The Relief of Explanation

Nietzsche, Formal Necessity, and the Feeling of Understanding

Flyxion
Independent Researcher

October 2026

Abstract

Certain accounts feel explanatory before anyone has established that they explain anything. This essay proposes that much explanatory relief is the felt closure at the point where an account is adopted as a classification of situations. Adopting a classification discards the differences among the situations it identifies, and the essay states a necessary condition on what may be discarded: any two situations the account identifies must have the same continuations for the question asked, so that no admissible observation or intervention bearing on the question could later tell them apart. The condition constrains what an explanation may leave out, and it does not by itself make an account explanatory. Familiarity, simplicity, practical reassurance, and mathematical demonstrability are each genuine achievements, and each is a distinct route by which a compression can be adopted without meeting this condition. Nietzsche's suspicion that the relief of recognition is mistaken for knowledge supplies the provocation, and two small cases carry the argument: a cracked wall, where a familiar cause ends inquiry before it separates cases that further inquiry would separate, and a pole's shadow, where a correct equation is read in a causal direction it does not supply. The resulting test also explains why better understanding so often begins by making an account less simple, and why an explanation earns trust to the extent that it can be refined when a collision is found.

1. The Moment When an Explanation Feels Complete

An event troubles us until someone supplies a cause, and then it seems settled, though the evidence available has scarcely changed. A wall cracks along a diagonal above a window, a machine stops in the middle of a shift, a colleague who is usually punctual fails to arrive and offers no word. In each case a short phrase is available that closes the matter: weak mortar, a faulty component, that is simply the kind of person he is. The phrase supplies a recognizable category, the category carries with it a customary response, and the pressure to keep investigating subsides. Anyone who has watched a meeting pivot from puzzlement to agreement on the strength of one confident sentence will know how quickly the transition occurs, and how little of the transition is visible to the people undergoing it.

What this release establishes is the first question to ask. Something has certainly happened to the inquirer: uncertainty has diminished, the event has acquired a familiar shape, and action once again feels possible. Whether something corresponding has been discovered about the event itself is a separate matter, and the experience of relief does not by its own character say which has occurred. The relief that follows a correct diagnosis and the relief that follows a plausible but mistaken one are, from the

inside, close to indistinguishable. This is not an exotic observation. Studies of what people believe they understand have found that confidence in one's grasp of ordinary mechanisms, such as how a zipper or a flush toilet works, regularly exceeds the ability to produce an actual explanation when asked to give one [12]. The feeling of having understood is a cognitive event with its own causes, and those causes are not guaranteed to coincide with the causes of the thing understood.

It is useful to name the two things whose coincidence is in question.

Definition 1 (Explanatory satisfaction)**.** Explanatory satisfaction is the experience, in an inquirer, that an account makes sense of an event: the sense that a question has been answered, that the event is no longer anomalous, and that inquiry may stop.

Definition 2 (Explanatory adequacy)**.** Explanatory adequacy is the success of an account in answering a specified question about an event, under conditions that can be stated and defended, and against the alternatives that could plausibly be offered in its place.

The two can coincide, and in well-functioning inquiry they frequently do. The thesis of this essay is not that the feeling of understanding is worthless or that it always misleads. It is that their coincidence is a result to be accounted for, and that a long history of accounts which felt complete and were not should make us attentive to the circumstances in which the feeling is produced by something other than adequacy. The proposal defended here can be stated more exactly. An accepted account classifies situations, and in doing so it discards the differences among the members of each class. This discarding is what the essay calls compression, in a specific sense: the details of a situation, and with them the parts of an inquiry's history that bear only on those details, are replaced by a class label that is thereafter treated as sufficient for the question. The term does not mean merely the shortening of a description. The proposal is that much explanatory relief is the felt closure at the point where such a classification is adopted, although relief can occur without a newly accepted account and an account can be accepted without relief, and the proposal concerns the cases in which the two go together. What the classification owes the inquirer can be put as a condition on what it identifies: if the account treats two situations as the same, then no further inquiry bearing on the question should be able to tell them apart.

Definition 3 (Continuation of a situation)**.** Fix a question Q, a domain D of situations, a family Π_Q of admissible inquiry protocols (sequences of observations and interventions whose later steps may depend on earlier results), and a record R_Q of the outcomes relevant to Q. For $s \in D$ and $\pi \in \Pi_Q$, the continuation is the distribution $K_Q(s,\pi) = \mathscr{L}(R_Q \mid s,\pi)$ of those outcomes when π is carried out in s. A deterministic situation yields a distribution concentrated on a single result.

Definition 4 (Continuation-sufficiency)**.** An account A, which assigns to each situation s a state $A(s)$, is continuation-sufficient for Q on D when, for all $s, s' \in D$, $A(s) = A(s')$ implies $K_Q(s,\pi) = K_Q(s',\pi)$ for every $\pi \in \Pi_Q$. A pair that violates the implication is a collision: two situations the account merges that some admissible inquiry bearing on Q separates.

