Scientific Explanation: a predictive framework

Beatriz Borges

School of Computer and Communication Sciences

Project SHS, 1st year master

Professor: Christian Sachse

Co-supervisor: Frida Trotter

Philosophy of life sciences

Submission date: June 3, 2020

Lausanne, academic year 2019-2020

EPFL

Contents

1 Introduction **2**

2 General notions **3**

3 Predictors **5**

3.1 Introduction 5

3.2 Domain of applicability 7

3.3 Construction 8

3.4 Evaluation 9

3.4.1 General metrics 9

3.4.2 Uncertainty 12

4 Explanatory adequacy **13**

4.1 Comparing the adequacy of different predictors 14

4.2 Comparing Newtonian Gravity and General Relativity 15

5 Conclusion **16**

1 Introduction

It is common, in the scientific community, for there to be multiple theories that address the same phenomena. For example, there were two famous contrasting theories that attempted to explain Mercury's perihelion – Newton's classical mechanics and Einstein's relativity (the latter having superseded and falsified the former). In this essay, the explanatory power of scientific theories will be analyzed, looking to establish a framework that ultimately allows their ranking in explanation adequacy.

The criteria for comparing two scientific theories will first be discussed, arguing the need of incorporating probabilities and prediction in such a task. This will lead to the construction of a philosophical framework for contrasting scientific explanations by comparing their predictive power, and to analyze the consequences of such an approach. This framework will be referred to as the *Predictor Framework*, within which a model for explanation is delineated. Finally, the examples of Newtonian Gravity and General Relativity will be studied in light of this newly developed system.

The focus of this essay will lay with the need of predictive mechanisms for comparing scientific theories. When dealing with predictions, probabilities are often being dealt with as well. Indeed, one of the major arguments for this need can be established by the fact that some theories of scientific explanation cannot escape the need for a probabilistic approach before they have even been established. This gives rise to several questions:

1. *Are phenomena which cannot be physically deconstructed as explainable as the ones that can be physically taken apart? (i.e. understanding the pressure in a gas container vs. understanding a clock's mechanisms)*

   In the first case, it is impossible to account for every individual particle's location and velocity, and only through inclusion of probabilities can sensible explanations for the behaviour of the whole be constructed.[1]

2. *Can non-exact phenomena (to the best of our current scientific understanding), like atom decay, be explained? (Salmon 1990, pp. 8, 15-18)*

   Once again, the explanations that can be constructed for this type of phenomena are irreducibly statistical.

3. *And finally, by design, probabilities directly address uncertainty. Thus the question*

[1] For the specific case of thermodynamics, this is due to the fact that for each *macrostate* (like pressure, volume, temperature, among others), there are many *microstates* (possible particles' configurations, through their position and velocity) which constitute the same *macrospace*. For example, interchanging the velocity (but not position) of two gas particles does not alter the *macrostate* though the *microstate* is different.

of whether an imperfect theory or prediction can ever truly offer an explanation arises.

This last topic will warrant our attention throughout the third part of this essay. The relationship and boundaries between explanation and prediction will first be explored. Then, the focus will lie on the analysis of two different predictors' explanation's quality and validity. Concretely, as mentioned above, the comparison between the explanations Newtonian Physics and General Relativity can offer will be focused on.

To tackle the topic of predictors – including their handling of different categories of scientific explanation, and of uncertainty and probabilities – this essay will start by briefly discussing the concepts of explanation, prediction, and causality. This essay will then work towards the definition of a new prediction-based model for explanation. To achieve this, predictors will be presented in detail – from how they allow the comparison of theories from different categories of explanatory models, to their definition, main components, type and construction. After tackling these concepts, it becomes possible to define the *Predictor Framework* – which explores both the predictors and their predictions, as well as both metrics and a procedure for their evaluation and comparison. Finally, drawing from the previous section, the explanatory adequacy of predictors will be further examined, as well as the criteria for preferring one predictor over another. Throughout the essay, the aforementioned example of the classical and relativistic gravity theories will referred to, in order to illustrate in deeper detail some aspects of predictors and their comparisons.

