What Makes a Theory Good? A Reconnaissance of Explanatory Virtue

What makes a theory good? In his canonical 1991 book *Inference to the Best Explanation*, Peter Lipton attempts to answer this fraught question. The philosopher identifies eleven explanatory virtues that are often placed within four groupings: evidential, coherential, aesthetic, and diachronic. Two others, James Beebe[1] and Kenneth Dillon[2], draw upon the same categorical schema to present four other virtues for consideration. All fifteen are listed and defined in the following table:

Virtue	**Definition**
Simplicity	Explains the same facts as rivals, but with less content. Exhibits ontological parsimony in that it lacks ad hoc hypotheses.
Elegance	Exhibits syntactic simplicity. Has a small number of concise basic principles.
Precision	Explains phenomena exactly and accurately.
Predictive Power	Predicts observed outcomes.
Testability	Falsifiable.
Fruitfulness	Generates additional discovery by disclosing new phenomena or revealing new relationships among known phenomena.
Wide Scope	Explains a great variety of facts.
Specification of Mechanism	If P occurs because of Q, specifies the mechanism linking P to Q.
Unification of Existing Facts and Theories	Explains more kinds of facts than rivals with the same amount of theoretical content. Fits with other phenomena to form an integrated whole.
Fit with Existing Background Beliefs	Coherent with other existing knowledge.
Variety in the Sources of Evidence	Does not rely on one piece or origin of evidence.
Depth/Illumination	Provides a depth of analysis related to causal history and passes counterfactual tests.
Intrinsic Plausibility	Intuitively plausible.
Resolution of Anomalies	Explains regularities that aren't explained by existing theories.

[1] Beebe identifies depth and intrinsic plausibility.
[2] Dillon contributes resolution of anomalies and coherence.

Coherence	Internally consistent.

Worthy of a second thought are the criticisms corresponding to the absence of each explanatory virtue. I put forth a few:

a. Insufficiently accounts for the phenomenon or context it's embedded in
   i. Wide scope
   ii. Specification
   iii. Unification
   iv. Fit
   v. Resolution of anomalies

b. Overcommunicates
   i. Simplicity
   ii. Elegance

c. Logically inconsistent (in relation to its own hypotheses or in relation to the realities of the phenomenon)
   i. Specification
   ii. Unification
   iii. Fit
   iv. Coherence

d. Undercommunicates
   i. Precision

e. Ineligible for evaluation within the scientific domain
   i. Testability

f. **Excessively narrow or under-elaborated**
   i. Fruitfulness
   ii. Wide scope
   iii. Variety
   iv. Depth/illumination

In a theory, what properties of language may be proxies for inferiority?

g. Excessive information
   i. More information provided -> more potential for error

h. Ambiguity
   i. Ambiguous language -> more potential for misrepresentation
   ii. Misrepresentation of truth ≈ falsity

i. Usage of uncommon vocabulary
   i. Commonly: sophistry is occurring and advanced vocabulary is used to obscure or compensate for content inadequacy

How might one go about ranking these virtues, if such a thing is possible? First I maintain that no definition or evaluative system is ontologically privileged. Yet, it seems that one can still rank each virtue as indexed by a preselected, intersubjective definition and value hierarchy. In the first section I'll engage in some pedantry and methodically take this project on. In the second, speculative happy hour will usher in open-ended questions.

Virtue Ranking: An Astringent and Spockish First Pass

At risk of veering into tautological territory, I'll assume that explanatory power is the most fundamental element of a theory with explanatory virtue. Under this assumption, the explanatory virtue of a theory is disproportionately degraded by the absence of explanatory virtue relative to any other possible missing element. In fact, it can be said that a theory without explanatory power cannot have explanatory virtue unless a preponderance of other virtue-contributing but secondary factors are present in number and/or degree to compensate. Finally, the presence of explanatory power disproportionately increases the explanatory virtue of a theory relative to other factors. Ceteris paribus, the evaluative model governing the assignment of explanatory virtue calls for the maximization of explanatory power. However, even when all else is not equal, the promotion of explanatory power is called for to a greater extent than the promotion of any other factor in a theory with explanatory virtue.