Explanatory satisfaction attaches to the adoption of A, and continuation-sufficiency is a property of A. The condition is a constraint on what an account may discard, and it is not a definition of explanation. An account that assigned every situation its own identifier would have no collisions, and a complete

archive of situations would pass for the same reason, yet neither need explain anything. Explanatory relevance, empirical warrant, and intelligibility remain further obligations, and the claim is only that an account used as sufficient for a specified inquiry must preserve the distinctions that could change that inquiry's answers. Accepting such a classification allows an inquirer to set aside parts of the inquiry's history, and this is where the condition bears on the felt closure described above, without the present map being a formalization of that replacement itself. In practice the condition can be checked only approximately and in degrees, by searching for collisions and estimating how costly the ones found would be, so it offers a standard for the conduct of inquiry and not a procedure that certifies explanations. The next sections examine the routes by which a map can be adopted while collisions remain undetected, beginning with the one Nietzsche made famous, and then turn to a less obvious case: the confidence that attends a formula or a proof. A later section returns to the condition and applies it to all four routes at once.

2. Nietzsche's Suspicion: Recognition Masquerading as Knowledge

Nietzsche's challenge, as reconstructed here, is a question about the standards by which explanations are accepted. In the fifth book of *The Gay Science* he takes up the origin of our concept of knowledge and proposes that to know, in the ordinary sense, is to trace something strange back to something familiar [7]. The unfamiliar is felt as a burden, perhaps even as a threat, and the moment it is assimilated to what is already known the burden lifts. He suggests that this lifting is then treated as a mark of truth: what has relieved us is taken to have been understood. A related passage in *Twilight of the Idols*, in the discussion of what he calls the error of false causes, makes the same point about causal explanation in particular, observing that the first thought which accounts for an unknown event in terms of something known tends to be accepted simply because it brings comfort [9]. What follows in this essay is a reconstruction of that line of thought in paraphrase, together with a good deal that Nietzsche did not say. Where the argument goes beyond his texts it is offered as the essay's own.

The central step in the reconstruction is that familiarity is a fact about the inquirer's history, and that the structure of an occurrence is a fact about the occurrence. An account may feel natural because it resembles accounts we have given before, because the vocabulary it uses is the vocabulary of our trade or our upbringing, or because the category it invokes has been rehearsed until it requires no effort. None of these properties bears on whether the category fits. A trained mechanic, a physician, and a structural engineer each possess a repertoire of recognizable failures, and in each case the repertoire is a hard-won achievement. It is also a set of grooves into which a case may slide without having been examined. The same fluency that makes expert diagnosis fast makes expert error confident.

It follows that two activities should be kept apart. Locating an event within an established vocabulary is one, and discovering why it happened is another. Naming a process can orient inquiry, since it tells us where to look, what has been seen before, and what to try. The same act of naming can terminate inquiry when the name is mistaken for a mechanism, and the temptation is strongest when the name is itself an explanation-shaped word. To say that a medicine works because it has a soporific virtue, to say that a market fell because of panic, or to say that a student failed because of laziness is to supply a noun where an account of operation was wanted. Recognition is indispensable to knowledge. No inquiry begins from nothing, and a great deal of what is called understanding consists precisely in assimilating the new

to the old in a disciplined way. The question is under what circumstances recognition substitutes for knowledge rather than serving it.

Acceptability and adequacy are therefore governed by different conditions. The psychological conditions that make an explanation acceptable (familiarity, fluency, a sense of closure) need not be the evidential conditions that make it adequate (discrimination among alternatives, sensitivity to the facts of the case, exposure to correction). When the first set is mistaken for the second, relief is read as discovery.

3. Comfort, Strength, and the Acceptance of Belief

Beliefs can restore confidence, support coordinated action, and make a threatening situation manageable. These benefits are real, and any account that dismissed them would misdescribe why people hold the beliefs they do. A community that believes its water is safe will drink it and carry on with its work, a team that believes a plan is sound will execute it with energy, and a patient who believes a treatment will help may adhere to it more faithfully. Each of these describes something a belief enables its holder to do. Each is a different matter from whether the belief accurately represents its subject, and the practical value of a belief gives no direct evidence about the second question. A belief can be extremely useful and false, as when a confident navigator steers with the wrong chart and arrives by luck, and a belief can be true and disabling, as when an accurate prognosis paralyzes the person who receives it.

It would be a mistake to reduce this point to conscious wishful thinking. The more instructive cases are those in which an explanation reassures precisely because it appears impersonal and rigorous. Its comfort may consist in making events seem ordered, calculable, or attributable to identifiable agents, with no promise of any pleasant outcome. A grim account can relieve uncertainty as effectively as a cheerful one. A person who learns that a loss was caused by a specific, nameable decision may find the knowledge painful, and yet find it easier to bear than the prior condition in which the loss had no assignable source. A theory according to which events are driven by hidden coordinating powers offers a bleak world, but also a legible one, and the legibility is a form of consolation that has nothing to do with optimism. Even an explanation that frightens can relieve in this way, because fear of a known thing is often easier to carry than the unspecific dread of an unknown one.