2 General notions

Can there be **explanation** in total randomness? If a system behaves truly randomly – without any causal interactions between its components, itself, the outer world and past behaviors exhibited by itself – can it be explained, or just merely described? Are there underlying principles it follows, like the statistical regression to the mean, or causal interactions and relationships between the system, its components, and possibly even the outside world? If a system behaves in a purely random fashion, then it cannot, by definition, follow any such principles, and thus it cannot be accurately predicted nor explained in any manner that entails deeper understanding of its inner workings. Indeed, only that which is immediately observable can be understood – that this is a system whose behavior can only be described as one which is dictated by randomness and probabilities. This simple thought experiment highlights an important point – the importance of **prediction** in explanation, which will later be explored, as well as the distinction between **description** and explanation, which will now be discussed.

Indeed, it is not enough to be able to know the outcome of a given scenario in order to be able to explain it. Instead, true understanding is required to provide an explanation instead of simply a descriptive account – for example, knowing *herd immunity*[2] will prevent a disease's diffusion does not confer any kind of explanation as towards why that is the case. (Salmon 1982, pp. 4-8, 11) It is also worthy of mention that different models of explanation have differing compromises regarding the distinction between the explanatory and merely descriptive accounts.

Also highlighted in the above thought experiment, is the need for **causality** in the system's inner workings. Put differently, the system must have some kind of causal structure, whether in regards to its components, external influences, its own history or past behaviors, or any combination of these factors.

Otherwise, if there are no causal links that, even if not certain, favour some outcomes over others, then the system can only be observed. If there is no such causality to the system, then no prediction can be made with even the smallest amount of certainty, because the system is composed of irreducible uncertainty. However, when there are some causal (even if probabilistic) relationships to the system's behavior, then those reduce the associated uncertainty, and progress towards explaining it can thus start. A connection between causality and explanation can then be established. Consequently, a distinction between them is required. Several possible definitions exist for both concepts. For this essay, causality will be defined as the result of some regularity or law, that links cause to effect. In other words, referencing back to the simple thought experiment, causality can be understood as the influence (that can originate from some components of the system, or even from the outer world) that results in at least some lack of randomness in outcome likelihood, whereby, at the end, some possibilities have higher or lower probabilities of occurring. Explanation will be understood as a record of why a given event or regularity or pattern happened (Woodward 2017, sec. 1).

Focusing back on the relationship between explanation and prediction, it will also be argued that the lack of predictive power implies lack of adequate explanation, as it becomes simply a past case of chance correlation. To make use of Hempel's (Hempel 1965, pp. 335-344) own examples, given two generalizations, *'all members of the Greensbury School Board for 1964 are bald'* and *'all gases expand when heated under constant pressure'*, the first is only accidentally true, while the second is a law of nature, characterized as a 'non-accidental generalization' (Woodward 2017, sec. 2.2). While this distinction is intuitive, it is also very hard to formalize traditionally – but under the *Predictor Framework*, formally introduced in section 3.1, by recurring to the predictive power of theories,

[2]Herd immunity is the phenomenon by which individuals who have not developed immunity from a given disease are protected through the immunization of a critical mass of the community.

it becomes easy to systematize it.

3 Predictors

3.1 Introduction

A **predictor** is any kind of system, which, given pertinent information about an environment (henceforth referred to as **features**), can make a prediction concerning an event or pattern. This event or pattern is the **explanandum**, that is, the phenomenon to be explained. The predictor constitutes the **explanans**.

It is worthy of note that a predictor can use multiple processes to generate a prediction. In the philosophy of scientific explanation, the five most popular categories of explanation are the following:[3] (Woodward 2017, sec. 2.1-.2, 3.1, 3.3-.4, 4.1-.2, 5.1-.3) (Douglas 2009, pp. 4-10, 12-14) (Kitcher 1989, pp. 2-4, 10-12)

1. **the covering law model** – which deduces from the context a set of regularities – laws – that can explain the events that occurred in the aforementioned context; this model however, risks not separating correlation from causation, and admitting chance regularities to be seen as explanatory when they are not, like the *'all members of the Greensbury School Board for 1964 are bald'* example presented above,

2. **the statistical relevance model** – which states that statistically relevant properties or relationships – statistically relevant here meaning they have an impact on the explanandum[4] – have explanatory power, unlike those which are not statistically relevant; however, it is difficult to provide a causal account of explanation using statistical relevance as a basis (Hitchcock 1995, pp. 6-10) – like in the case of the covering law model, separating correlation from causation becomes very difficult, reason why Salmon himself moved on to develop the causal mechanical model,[5]

[3] Though it should be mentioned those five categories do not encompass all the proposed theories of explanation, like Van Fraassen's and Achinstein's pragmatic approaches, which will not be discussed at the same depth since predictors will be evaluated from a normative, objective perspective.