As we must rely on intersubjective definitions here, consensus will guide us in establishing a working definition of explanatory power. It is my belief that all definitions of power are teleological, in other words purposive. There is no power in abstract, but rather power *in relation to* desiderata. By current cultural consensus, the telos of a theory with explanatory power is to clarify the nature of the phenomena established within its purview.

When the nature of a phenomenon is clarified, it is both 1) accounted for in a way that doesn't fail our best empirical tests and 2) made comprehensible. It is important to note that I consider our rational faculties to be one of our empirical tools, and logical evaluations (along with others relying on the five senses) to be empirical tests. I also believe it to be inconclusive that our empirical tests are sufficient to access objective truth, and thus to say a phenomenon is clarified because an associated theory fails no empirical tests is not to say it is settled. Clarification can only be sought for apparent truth, which is useful to many ends but not representative of complete truth.

Some may claim that only #1 is essential to explanatory power, but I disagree. Those who hold this view may be confusing explanation, which I believe to be a form of communication, with cause-effect linkage, which is independent of human faculties.

Being both a form of interpersonal communication and a form of depersonalized linkage between theory and phenomena, an explanation has two elements: its relation to its subject (the phenomena) and its relation to its recipient. Considerations of comprehensibility are not relevant to the subject relation, but they are relevant to the recipient relation. Thus, there are two separate modes of consideration as regards a theory's explanatory power: considerations pertaining to the theory itself and considerations pertaining to the theory's capacity to contribute to understanding.

In summary, a theory with explanatory virtue, and thus explanatory power, is clarificatory in terms of: 1) the apparent accuracy and comprehensiveness of the cause-effect relationship it identifies in its subject 2) its ease of comprehension. The aforementioned explanatory virtues will be evaluated along these two

dimensions, with a weighting towards the first. A theory that succeeds at accounting for the phenomena within its purview cannot be deemed an explanation if it fails to be comprehensible. However, it is arguable that it is not possible for a human being to generate a successful account-generating theory that is 100% incomprehensible. It is, however, possible for one to generate a comprehensible theory that is 100% inaccurate. A theory that is malformed but easy to understand renders itself ineligible to have explanatory power, and thus fails to have the foundation of explanatory virtue by my definition. On the other hand, an arcane theory that accounts for its subject can still acquire explanatory power at any time as soon as it is formatted for simplicity, if it does not already have this already by virtue of being human-generated. For this reason, I'll assume that a theory's cause-effect account is more essential to its explanatory virtue than its comprehensibility.

It is not the case that accuracy and comprehensiveness are roughly twice as important as comprehensibility, and that this ratio holds for all related properties. There are a number of more precise scenario-based weightings that can be assigned based on contingencies and interactions. A theory extremely flawed in any given category, even one that would normally be underweighted, may become untenable. A "fit test" may be warranted for each theory, creating relational dynamics between the virtues that differ from case to case and ratio to ratio. It is difficult to establish a firm relative value or exchange rate, so to speak, between the metrics. Furthermore, it cannot be said that these are the only two metrics factoring into explanatory virtue. Nevertheless, it is still true that generally, factors pertaining to the integrity of a theory's account are more significant relative to factors concerning comprehensibility.

Another thing to mention:
The underweighting of virtues on the grounds of redundancy works in a closed system of fifteen preselected virtues, but it doesn't necessarily represent the value of these virtues in their own right. Both relative and objective weightings can be derived. Discounted weighting based on contingent properties of the virtues should be less than discounted weighting based on fundamental properties of the virtues themselves. Furthermore, inadequacies that relate to intersubjective properties may need to be discounted more than inadequacies that relate to subjective ones. This has implications for the virtues that represent and buffer against their absence.

Important to ask is: what flaws does this virtue buffer against, and what flaws does it not? How much more vulnerable is a theory if this virtue is not manifested? If it is a lot, this indicates that the virtue is especially important relative to others. However, because this vulnerability itself may be partially evaluated based on the same virtues, there's a closed-system risk at play here that is worth noting.