The corresponding asymmetry deserves emphasis. A useful belief can be false, and a well-supported account can initially weaken our confidence by exposing dependencies and limitations that the cruder story concealed. An engineer who learns that a structure's safety depends on three assumptions, none of which has been tested, is less comfortable than before, and is also better informed. A physician who replaces a confident diagnosis with a differential of four possibilities has made the clinical picture more difficult to hold in mind, and more accurate. Neither reassurance nor discomfort is therefore a reliable sign of truth. Reassurance can follow adequacy, but so can it follow the opposite. Discomfort can signal that an account has become more honest, and it can equally signal that an account has become merely more alarming. The felt valence of an explanation is evidence about the inquirer's state and, at most, indirect evidence about the explanation's standing.

4. Simplicity and the Cost of Leaving Things Out

The claim that simple explanations are suspect needs to be qualified at once, because it is false as a general proposition. Explanation necessarily selects, abstracts, and compresses. An account that reproduced every detail of an event would typically obscure the relationship the inquirer wanted to understand, in the way that a map at the scale of the territory is a perfectly accurate and entirely useless object. When a physicist reduces the motion of a planet to a point mass moving under a central force, or an epidemiologist reduces a population to a few compartments, the reduction is a substantial intellectual accomplishment and the simplicity is part of what the accomplishment consists in. Simplicity can be an achievement of understanding, and the history of science records many cases in which a laborious and heterogeneous description was replaced by a compact one that revealed what the laborious one had hidden.

The relevant question is therefore what the simplification discards. Return to the cracked wall. "Weak mortar" may identify a genuine contributing condition, and in a case where the mortar is demonstrably deficient it may even be the main one. It can also conceal differential settlement, moisture cycling, thermal movement, or the way loads are distributed through the openings of the wall. The diagonal crack above a window corner is a standard signature of concentrated stress at the corner of an opening, and this fact is invisible from within the vocabulary of mortar strength. An account adequate for explaining why mortar failed a strength test may be wholly inadequate for explaining why this wall cracked at this location, and the two questions are easily conflated because both can be answered by the same short phrase.

Principle 1 (Relative adequacy of simplification)**.** A simplification is adequate insofar as the distinctions it omits do not defeat its answer to the question being asked.

Stated in ordinary prose, the principle says that the explanatory acceptability of a simplification depends on the question it is asked to answer and the domain in which it is applied. Simplicity must be assessed relative to a purpose, a domain, and the plausible alternatives that the simplification must outperform. A model that ignores friction is adequate for a thought experiment about inertia and inadequate for a braking distance. An explanation that omits the mortar's moisture history is adequate if the wall's failure does not depend on moisture and inadequate if it does, and discovering which is the case requires the very investigation the simple answer was about to forestall. The cost of leaving things out is therefore not fixed in advance. It is paid, or not, depending on which of the omitted distinctions turns out to matter, and an account that does not state what it has omitted has withheld the information needed to judge whether the price was acceptable.

5. The Grammar of Doers and Deeds

Attention can now turn from the attractiveness of familiar causes to the form that familiar explanations take. Ordinary sentences commonly organize events around subjects performing actions on objects: the hunter kills the deer, the wind breaks the branch, the market punishes the firm. This structure is convenient, and a language without something like it would be hard to use. Its convenience does not independently establish that every process consists of a discrete doer, a separable deed, and a passive recipient. In the first essay of *On the Genealogy of Morality*, Nietzsche presses this point with the image

of lightning, observing that ordinary speech separates the flash from the lightning and treats the flash as an action performed by a subject that might equally have refrained [8]. The grammatical habit generates a picture of a neutral substrate standing behind its manifestations, free to manifest or withhold, and the picture then feels like an observation about the world.

The temptation that results is to posit a stable agent behind an action and then to treat that agent as the explanation of the action. “He acted aggressively because he is aggressive” turns a description of conduct into an apparently independent cause, and does so by the simple device of converting an adjective into a noun-like property that resides inside its bearer. The circularity is easy to overlook because the sentence has the grammatical shape of an explanation. The difficulty is not that dispositions never explain behaviour. Many dispositions do, and some, such as the solubility of a salt or the brittleness of glass, are paradigms of respectable causal properties. The point concerns what has to be added for a disposition to carry explanatory weight: it must be characterized independently of the very behaviour it is invoked to explain, it must be tied to conditions under which it does and does not manifest, and it must be shown to be stable enough to support the predictions it implicitly makes. Without such additions the disposition is a redescription, and the explanatory work remains to be done.

A connection to the fundamental attribution error can be drawn here, though it should be drawn with care. Social psychologists have documented a tendency to privilege the agent over surrounding conditions when explaining another person’s behaviour, attributing to character what is better attributed to circumstance [11]. The resemblance is real, since both involve over-reliance on the agent as the locus of explanation. Nietzsche’s question reaches further, because it concerns the conceptual construction of agency itself, the thought that there is a doer apart from the doing, and not merely the empirical question of how much weight to give character relative to situation. The modern psychological concept and the philosophical criticism should therefore remain distinct. The first is a finding about a bias within a framework that takes agents for granted, and the second is a question about whether that framework’s central device is an observation or a habit of grammar.

