Story Appended to It)'. *Journal of Economic Behavior & Organization* 39 (1999): 29–40.

——, D. Faust, and P. Meehl. 'Clinical Versus Actuarial Judgment'. *Science* 243 (1989): 1668–74.

DeVaul, R. A., F. Jervey, J. A. Chappell, P. Carver, B. Short, and S. O'Keefe. 'Medical School Performance of Initially Rejected Students'. *Journal of the American Medical Association* 257 (1957): 47–51.

Disraeli, B. 'University Education Bill' (A public speech). 1873.

Fischhoff, B. and R. Beyth. '"I Knew It Would Happen": Remembered Probabilities of Once-Future Things'. *Organizational Behavior and Human Performance* 13 (1975): 1–16.

——, P. Slovic, and S. Lichtenstein. 'Knowing with Certainty: The Appropriateness of Extreme Confidence'. *Journal of Experimental Psychology: Human Perception and Performance* 3 (1977): 552–64.

Fodor, J. 'Special Sciences' (1974). *The Philosophy of Science*. Eds. R. Boyd, P. Gasper, and J. D. Trout. Cambridge, MA: The MIT Press, 1991. 429–41.

Friedman, M. 'Explanation and Scientific Understanding'. *Journal of Philosophy* 71 (1974): 5–19. Reprinted in *Theories of Explanation*. Ed. J. C. Pitt. New York, NY: Oxford UP, 1988, 188–98. Page references are to this reprint.

Garfinkel, A. *Forms of Explanation*. New Haven, CT: Yale UP, 1981.

Giere, R. *Explaining Science: A Cognitive Approach*. Chicago, IL: U of Chicago P, 1988.

Gigerenzer, G., and P. Todd, and the ABC Group. *Simple Heuristics that Make Us Smart*. New York, NY: Oxford UP, 1999.

Goldinger, S. 'Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory'. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22.5 (1996): 1166–83.
Gopnik, A. 'Explanation as Orgasm'. *Minds and Machines* 8 (1998): 101–18.
Grimm, S. 'Is Understanding a Species of Knowledge'. *The British Journal for the Philosophy of Science* 57.3 (2006): 515–35.
Hardcastle, G. 'A Problem-Solving Account of Scientific Explanation'. Forthcoming.
Hardwig, J. 'Epistemic Dependence'. *Journal of Philosophy* 82 (1985): 335–49.
——. 'The Role of Trust in Knowledge'. *Journal of Philosophy* 88 (1991): 693–708.
Hastie, R. and N. Pennington. 'Explanation-Based Decision Making'. *Judgment and Decision Making: An Interdisciplinary Reader*. Eds. T. Connolly, H. Arkes, and K. R. Hammond. 2nd ed. New York, NY: Cambridge UP, 2000. 212–28.
Hempel, C. *Aspects of Scientific Explanation*. New York, NY: The Free Press, 1965.
Henrion, M. and B. Fischoff. 'Assessing Uncertainty in Physical Constants'. *American Journal of Physics* 54 (1986): 791–8.
Humphreys, P. *Chances of Explanation*. Princeton, NJ: Princeton UP, 1989.
James, W. *Principles of Psychology: Volume 2*. Cambridge, MA: Harvard UP, 1890/1981.
Keil, F. 'Explanation and Understanding'. *Annual Review of Psychology* 57 (2006): 227–54.
Kelemen, D. 'Beliefs about Purpose: On the Origins of Teleological Thought'. *The Descent of Mind*. Eds. M. Corballis and S. Lea. Oxford: Oxford UP, 1999. 278–94.
Kitcher, P. *The Advancement of Science*. New York, NY: Oxford UP, 1993.
——. 'The Division of Cognitive Labor'. *Journal of Philosophy* 87 (1990): 5–21.
——. 'Explanatory Unification'. *Philosophy of Science* 48 (1981): 507–31. Reprinted in *Theories of Explanation*. Ed. J. C. Pitt. New York, NY: Oxford UP, 1988. 167–87. Page references are to this reprint.
——. 'Explanatory Unification and the Causal Structure of the World'. *Scientific Explanation: Minnesota Studies in the Philosophy of Science, Vol. 13*. Eds. P. Kitcher and W. Salmon. Minneapolis, MN: U of Minnesota P, 1989. 410–505.
Koriat, A. 'Dissociating Knowing and the Feeling of Knowing: Further Evidence for the Accessibility Model'. *Journal of Experimental Psychology: General* 124.3 (1995): 311–33.
——. 'How Do We Know that We Know? The Accessibility Model of the Feeling of Knowing'. *Psychological Review* 100.4 (1993): 609–39.
Lewis, D. 'Causal Explanation'. *In Philosophical Papers, II* (1986): 214–40. New York, NY: Oxford UP. Reprinted in *Explanation*. Ed. D. Ruben. New York, NY: Oxford UP, 1993. 182–206. Page references are to this latter reprint.
Lipton, P. 'Contrastive Explanation'. *Explanation and Its Limits*. Ed. D. Knowles. Cambridge UP, 1990. Reprinted in *Explanation*. Ed. D. Ruben. New York, NY: Oxford UP, 1993. 207–27. References are to this latter reprint.
Lloyd, E. *The Case of the Female Orgasm: Bias in the Science of Evolution*. Cambridge, MA: Harvard UP, 2005.
Lombrozo, T. 'Simplicity and Probability in Causal Explanation'. *Cognitive Psychology* 2007, in press.
——. 'The Structure and Function of Explanations'. *Trends in Cognitive Science* 10.10 (2006): 464–70.
——. 'Why Adaptationist Explanations are So Seductive'. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 2005. Erlbaum. 2516.
—— and S. Carey. 'Functional Explanation and the Function of Explanation'. *Cognition* 99 (2006): 167–204.
Miller, R. W. *Fact and Method*. Princeton, NJ: Princeton UP, 1987.
Milstein, R. M., L. Wildkinson, G. N. Burrow, and W. Kessen. 'Admission Decisions and Performance During Medical School'. *Journal of Medical Education* 56 (1981): 77–82.
Newton-Smith, W. H. 'Explanation'. *A Companion to the Philosophy of Science*. Ed. W. H. Newton-Smith. Malden, MA: Blackwell, 2000. 127–33.
Nygaard, L. C., M. S. Sommers, and D. B. Pisoni. 'Speech Perception as a Talker-Contingent Process'. *Psychological Science* 5.1 (1994): 42–6.
Oskamp, S. 'Overconfidence in Case Study Judgments'. *Journal of Consulting Psychology* 63 (1965): 81–97.
Paracelsus, T. *Of the Tincture of the Philosophers*. Trans. J. H. Oxon. London: J. H. Oxon, 1665.