[4] More concretely, if an outcome's probability differs, in a statistically significant way, from itself when conditioned on the value of a given property of the model, then this property is said to have an impact on the explanandum, and be statistically relevant, explanation-wise.

[5] As a counterexample, it is possible to show that statistical relationships are not necessarily at all connected to causal ones. Consider, for example, that in the US there is a strong correlation between high mortality rate and Social Security benefits (Rosenbaum 2019, pp. 70-71). In reality, the latter obviously poses no threat to public health. An inhabitant will simply receive more benefits when they are retired, older citizens, who have a much higher risk of death than the average adult. There is a hidden confounder in this model, the citizen's age and general health condition, which impacts both Social Security benefits and mortality rate. As such, despite the strong statistical correlation between Social Security benefits

3. **the causal model** – which employs causal relations as explanatory ones, creating a process composed of a chain of causal inferences; however, as Lipton (Lipton et al. 2004, pp. 30-33) points out, "a causal model of explanation cannot be complete", since there are several scientific explanations which cannot have a causal character, but, for example, follow statistical tendencies instead,[6]

4. **the unification model** – which joins together disparate theories and phenomena under a single banner, allowing for the emergence of explanation and understanding; however, this model appears to not truly explain singular events, where it is often needed to look at specific causes rather than the unifying law (Lipton et al. 2004, pp. 28), or even offer no explanation at all, for example, combining various instances of metals expanding when heated into the more general rule of 'metals expand when heated' offers no explanation as to why that is,

5. and **the mechanism model** (Machamer, Darden, and Craver 2000, pp. 3-8) – which allows the emergence of functional patterns in a given context, and the subsequent inference of the outcome in similar contexts; however, the authors themselves admit this model was solely designed to provide insight for specific scientific areas, with no claim of being a universal model.

In his 2009 paper, Douglas (Douglas 2009, pp. 14-16) argues all these explanation model categories can be unified under a predictive umbrella – by viewing these models as tools with which to generate testable (and novel) predictions, one becomes able to validate, refine or reject the model in question. It is assumed that any model that explains a phenomenon can be used to make predictions about it. In this way, predictions act as lenses into the implications of theory being studied.

The fact that it is possible to abstract away from the underlying mechanism of a model, and view it simply as a prediction-generating black box is one of the crucial pillars of the *Predictor Framework*. It allows us to compartmentalize the conceptual differences of the inner workings of each type of model, while simultaneously remaining firmly rooted in a tangible domain due to the comparable predictions generated by them.

And so it is finally possible to arrive at the full definition of a predictor – a predictor acts as wrapper for a given explanatory model, incorporating three aspects:

1. the **theory** or hypothesis itself, which acts as the model's explanatory device,

and mortality, it is obvious that there is no causal relation there, only a statistical one.

[6]One of the examples provided by Lipton is the effectiveness of criticism and inefficacy of praise on air force pilots' performance – they appeared to always improve after criticism, but to always worsen after praise. Regression back to the mean is a better explanation for this phenomenon than the powerful positive and negative impact of criticism and praise, respectively.

2. the **domain of applicability**, that is, the specific domain for which this theory was built or since found to be applicable to[7], and for which it claims to offer pertinent insight,

3. and lastly, the **predictive mechanism** by which it is possible to go from theory (and whichever features it requires) to prediction.

It is worthy of mention that it is possible that the first and third components are merged in a single package. Both Newtonian and Relativistic gravitation theories are examples of this case. Due to their mathematical backbone, given all the relevant input features, they explain not only the impact of each one on the gravitational forces applied to the object being studied, but also predict these forces. A counterexample is, for example, germ theory. This theory aims at explaining the spread of several diseases via microorganisms (pathogens), but does not contain any intrinsic predictive mechanism. Despite this, several can be proposed and adapted from it, including the one of the first used to create this theory – Semmelweis' observation that women died much more often during childbirth if their doctor had not washed their hands before attending them, compared to both a midwife or a doctor that had previously washed their hands. As such, a childbirth survival prediction mechanism could be constructed to verify the impact of hand-washing on the contamination of infectious diseases. More elegant and refined tests and predictions can obviously be constructed as well.