6. Intention as a Template for Causation

A further possibility concerns the source of the pattern. Our experience of intending and acting may supply a model through which other events are interpreted. When we decide to lift a cup and the cup rises, the experience has a distinctive structure: a prior state of wanting, a directing of effort, a result that answers to the directing. An occurrence elsewhere appears explicable once something can be identified as directing, producing, or compelling it. Experimental work going back to the study of apparent behaviour in moving geometric figures has shown how readily observers attribute goals, pursuit, and evasion to shapes whose motion is specified only by simple kinematic rules [4]. The readiness is not confined to the laboratory. It appears in talk of a river seeking the sea, of an economy that wants growth, of selection that favors one trait, and of an algorithm that has decided something.

Two cases should be distinguished. Intentional explanation is appropriate for action: when a person closes a window because the room is cold, the account in terms of belief and desire is a good account, supported by the person’s testimony, by their other behaviour, and by their response to changed circumstances. Its propriety there does not license the projection of intentional structure onto processes for which it has not been established. Even an account framed without conscious agents can retain the

imagery of commands, obedience, purposes, or inherent striving. A system is said to "obey" a law, a substance to "seek" equilibrium, a gene to "instruct" the cell, and in each case the language imports a template of directing and compliance that the underlying description may not contain. Whether the imported structure is harmless shorthand or a quietly operative assumption is the sort of thing that has to be checked, and the felt naturalness of the phrasing is a poor guide.

None of this implies that every causal concept originates in intention. That thesis is far stronger than the evidence supports, and it would be a version of the simplifying error under discussion. The narrower claim is that familiar experiences of agency can shape causal expectations, and that those expectations require independent scrutiny when extended to other domains. An account whose persuasiveness derives from its resemblance to the experience of willing is borrowing credit from a source that has not been shown to apply. The scrutiny consists in asking what, besides the resemblance, supports the attribution of directing, producing, or compelling to the process in question. The three requirements stated for dispositions apply here as well: an attribution of belief and desire must be characterized independently of the behaviour it explains, tied to conditions under which it would and would not manifest, and stable enough to support predictions about what the agent would do in other circumstances.

7. When Calculation Becomes Explanation

The essay's central formal question can now be put: why does possession of a formula make us feel that we have discovered a cause? Mathematical representation brings extraordinary order to experience. It connects measurements that appeared unrelated, generates predictions of great precision, and exposes relationships that unaided observation misses. The achievements are not in doubt. Because they are substantial, they invite an additional inference that has not yet been earned, namely that the quantities appearing in the equation stand to one another as causes to effects, and that the equation therefore tells us why things happen as they do.

An equation may describe how quantities covary without identifying which changes produce which others. Accurate calculation establishes something important, but what it establishes depends on the model, on its interpretation, and on the evidence supporting that interpretation. The distinction is old in the philosophy of explanation, where it has long been observed that derivations can be impeccable and still fail to explain [5, 13]. A familiar illustration is the relation between a pole, its shadow, and the sun. Let h be the height of the pole, s the length of its shadow on level ground, and θ the elevation of the sun above the horizon. Elementary geometry gives

$$h = s\tan\theta, \qquad s = \frac{h}{\tan\theta}. \tag{1}$$

The first form allows us to calculate the height from the shadow, and the second allows us to calculate the shadow from the height. Both are valid, and both are correct as mathematics. Neither arrangement, considered alone, is an account of what produces what. The relation describes an unobstructed shadow on level ground under a specified illumination, and rearranging it does not establish that changing the shadow will change the pole. Shading the ground with a screen does alter the shadow, but it does so by changing the conditions under which the relation applies, because the shadow is then no longer the unobstructed one the relation describes. The screen is therefore not an intervention on shadow length

that preserves the model, and the pole's height is unaffected for reasons the equation does not itself state. The same mathematical relation supports different inferential directions without supplying equivalent causal directions. In the usual why-question, why the shadow has the length it has, the pole's height and the sun's elevation are the independent conditions and the shadow length is what they determine. That ordering is a fact about the physical situation, and it does not appear in the geometry. The asymmetry is contributed by knowledge of illumination, obstruction, and geometry, and a causal account built on that knowledge explains both why the relation holds under the stated conditions and why adding a screen invalidates its application. A reader who has forgotten that knowledge cannot recover it from (1). This is the example that Bromberger used, in the form of a flagpole, to show that a deduction can answer a why-question in one direction and fail to answer it in the other [1].

The lesson is not that calculation is a lesser achievement than explanation. It is that the feeling of having captured something, which attends a neat formula, belongs to the formula as a compression of a relationship. Whether the relationship is one of dependence, and which way the dependence runs, is a further question, answered by further evidence.

8. What a Proof Makes Necessary

The necessity of a proof must be distinguished from a claim of necessity about nature. Given specified premises, definitions, and rules of inference, a conclusion may follow necessarily, and this is among the most secure kinds of knowledge that human beings have. Applying such a conclusion to an empirical situation is a different undertaking. It requires additional judgments about whether the situation satisfies the premises, about whether the mathematical construction is the right one for the target, and about whether the interpretation placed on its terms is appropriate. The necessity belongs to the conditional: if the premises hold, the conclusion holds. It does not travel with the conclusion into the application, because the antecedent is what the application has to earn.