Below I'll rank the fifteen explanatory virtues based on what I've defined to be explanatory power. I'll also assign them notional weightings with a base reference range from -1 to 1, with the disclaimer that these weightings are to a great extent arbitrary. If it can be objectified at all, the proposition that the explanatory value of a theory can be represented by a continuous variable (rather than say, a categorical one) is not certain. However, if one virtue contributes more to the explanatory value of a theory than another, a natural conclusion is that this value can *have* a value. Why might coherence deserve .5 and not .51? I don't have any good answer, so the subjectivity of this exercise should not be misgauged. However, the following is my attempt to undertake it anyway:

Coherence

Ranking: 1
Notional Weight: N/A, as categorical (if absent, theory is not complete. If present, theory is complete)
Internally consistent.
Internal contradictions destabilize a theory, rendering it an incomplete prototype that is unready to be evaluated for its explanatory power vis a vis what is outside of itself.

Testability
Falsifiable.
Ranking: 2
Notional Weight: N/A, as categorical (if absent, theory has no explanatory power until future. If present, theory is eligible to have explanatory power)
Testability retains weighting as a precondition for the evaluation of a theory's *apparent* accuracy and comprehensiveness. It cannot be known (apparent) that a theory explains phenomena if the mechanism it posits is imperceptible or unverifiable. However, it must be stressed that this problem is epistemic rather than ontological. It is, in fact, the reason that I draw a distinction between explanations and cause-effect linkages. There may be theories with great potential explanatory power that are at present untestable. These theories may even identify objective cause-effect linkages, though this cannot ever be known. In the future, such theories become testable, revealing their explanatory merit, or they may fail to do so while retaining their unverifiable truth. Because this latent value is present independently of current tools to uncover it, untestability loses some sway on explanatory power if temporary, but retains it if necessary and/or permanent. **It is important to note that testability's influence on apparent *explanatory* power does not necessitate a devaluation or discontinuation of untestable theorizing at large. The formulation of untestable theories may be valuable as an end in itself or have other merits and benefits unattached from demonstrability.**

Predictive Power
Predicts observed outcomes.
Ranking: 3
Notional Weight: 1
Predictive power retains its weight as a basis of a scrutiny-withstanding comprehensive and accurate account. A theory that offers this accounts for the underlying mechanisms of a phenomenon by being generalizable in relation to it and thus having predictive power over its future occurrences.

Depth/Illumination
Ranking: 4
Notional Weight: 0.8
Provides a depth of analysis related to causal history and passes counterfactual tests.
In the closed system, depth is somewhat mirrored in specification of mechanism and thus, precision. However, it is still true that the mechanism of causation is the basis of an explanation.
Additionally, failing counterfactual tests reveals contradictions, even one of which can invalidate the accuracy of an account and render it untenable.

Precision
Explains phenomena exactly and accurately.
Ranking: 5
Notional Weight: 0.7
Precision loses some weighting insofar as inexact explanation contributes to incomprehensibility alone, but maintains weighting insofar as it jeopardizes accuracy. Inexact explanation has the potential to

create a discrepancy between posited and conveyed ideas, altering the explanatory power of a theory by error in its manifest form. Precision then gains significant weighting given that accurate explanation is not merely a direct input into accurate explanation like some other virtues, but in fact equivalent to it. While accuracy seems to a large degree exempt from the problem of arbitrariness, less so for exactitude.

Unification of Existing Facts and Theories
Explains more kinds of facts than rivals with the same amount of theoretical content. Fits with other phenomena to form an integrated whole.
Ranking: 6
Notional Weight: 0.5
Within the closed system, the fit dimension of unification is encapsulated within fruitfulness. Unification loses some weighting for relying on categories, which fall victim to the problem of arbitrariness and make consistent evaluation difficult if not impossible. That said, the same obstacle presents itself everywhere that definitions are required. In a way, it's arbitrary to give it excess influence here. Unification gains some weighting insofar as fit tests buffer against contradiction, an enemy of apparent accuracy and comprehensiveness. It is also a unique virtue in that it has a mechanism not only to validate new theories but also to identify superfluous content in existing ones.