Predictors themselves are one of the key components of the *Predictor Framework*. Having created an abstraction that provides equal footing for the different categories of explanation, only two other major aspects are still needed to complete it – namely, how to construct predictors, and how to evaluate and compare them. These are discussed in detail in section 3.3 and 3.4, respectively. Before breaching those topics, however, the domain of applicability of a predictor will also be presented in larger depth, in section 3.2.

3.2 Domain of applicability

But how exactly are these predictions defined? Let us go back to the *'all members of the Greensbury School Board for 1964 are bald'* example. Assuming this is our predictor for

[7]The domain of applicability contains, then, at least the specific domain that lead to the theory's creation. However, it can also have grown to include more than that original domain, if the theory has later found to be equally acceptable to another domain. One example are Maxwell's equations, which originally comprised the domain of classical electromagnetism, electric circuits, and optics – but which were later incorporated by Albert Einstein into his special relativity, thus expanding the domain of applicability of the original equations.

baldness, then it is possible to generate predictions of the form:

If person p is a member of the Greensbury School Board for 1964, then p is bald.

But this predictor does not, cannot, tell us anything about anyone else in the world, both in domains of time and space. Indeed, its **domain of applicability** is the board of the Greensbury School for 1964. This is, as it is possible to see, an extremely small domain of applicability – implying from the outset the limited potential of this model, if there is any.

In the limit, it is trivial to show that for an arbitrarily small domain, any model can be made correct, generating perfect predictions. But how then can the predictor's (and, by proxy, the model's) worth be measured?

The answer is as far from trivial as it appears. Nevertheless, there are several criteria by which it is possible to judge a model both inside and outside its domain of applicability. A few metrics, which compose the smallest viable subset of tools by which a predictor can be evaluated, will be presented below in the section 3.4.

3.3 Construction

At their essence, predictors can make one of two types of predictions:

1. predictions that are **continuous** and fall in a spectrum (for example, predictions such as:

   (i) *if I throw an object with mass m, at height h and angle a relative to the ground, assuming no air resistance, it will take* ***{continuous prediction}*** *seconds for it to fall to the ground*, or

   (ii) *baldness and stress have a level of correlation of* ***{continuous prediction}****, (on a 0-1 scale)*.

2. predictions that are **discrete** and have a fixed value (for example, predictions such as:

   (i) *by analyzing how unexpected words are in the context of a given news headline in English, it is possible to determine the news type,* ***{discrete prediction}****, from the three possible types of humorous, critical or neutral*, or

   (ii) *a person will become bald if exposed to above-average stress in their daily lives*, which contains an implied **binary** (that is, either yes or no) **discrete prediction**).

It should also be noted any continuous predictor can be made into a discrete one, by grouping its predictive outcome into *bins*, or ranges of possible values for the outcome. For the *1.i* example above, the bins could be defined as $[0, 1[$ (less than 1 second), $[1, 5[$ (at least 1 second, but less than 5 seconds), $[5, \infty[$ (5 seconds or more), for instance.

Having introduced this, it is possible to construct the predictors of the by-now-familiar scientific theories of Newtonian Physics and General Relativity:

1. **Newtonian Physics**

   1.1. **Prediction** – given two bodies with mass m_1 and m_2, they exert upon each other a force of the form $F = G\frac{m_1 m_2}{r^2}$ (where G is Newton's gravitational constant and r the distance between the centers of mass of the two bodies), in the direction of the straight line that connects their centers of gravity, in opposite (but mutually attracting) directions. These forces are exerted instantaneously.