Three connected questions can be separated. The first concerns the derivation: whether each step is valid, so that the conclusion does follow from the premises. The second concerns the representation: whether the mathematical construction represents the target adequately, so that its premises are true, or true enough, of the system to which it is applied. The third concerns the interpretation: whether the representation supports the causal reading being proposed, so that a functional dependence among variables in the model is warranted as a dependence among processes in the world. These questions have different kinds of answers. The first concerns consequence relative to a specified formal system: a valid derivation establishes that the conclusion follows within that system, although checking that a particular inscription actually constitutes such a derivation remains a fallible activity, whether it is performed by a reader or by a proof-checking program, whose verdict depends on assumptions about its own implementation and about how the formal system was encoded. The second is settled, if at all, by measurement and by testing the assumptions against the system. The third is settled by evidence about how the system behaves when it is interfered with, observed over time, or compared across circumstances. A successful answer to the first cannot automatically settle the others.

A proof can establish the consequences of an account without establishing that the account identifies how the world produces those consequences. Deriving an orbit from a stipulated force law establishes what the model predicts; it does not independently establish the law's empirical accuracy, domain of

validity, or fundamental status. This position preserves the full force of mathematical demonstration. The theorem is as secure as it ever was. What it locates is the further obligations involved in its empirical use, which are obligations of a different kind and are not discharged by pointing to the rigor of the proof.

9. Several Accounts, the Same Observations

Predictive success may leave causal questions open, and the reason is structural. Different accounts can reproduce the same observed pattern while disagreeing about unobserved conditions, about underlying mechanisms, and about how the system would respond to intervention. This is a version of the familiar thought that theories are underdetermined by the evidence available at a given time [2]. Its force in the present setting is that a formula fitted to data does not carry with it a reading of the data, and the feeling of explanatory completion that follows a good fit is silent about the alternatives that fit equally well.

A simple common-cause example shows the structure. Suppose two quantities, X and Y, are observed to move together across many cases. At least three accounts are available. One holds that X influences Y. A second holds that Y influences X. A third holds that both respond to a further factor Z, so that the association is produced by a shared source and neither quantity affects the other. In a linear idealization the third account can be written

$$X = aZ + \varepsilon_X, \qquad Y = bZ + \varepsilon_Y, \tag{2}$$

with the noise terms independent of each other and of Z. Provided that Z has positive variance and a and b are both non-zero, X and Y are correlated. Provided the relevant variances exist and $\mathrm{Var}(X) > 0$, the population least-squares slope of Y on X is

$$\beta_{Y|X} = \frac{ab\,\mathrm{Var}(Z)}{a^2\,\mathrm{Var}(Z) + \mathrm{Var}(\varepsilon_X)}. \tag{3}$$

This coefficient can support useful prediction when the shared signal is sufficiently strong, although the structural equations contain no causal influence from X to Y. Now consider the stipulated intervention that replaces the equation for X with $X = x$, preserving the equation for Y and the exogenous distribution. It leaves the distribution of Y unchanged. Using the notation of interventional reasoning, the observational conditional and the interventional distribution generally come apart, as distributions and under the nondegeneracy conditions just stated, though not necessarily at every particular x:

$$P(Y \mid X = x) \;\neq\; P\big(Y \mid \mathrm{do}(X = x)\big), \tag{4}$$

where the left-hand side varies with x under those conditions and the right-hand side does not [10]. Selection can produce the same appearance without any common cause at all, if the cases observed were chosen in a way that depends on both quantities. Such selection-induced dependence is a further case in which the class of models compatible with the data contains members whose interventional continuations differ, here because of the selection mechanism and not because of a common cause. The formula fitted to the observed association is compatible with each of these accounts, and its predictive adequacy as a tool for forecasting Y from X is real in all of them. What differs is the answer to the question of what would happen if X were changed, and on that question the fit has nothing to say. In

the terms of the principle of relative adequacy, the fitted formula is an adequate simplification for the question of forecasting and an inadequate one for the question of manipulation, and the cost of what it omits is paid at the point where the question shifts from the first to the second.

The point is not that observation is powerless. An observation discriminates between competing accounts insofar as those accounts assign it different probabilities or make different commitments about its occurrence. Evidence equally expected under both does not favour either over the other, although it may support both against further alternatives. Repeated observations are not worthless on this view, since their value depends on whether the competing accounts remain equivalent across those observations and their conditions, and a new setting in which the accounts diverge can discriminate where the earlier ones could not. An explanation becomes stronger when it succeeds where a plausible competitor would fail [6], and it becomes no stronger by repeating a pattern that both can accommodate. This is why the felt adequacy of an account should be tested by asking what a rival would have predicted, and why the ease with which a pattern is reproduced by a single account is a weak indication that no other account could reproduce it.

10. How Formal Models Acquire Causal Warrant

The preceding criticism requires a constructive answer, and the answer is that mathematics contributes to causal understanding when it is joined to evidence supporting the interpretation of its variables, assumptions, and dependencies. A model is a structured set of claims about what depends on what, and those claims have to be checked against the world in ways that the formalism alone cannot supply. The formal apparatus then does what it does best: it makes the consequences of the claims explicit, shows which are testable, and exposes the assumptions on which a conclusion rests. A model joined to such evidence can be far more powerful than either mathematics or observation alone, and the history of the exact sciences consists largely of such unions.