Simplicity
Explains the same facts as rivals, but with less content. Exhibits ontological parsimony in that it lacks ad hoc hypotheses.
Ranking: 7
Notional Weight: 0.3
Insofar as it is a virtue of comprehensibility, simplicity is relatively underweighted. However, insofar as Occam's razor mediates the probability of a theory's accuracy, simplicity remains an input to the most essential dimension of explanatory power. It is not a scientific fact that a simpler explanation is more likely to be true or have explanatory power. However, a simpler theory may give a rational person more of a reason to believe it. Popper's falsifiability criterion is more amenable to simpler theories, which are comparatively unburdened. A theory with fewer hypotheses has less content that is vulnerable to falsification, and thus less content that can render it inaccurate, damaging its explanatory power and virtue. Because simplicity is a buffer against inaccuracy, which is a destructor of explanatory power, it is a relatively important aspect of explanatory virtue. Lack of simplicity often contributes to inaccuracy through the inclusion of inaccurate hypotheses. Still, it remains a proxy for factors contributing to poor explanatory power rather than a direct contributor.

Fit with Existing Background Beliefs
Coherent with other existing knowledge.
Ranking: 8
Notional Weight: 0
In the closed system, fit has somewhat of a Russian nesting doll relationship with unification and fruitfulness. However, it retains some weighting due to the nuances between a fit in abstract and a fit that "forms an integrated whole". The latter is twice as vulnerable to the problem of arbitrariness as the former, though both indeed are.

Specification of Mechanism
If P occurs because of Q, specifies the mechanism linking P to Q.
Ranking: 9

Notional Weight: 0
This virtue loses some weighting for being a virtue of comprehensibility, and is to some degree redundant within a closed virtue system featuring precision. If an account, which encompasses both P and Q, is precise, it specifies the mechanism linking the two. Closed system considerations notwithstanding, insufficient specification can similarly render an account vulnerable to inaccuracy. However, while precision arguably mediates the accuracy of an account alone, specification may also mediate comprehensiveness depending on the relational elements and scope of P and Q.

Wide Scope
Explains a great variety of facts.
Ranking: 10
Notional Weight: -0.1
There are two ways for a theory to have high explanatory power through the account it provides. It can explain a few facts to a great extent, or explain many facts to some extent. Through the second channel, this virtue is a direct input to the comprehensiveness of an account, however the determination of a "great variety" is subject to the problem of arbitrariness.

Variety in Sources of Evidence
Does not rely on one piece or origin of evidence.
Ranking: 11
Notional Weight: -0.2
A theory may have high comprehensibility, comprehensiveness, and accuracy while relying on only one piece of evidence. Variety in evidence is not necessary or sufficient toward those ends. One piece of evidence can render all understanding void. However, insufficient or inaccurate evidence may lead to inaccurate principles, assumptions, predictions, posited mechanisms, and so forth, and thus inaccurate theories. Data collection impacts all aspects of data interpretation, and thus theory crafting.

I put forth another possible Occam parallel: ceteris paribus, the more facts filter into a hypothesis, the more there is more reason to believe it. Variety may be a buffer against inaccuracy in many cases due to its impact on hypothesis formation. If a hypothesis relies on one piece of evidence only, the reinterpretation of that evidence can render it baseless. By diversifying evidence, one protects their hypotheses from being immediately robbed of all legitimacy. Variety may help defend contributors to explanatory power from being instantly undermined, but it is not a requirement for truth apparent or otherwise.

Elegance
Exhibits syntactic simplicity. Has a small number of concise basic principles.
Ranking: 12
Notional Weight: -0.3
Not all elegance is created equal. The syntactic simplicity dimension of elegance relates to comprehensibility and beauty. The basic principles element, on the other hand, relates to accuracy to the extent that ontological parsimony does. Both are vulnerable to the problem of arbitrariness. Excessive principles run into the aforementioned threats posed by the falsifiability criterion and Occam's razor. Elegance loses some weighting insofar as its element of conciseness relates more to relatively inessential comprehensibility than relatively essential accuracy. It loses additional weighting by failing to impact explanatory power in a way that isn't already accounted for by the simplicity category, which itself only mediates explanatory virtue through proxies. This is only true, however, in the closed system

of these fifteen preselected virtues. When elegance, like anything on the list, is evaluated either in itself or in relation to different virtues, a new calculation is in order.

Resolution of Anomalies
Explains regularities that aren't explained by existing theories.
Ranking: 13
Notional Weight: -0.4
This is a direct input to comprehensiveness, although a theory that resolves anomalies demonstrates the potential inadequacy of others more-so than its own merit.