   1.2. **Domain of applicability** – bodies of mass, unless $(v/c)^2$ and ϕ/c^2 are both much less than one (where c is the speed of light (in vacuum), v is the velocity of the bodies of mass, and ϕ is the gravitational potential). (Misner, Thorne, and Wheeler 1973, pp. 1047-1049, 1066-1073)

2. **General Relativity**

   2.1. **Prediction** – a more general theory than Newtonian gravity, general relativity can be described using Einstein's field equations, which in turn can be written in several different ways, such as $R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}$ (Grøn and Hervik 2007, pp. 179-184), which will not be explained here in detail due to their inherent complexity. By using them, as well as the geodesic equations (Weinberg 1972, pp. 67-73), $\frac{d^2x^\mu}{ds^2} + \Gamma^\mu{}_{\alpha\beta}\frac{dx^\alpha}{ds}\frac{dx^\beta}{ds} = 0$, and solving the system by them generated, it is possible to predict how freely-falling matter will move through space-time.

   2.2. **Domain of applicability** – all of the universe, above the quantum scale.[8]

3.4 Evaluation

3.4.1 General metrics

Having defined our predictors, the need to evaluate them arises. Below several metrics are introduced, which constitute a sensible group of tools with which it is possible to evaluate

[8] General Relativity is not compatible with quantum theory. Aggravating this problem is the fact that it is not (currently) possible to run experimental tests so as to test the predictions of possible reconciling theories bridging the two domains. (Schwarz 2007, pp. 1-2, 7-9) (Ashtekar 2008, pp. 1-5, 10, 12)

the predictor.[9]

1. **Criteria applicable to all types of discrete predictors**

   1.1. **domain accuracy** – a measure of the proportion of correctly predicted outcomes, in the predictor's domain of applicability.
   More concretely, for the *2.i* example in section 3.3, it would be the number of correctly classified types of news for the studied corpora. This measure can be skewed when the domain is unbalanced, that is, when one class is much more frequent than the other in terms of observations – in other words, when one class occurs much more frequently than the other.[10]

   1.2. **out-of-domain accuracy** – the same measure as the previous one, but applied to domains that are at least partially outside (meaning they can contain) the domain originally defined by the model. In a sense, it can be interpreted as the generalization accuracy.
   Using the *2.i* example again, these out-domains can be news headlines that are not in English, or sentences that are not news headlines, for example.

2. **Criteria applicable to binary predictors** [11]

   Henceforth, the convention that one of the classes is the *yes* class, and the other the *no* class will be adopted. Which is which is arbitrary for the purposes of presenting these metrics, given the decision is consistent throughout their application.

   2.1. **precision** – a measure of the proportion of predictions that were relevant, that is, a measure of correct predictions for the *yes* class, out of all *yes* class predictions (which include the correct predictions and the incorrect *false positives*).[12]
   More concretely, for the *2.ii* example in section 3.3, it would correspond to the number of correctly predicted cases of baldness, not only out of those correct

[9] It should be emphasized the metrics presented constitute but a sample of the possibilities, many more metrics can be developed for more specialized model evaluations. Here, the more general metrics were focused on.

[10] For example, if the odds of having cancer for any given person are only 1%, the predictor that always predicts no one ever has cancer will have 99% accuracy.

[11] It should be noted that, technically, any predictor can be decomposed into a set of binary predictors. This is achieved first by discretizing them into a predictor of an arbitrary number of N possible classes, and then by turning a discrete predictor into $N-1$ binary ones, of the type *input is in class n, or input is not in class n* (where, by exclusion of parts, if the input does not belong to any of the first $N-1$ classes, then it can only belong to the last, N-th one, thus exhausting all the possibilities).

[12] A false positive is a prediction that should have belonged to the *no* class, but was incorrectly classified (hence the false) as belonging to the *yes* class (hence the positive). Following the same logic, a false negative is a prediction that should have been classified as *yes*, but was instead *no*.

predictions but also the actually-haired people who were incorrectly predicted to be bald.

As such, it is possible to see precision reflects the proportion of correct predictions, for the *yes* class.

2.2. **recall** – a measure of how exhaustive the predictor is, this is a measure of the number of correct predictions for the *yes* class, out of all *yes* class predictions (which include the correct predictions and the incorrect *false negatives).* Put another way, it captures how many misclassified entries (of the *yes* class) would need to be recalled in order to be corrected, so that at the end all entries of the *yes* class would be perfectly classified.

More concretely, for the *2.ii* example in section 3.3, it would correspond to the number of correctly predicted cases of baldness, out of both those correct predictions and actually-bald people who were incorrectly predicted to have hair.