Several sources of discrimination deserve mention. Interventions provide the most direct, since a dependence that survives deliberate manipulation of one variable and observation of another is a dependence of a stronger kind than a covariation [15]. Temporal relations constrain the direction of dependence, because effects do not precede their causes, although the ordering of measurements need not match the ordering of processes and care is needed. Mechanisms, understood as the intermediate steps by which an influence is transmitted, allow an inference to be checked at several places and not only at its ends. Natural experiments exploit circumstances in which something resembling an intervention occurred without being arranged. Predictions under changed conditions test whether a model captures the structure that produced the data or only the regularity the data displayed under the circumstances in which they were gathered.

None of these should be presented as a universal recipe. Opportunities for controlled intervention vary enormously across fields, and in astronomy, geology, and large parts of social science they are scarce or unavailable. Interventions themselves depend on background assumptions, such as the assumption that the manipulation affects only the variable intended, and those assumptions can fail. Mechanistic knowledge can be incomplete or wrong in ways that are hard to detect. The governing question is what entitles us to move from a relation within the representation to a dependence in the represented process. A causal account gains warrant through the quality and convergence of those connections, through

several independent lines of evidence agreeing where a single line might mislead, and it does not gain it through formal elegance alone. Elegance may be a reasonable guide to which hypotheses deserve testing, but a guide to testing is not a substitute for it.

11. When Better Understanding Makes Things Less Simple

The experience of explanatory relief from which the essay began can now be reversed. Inquiry can improve an account by making it initially less satisfying. A single cause becomes a conditional interaction, in which a factor matters only in the presence of another. An apparent exception reveals a hidden dependency that had been masked by the way cases were grouped. A stable category separates into several mechanisms that share a name and a presentation while differing in origin and in what repairs them. In each of these the later account is harder to state, harder to hold in mind, and less comforting than the earlier one, and in each it is a better account because of the distinctions it has learned to make.

Productive complication needs to be distinguished from indiscriminate complexity. Adding detail helps when the detail changes what we can explain, predict, distinguish, or repair. A new variable that separates two cases previously confused, or a boundary condition that predicts where a rule stops applying, is a gain. Complexity that contributes none of these benefits can be as intellectually complacent as an oversimplified answer, because it replaces the work of discrimination with the appearance of thoroughness. A report that lists forty contributing factors without indicating which matter, under what conditions, and how they interact is not more adequate than one that names a single factor. It has merely moved the failure from omission to indiscrimination, and it offers its own kind of relief: the sense that nothing has been left out.

This suggests an account of intellectual maturity in explanation. Understanding includes knowing which questions an explanation leaves open and which changes would require us to reconsider it. A mature inquirer can say not only what the account claims but where it is thin, what it assumes that has not been tested, and what observation would count against it. Confidence becomes better calibrated as the account's limits become more explicit. The result is not a reduction in confidence across the board. It is a redistribution, in which firmer commitment is placed where the evidence is convergent and tentativeness is placed where it is not, and the inquirer's feeling of understanding comes to follow, imperfectly but with practice, the actual state of the evidence.

12. Four Routes and One Condition

The continuation condition can now be applied to each route by which relief has been seen to outrun adequacy. In every case the question is which identification the route performs, and whether the situations it identifies share their continuations.

Familiarity identifies situations by the inquirer's history. Two cases fall under one heading because they remind us of the same earlier cases, and the equivalence is produced by what has been rehearsed, not by what the cases would do under further inquiry. This is the precise sense in which recognition can masquerade as knowledge: the partition is real and useful as a record of what the inquirer has encountered, and it has no necessary relation to the partition that the event's continuations would induce. The situations at issue are events as they present themselves to the inquirer, together with the inquirer's

state of information about them. Two events fall under one heading when the inquirer's history makes them indistinguishable, and a collision occurs when some inquiry the inquirer has not considered would distinguish them. Simplification coarsens a partition on purpose. Ignoring friction identifies all situations that differ only in friction, which is harmless for a question about inertia and costly for a question about braking distance, because the continuations separate those situations in the second case and not in the first. The cracked wall is a collision of this kind. A wall whose mortar is weak and a wall whose foundation is moving are identified by the phrase "weak mortar," and a single continuation, replacing the mortar and waiting, separates them. Practical reassurance identifies situations by the response they call for, so that all cases in which the same plan can go ahead are treated as one. The merger is harmless for the question of whether to proceed under ordinary conditions, and it becomes a collision where some inquiry, such as a test of the water or a change of weather, separates cases that the plan's past success had merged. Success under past conditions is evidence about the continuations that were tested and says nothing about those that were not. Mathematical demonstrability enters by two doors. A fitted formula identifies every model that agrees on the features of the data that the formula captures, which is generally less than the full joint distribution. These classes are at least as coarse as observational equivalence classes, and the common-cause example shows that they contain models whose continuations under intervention differ, since the inequality in (4) holds for one member of such a class and fails for another. A derivation, finally, preserves what it was asked to preserve within a formal system, namely the consequences of the stated premises. The relation between the formal object and its target is a further map, and the proof is silent on whether the identifications that map induces match the continuation structure of the target. A model can be verified without fault and still answer a question substituted for the one that was asked, because the verification concerns the derivation and the substitution occurs in the translation from the situation to its formal image.