Pearl, J. *Causality: Models, Reasoning, and Inference*. New York, NY: Cambridge UP, 2000.
Peirce, C. S. 'Doctrine of Chances' (1878). Reprinted in Charles S. Peirce, *Writings of Charles Sanders Peirce: A Chronological Edition, Volume 3*. Bloomington, IN: Indiana UP, 1982. 276–89.
——. 'A Neglected Argument for the Reality of God'. *The Hibbert Journal* 7 (1908): 90–112.
Perrin, J. *Les Atomes*. Paris: Alcan, 1913.
Preston, J. and N. Epley. 'Explanations vs. Applications: The Explanatory Power of Valuable Beliefs'. *Psychological Science* 16.10 (2005): 826–32.
Ptolemy, C. *Almagest*. Trans. G. J. Toomer. New York, NY: Springer-Verlag, 1984.
Railton, P. 'Probability, Explanation, and Information'. *Synthese* 48 (1981): 233–56. Reprinted in *Explanation*. Ed. D. Ruben. New York, NY: Oxford UP, 1993. 160–81. Page references are to this reprint.
de Regt, H. W. 'Discussion Note: Making Sense of Understanding'. *Philosophy of Science* 71.1 (2004): 98–109.
—— and D. Dieks. 'A Contextual Approach to Scientific Understanding'. *Synthese* 144.1 (2005): 137–70.
Rogeberg, O. 'Taking Absurd Theories Seriously: Economics and the Case of Rational Addiction Theories'. *Philosophy of Science* 71.3 (2004): 263–85.
Rothenberg, A. 'Creative Homospatial and Janusian Processes in Kekule's Discovery of the Structure of the Benzene Molecule'. *The Kekule Riddle*. Ed. John H. Wotiz. Clearwater, FL: Cache River Press, 1993. 286–309. (Translation of Kekule quotations by Francis R. Japp.)
Rozenblit, L. and F. Keil. 'The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth'. *Cognitive Science* 26 (2002): 521–62.
Salmon, W. 'Four Decades of Explanation'. *Scientific Explanation: Minnesota Studies in the Philosophy of Science, Vol. 13*. Eds. P. Kitcher and W. Salmon. Minneapolis, MN: U of Minnesota P, 1989. 3–219.
——. 'The Importance of Scientific Understanding'. *Causality and Explanation*. W. Salmon. New York, NY: Oxford UP, 1998. 79–91.
——. 'Scientific Explanation'. *Introduction to the Philosophy of Science*. S. Merrilee et al. Englewood Cliffs, NJ: Prentice-Hall, 1992. 7–41.
——. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton UP, 1984.
Scriven, M. 'Explanation, Predictions, and Laws'. *Minnesota Studies in the Philosophy of Science*. Eds. H. Feigl and G. Maxwell. Minneapolis, MN: U of Minnesota P, 1962. 170–230.
Shtulman, A. 'Qualitative Differences Between Naive and Scientific Theories of Evolution'. *Cognitive Psychology* 52.2 (2006): 170–94.
Solomon, M. 'Scientific Rationality and Human Reasoning'. *Philosophy of Science* 59 (1992): 439–55.
Spirtes, P., C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. 2nd ed. Cambridge, MA: MIT Press, 2000.
Strevens, M. *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard UP, forthcoming.
Thagard, P. 'Explanatory Coherence'. *Behavioral and Brain Sciences* 12 (1989): 435–67.
Toulmin, S. *Foresight and Understanding*. London: Hutchinson, 1961.
Trout, J. D. *Measuring the Intentional World: Realism, Naturalism, and Quantitative Methods in the Behavioral Sciences*. New York, NY: Oxford UP, 1998.
——. Paying the Price for a Theory of Explanation: de Regt's Discussion of Trout (2002). *Philosophy of Science* 72 (January 2005): 198–208.
——. 'Scientific Explanation and the Sense of Understanding'. *Philosophy of Science* 69.2 (2002): 212–33.
——. 'Theory Conjunction and Mercenary Reliance'. *Philosophy of Science* 59 (1992): 231–45.
van Fraassen, B. C. *The Scientific Image*. Oxford: Clarendon Press, 1980.
Watkins, J. 'Methodological Individualism and Social Tendencies'. *British Journal for the Philosophy of Science* 8 (1957): 104–17.
Watson, J. D. *The Double Helix*. New York, NY: Atheneum, 1968.
Woodward, J. *Making Things Happen*. New York, NY: Oxford UP, 2003.
——. 'A Theory of Singular Causal Explanation'. *Erkenntnis* 21 (1984): 231–62. Reprinted in *Explanation*. Ed. D. Ruben. New York, NY: Oxford UP, 1993. 246–74. Page references are to this reprint.
Wright, W. 'Explanation and the Hard Problem'. *Philosophical Studies* 132 (2007): 301–30.