Intrinsic Plausibility
Intuitively plausible.
Ranking: 14
Notional Weight: -0.8
Intrinsic plausibility is a paradoxical notion thus unfulfillable as a condition. All plausibility is perception or intuition-based and thus extrinsic. If we establish extrinsic plausibility as a separate virtue, it can regain some weighting. This is because there is possibility that intuitive reactions reflect subconscious processing mechanisms which can accurately evaluate explanatory merit. Separately, a theory eliciting strong feelings of implausibility can be considered incomprehensible if it shuts off the desire for understanding.

Fruitfulness
Generates additional discovery by disclosing new phenomena or revealing new relationships among known phenomena.
Ranking: 15
Notional Weight: -1
Fruitfulness loses some weighting by failing to directly relate to either one of the key inputs to explanatory power. As a possible counterpart to Occam's razor, a theory that is compatible with more true claims may be more likely to provide a true account. In this way, fruitfulness at first seems to factor into apparent accuracy as a proxy for proxies of accuracy. However, to reveal new relationships among known phenomena is not to demonstrate them.

In the following sections I'll transition from the indexed to the ambiguous, presenting some speculations and questions for further debate.

Getting Tipsy on Conjecture

A theory with explanatory virtue promotes understanding. But what even is that? What in particular must an "understander" walk away possessing? One interpretation is diminished curiosity: understanding is the ability to truthfully profess, if X is the phenomenon in question, "I know why X is the way it is." Now, why might a given X be the way it is? If Y is an important related fact, there seems to be two main possibilities:

(1) Because Y is the case, in which Y is a static property of X
(2) Because Y occurs, in which Y is a changing state of affairs related to X

These can further be separated into:

[A] Ongoing states of affairs in the present
[B] Completed states of affairs in the past

Arguably, all possible future occurrences are represented within present states of affairs, and thus accounted for within category A. Though they may be governed by static properties, these properties must have analogues in changing processes for future change to occur through them.

Are there different metrics that factor into explanatory power based on whether theories posit static properties or changing processes?

What are the conditions that a theory T must satisfy to show that a property Y(1) is indeed a property of phenomenon X? This relies on the definition of this property itself. Unexpectedly, the falsifiability criterion comes back into play. We find that all definitions (of the form “an A is B”) are unfalsifiable. Might this have broad-reaching epistemic implications?

Relatedly, what are the conditions T must satisfy to show that a process Y(2) occurs to produce X? For each occurrence posited to be part of a broader process, one must prove that it is linked to the others in a way that extends past coincidence. Thus, each occurrence that is tied to a hypothesized process creates an additional burden. The twin menaces of Occam’s razor and Popper’s criterion once again present themselves here. If each possible process element is a hypothesis, and each of those can be broken down nearly infinitely into other subprocesses, does the kaleidoscope of hypotheses get conceptually unmanageable?

Comprehensibility may have two dimensions:

1) The communicative dimension (“I understand what is being communicated”)
2) The aesthetic dimension (“The communication is presented in a way that is pleasing to me”)

When something is communicated in a way that is pleasing, it is easier to receive, and therefore perhaps easier to later comprehend. Comprehension may to some extent be a voluntary process that is facilitated by appeals to aesthetic faculties. Is beauty not given its due in discussions of the search for truth?

Relatedly, a theory or system may have aesthetic value that can be evaluated independently of its ability to be truth-tracking. It may make unfalsifiable claims about entities/processes that cannot be proven to exist precisely as described, but are understandable as archetypes. One example is the sixteen Jungian cognitive types. Most psychological theories in general, and all involving subconscious entities without observable correlates, fall into this category.

A Final Word

I've presented a ranking of fifteen explanatory virtues based on a provisional consensus framework of explanatory power. Though many more virtues and rankings can be identified, this project has yielded one possible hierarchy for those concerned with evaluating theories by accuracy, comprehensiveness, and comprehensibility. Ontological questions remain about these metrics themselves as well as the concepts they rely on and the contexts they're embedded in. Epistemic perimeters circumscribe the debate as a whole. Though it's possible to discuss the nature of what we seek in theories and link this to assumptions about what those theories should contain, the discussion remains asterisked and the topic largely unsettled. Its continuation if nothing else is aesthetically rewarding.