The same condition accounts for the earlier discussion of complication. Revising an account amounts to refining its partition by splitting a class into parts. Productive complication splits classes that some continuation relevant to the question separates: the single cause becomes a conditional interaction because the cases it merged diverge under a further observation or intervention. Indiscriminate complexity splits classes that no relevant continuation separates, or enlarges the state without removing any collision, and it offers relief of a different kind, the sense that nothing has been left out. Corrective openness can then be described in part. An account is open to correction when a collision, once found, can be repaired by splitting the offending class locally without discarding the identifications that remain sound. An account that can only be rejected whole forces the inquirer to choose between holding it despite the collision and losing everything it had correctly identified. Splitting is one form of correction among several, however. Some corrections withdraw distinctions that proved irrelevant, as when an account that distinguishes patients by an irrelevant characteristic must merge categories, and some replace the representation altogether, as when the wrong mechanism was assumed or the question and domain themselves must change. Refinement, coarsening, and replacement are all available where warranted, and the availability of local splitting is not a universal test of revisability.

A last consequence concerns the feeling of understanding. If relief is often the closure of a classification, then what it reliably indicates is that a classification has been adopted, and what it indicates about collisions depends on how the feeling came to be calibrated. Corrective experience is one route: an

expert's sense of fit can be recalibrated by repeated cases in which merged situations later separated, so that closure is withheld when a situation resembles past collisions. Instruction, transferred expertise, and independently checked reasoning are others. Where none of them has operated, the feeling of closure carries little information about collisions.

13. Explanations That Survive Correction

The argument can be brought into practical judgment by comparing an account that merely assigns a cause with one that supports an effective response. Replacing mortar may temporarily conceal a crack. If the underlying cause is movement in the foundation, the crack returns, and the repeated return is itself evidence that the first diagnosis was incomplete. Identifying movement in the foundation may explain why the crack returns and may alter what repair is appropriate, directing attention to drainage, underpinning, or the loads the wall was never meant to carry. The second account is better not because it is more elaborate but because it survives the test that the first failed: it continues to make sense of the wall after the repair has been attempted.

Successful repair can provide evidence for a causal account, but it is not conclusive by itself. An intervention may compensate for a problem without identifying its origin, as when a prop holds up a sagging beam without any understanding of why it sagged. Several mechanisms may respond to the same treatment, so that the success of an antibiotic does not tell us which of several organisms was responsible, and a treatment's working does not tell us why it works. The explanation must remain answerable to alternatives even when action succeeds, and the satisfaction of a repair that held is one more feeling that has to be distinguished from the evidence that would justify it.

Corrective openness is a virtue of explanation. An account should make it possible to recognize when its assumptions fail and should specify what evidence would change it. This is partly a matter of how the account is stated: a claim hedged so thoroughly that no observation could contradict it offers the inquirer nothing to correct, and a claim stated with its conditions of application leaves a visible place where it can fail. It is also a matter of how the account is held, with readiness to revise when the conditions fail. Treated in this way, uncertainty becomes a condition of disciplined inquiry and ceases to be an embarrassment to be concealed. The explanations that last are those that were built to be corrected.

14. Objections and the Essay's Own Explanatory Temptation

Several objections deserve a direct reply. The first is that simplicity is indispensable to science. The point is conceded, and simplification has been described above as often an achievement. What deserves resistance is the move from the value of economical representation to the acceptance of unsupported exclusion. A model that omits a variable because the variable has been shown not to matter in the relevant range is economical. A model that omits it because including it would be inconvenient has merely made a convenient assumption, and the two can look identical on the page. The difference lies in whether the omission was earned.

The second objection is that mathematical laws genuinely explain, and that the essay's emphasis on interpretation understates this. Laws do explain, and nothing above denies it. Their explanatory force depends on interpretation, scope, and empirical support. A law that unifies a wide range of phenomena under a small number of principles, and that has survived tests in circumstances where it could have

failed, has explanatory credentials that an arbitrary fitted formula lacks. The difference between them is the very difference between a relation that has been shown to hold up under pressure and one that merely describes what happened. This position is compatible with the high standing of the great laws of physics, and it supplies a reason for that standing.

A third objection is that the distinction between satisfaction and adequacy is itself too tidy. Feelings of understanding can be trained by expertise and may track genuine insight, a point that is easy to overlook in a discussion focused on their failures. The experienced clinician's sense that a presentation does not fit, or the mathematician's sense that a proof is missing something, are feelings of understanding and non-understanding that often prove correct, and an account that dismissed them would discard a valuable source of evidence. Debates over whether the sense of understanding is a reliable guide to explanatory success have been a persistent feature of the literature [14, 3], and they need not be settled here. What deserves criticism is the unexamined authority of the feeling, and its evidential usefulness need not be denied altogether. A trained feeling is a hypothesis about the state of one's own comprehension, and like other hypotheses it deserves testing against outcomes. Training can make the feeling track adequacy, and one route is that expertise recalibrates it against outcomes: each diagnosis that is confirmed or overturned adjusts what a sense of fit feels like, so that the feeling becomes a compressed summary of past discriminations between accounts that did and did not survive. The same account explains why the feeling is less reliable where no such feedback has occurred, as in the everyday mechanisms studied by Rozenblit and Keil, and it implies that a trained feeling is trustworthy only within the range of cases over which the recalibration took place.

A fourth objection is that the continuation condition is interventionism under another name. It is a close relative, and closer still to contrastive accounts of explanation in the tradition of Lipton. Woodward's account makes explanation a matter of information about what would happen under interventions, and Pearl's formalism supplies the machinery for the causal case, and both are presupposed above. What the present account adds is a different object of assessment and a wider range: it evaluates the act of accepting an account and not only the content of the account, and it applies to non-causal as well as causal cases. The object is the act of acceptance, the replacement of an inquiry's history by a state, so that the felt closure of explanation is itself something the condition evaluates. The range is wider because the same condition applies without modification to familiarity, to simplification, and to formal derivation, none of which is a causal claim, and because it treats the revision of accounts as refinement of one partition. A fifth objection is that question-relativity makes adequacy cheap, since any account is adequate for some sufficiently narrow question. That is correct, and it locates the usual failure. Accounts are rarely adopted for a question they cannot answer. They are adopted for a narrow question and then relied on for a wider one, as when the explanation that suffices for why a mortar sample failed is carried over to why this wall cracked here. Silent enlargement of the question is the characteristic way in which a compression comes to bear more weight than it was formed to carry.

Finally, the thesis applies to itself. "People accept explanations because explanations reassure them" could become the very kind of pleasing generalization the essay criticizes. It accounts for a great deal, it is easy to state, it casts the speaker as one who sees through the comforts that others accept, and it offers its own form of relief. It is therefore best treated as a partial account, tested against curiosity, intellectual discipline, discomfort-seeking, and the willingness to abandon familiar explanations. People

do abandon beloved accounts when evidence demands, and they do pursue questions that promise only trouble. A theory of explanatory relief that could not accommodate these facts would be failing in the way it describes, and the fair summary is that relief is one factor among several in the acceptance of explanations, and a factor that is easy to underestimate precisely because it is pleasant. The continuation condition is likewise a compression of a varied practice, and its own collisions should be sought among cases in which acceptance seems rational although no continuation has been examined, such as explanations of past events that admit no further intervention. There the condition must be read as a constraint on counterfactual and comparative evidence, and how far that reading extends is left open.

15. Conclusion: What Relief Can and Cannot Tell Us

Return, finally, to the opening moment when an explanation seems complete. The relief is evidence that our relation to an event has changed. We can describe it, place it, perhaps calculate it, and decide what to do next. These are genuine gains, and an inquirer who lacked them would be worse off. Whether our account identifies the event's causes depends on further work, of the kind sketched in the preceding sections: asking what alternatives the account excludes, what it has left out and at what cost, which direction its dependencies run, and what would happen if the conditions under which it was formed were to change.

The essay has distinguished several legitimate achievements, and it should end by preserving them. Familiarity enables recognition, without which inquiry could not start. Simplicity makes patterns tractable and allows reasoning to proceed at a scale a human mind can manage. Formulas express relationships with a precision that language cannot match. Proofs establish consequences with a security that nothing else in human knowledge rivals. The value of each becomes clearer when it is asked to bear the evidential weight it can actually support, and not the weight that its attendant feeling of completion suggests. Recognition supports orientation, simplification supports tractability, calculation supports prediction within a model, and demonstration supports conditional necessity. Causal adequacy is a further achievement that draws on all of these and is reducible to none of them.

In the terms developed here, relief often marks the moment at which a classification of situations was adopted, and continuation-sufficiency is a necessary condition on what that classification may discard: no pair of situations that the account identifies can be separated by an admissible inquiry bearing on the question asked. Explanatory relevance, empirical warrant, and intelligibility remain further obligations that the condition does not discharge.

An explanation earns our confidence when it helps us discover where our confidence was premature.

References

[1] Bromberger, S. (1966). Why-questions. In R. Colodny (Ed.), *Mind and Cosmos: Essays in Contemporary Science and Philosophy* (pp. 86–111). University of Pittsburgh Press.

[2] Duhem, P. (1906). *La théorie physique: son objet, sa structure*. Chevalier et Rivière.

[3] Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, 8(1), 101–118.

[4] Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57(2), 243–259.

[5] Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press.

[6] Lipton, P. (2004). *Inference to the Best Explanation* (2nd ed.). Routledge.

[7] Nietzsche, F. (1882/1887). *The Gay Science*, Book V, §355. Trans. W. Kaufmann, Vintage, 1974.

[8] Nietzsche, F. (1887). *On the Genealogy of Morality*, First Essay, §13. Trans. M. Clark & A. Swensen, Hackett, 1998.

[9] Nietzsche, F. (1889). *Twilight of the Idols*, "The Four Great Errors." Trans. R. J. Hollingdale, Penguin, 1968.

[10] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

[11] Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 10, pp. 173–220). Academic Press.

[12] Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521–562.

[13] Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

[14] Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, 69(2), 212–233.

[15] Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.