As such, it is possible to see recall reflects the ability to identify all relevant cases in its predictions, for the *yes* class.

When attempting to detect cancer in a patient, for example, this is a very important measure, because it would be preferable to incorrectly predict that a healthy patient has cancer, rather than incorrectly predict an actually sick patient is perfectly healthy, when deciding for whom to ask for further medical examinations, for example.

3. **Criteria applicable to all types of continuous predictors**

3.1. **domain error** – a measure of the average error of the predicted outcomes, in the predictor's domain of applicability. An example would be: the absolute value of the difference between the prediction and the actual time an object took to fall to the ground, $|t_{predicted} - t_{measured}|$.

3.2. **out-of-domain error** – the same measure as the previous one, but applied to domains that are at least partially outside (meaning they can contain) the domain originally defined by the model. In a sense, it can be interpreted as the generalization error.

Looking at Newtonian gravity, these out-domains would be gravitational fields with exceptional magnitude, for example, the prediction of Mercury's orbit around the Sun versus its actual, correct orbit.

3.4.2 Uncertainty

Both Newtonian Gravity and General Relativity are examples of non-probabilistic predictors. That is, even if they are imperfect predictors, they make a single prediction with a confidence level of 100% in it.[13] On the other hand, probabilistic predictors can also be defined. These are extremely similar to their non-probabilistic counterparts, with the single difference being that, for probabilistic predictors, their predictions reflect their level of certainty in them.

That is, for the *2.i* example in section 3.3, instead of outputting a prediction between the three possible types `[humorous, critical, neutral]` of the form `[0, 1, 0]` or `[0%, 100%, 0%]`, picking a single category with total confidence, it can output instead a prediction of the type `[0.23, 0.62, 0.15]`, or `[23%, 62%, 15%]`, indicating a most likely category prediction of the type '*critical*,' with a confidence level of 62%.

The evaluation metrics for these types of predictors take into account not only the prediction, but also its confidence level. For example, if the prediction was correct, but the confidence level was low, then this prediction still has some error in it, as a perfect predictor would have no doubt.[14] On the other hand, if the prediction is wrong, but its confidence level was low, then this error is smaller than a wrong prediction with higher associated confidence level, as the model was more cognizant of its biases.[15][16]

Since the difference between probabilistic and non-probabilistic predictors is fully handled by their evaluation metrics, and no other significant differences arise within this framework between them, they shall henceforth be jointly denominated as simply predictors once again.

[13] Non-probabilistic predictors can be viewed as probabilistic predictors whose predictions are always attached to a probability of 1. The reason it can be preferable to treat them separately, as non-probabilistic, has to do with the simpler operations involved, as prediction confidence can be ignored. Nonetheless, all the tools used for probabilistic predictors can still be applied to the non-probabilistic ones.

[14] To illustrate this case, let us imagine two doctors, diagnosing whether a patient is ill. The first doctor tells the patient that, according to her diagnosis, he is not sick, but that she is only 55% sure of that. The second doctor tells the patient he is not ill, and that she is 80% sure of her diagnosis. Assuming both doctors have approximately equal accuracy on correct diagnosing, the second doctor will be intuitively recognized as the better one at diagnosing this illness, despite the fact that both of them made a correct prediction.

[15] Similarly to the previous example, if the patient was actually sick, then suddenly, the first doctor stands as the better one – though both were wrong, the first doctor was much more uncertain of her diagnosis, and admitted there was much more margin for error, having possibly recognized the patient as an edge case.

[16] On a side note, it is worth mentioning that attaching confidence levels to predictions is not exclusive to a human diagnostic. Instead of predicting the outcome, by predicting the likelihood of each outcome, computers too can produce predictions coupled with confidence levels.

4 Explanatory adequacy

How adequate is an explanation? There are several possible criteria, one of which is its predictive power. If one was to ignore the predictive component of an explanation, it would be possible to ascertain other criteria: sense of understanding inspired by the explanation, degrees of explanation, unification power, precision, or even the beauty and simplicity of the theory that generated that explanation.

Is an explanation's adequacy inferred from the sense of understanding inspired in us by it? Some have argued this cannot be used as a reliable indicator of the theory's accuracy (Trout 2002, pp. 5-11, 13-18).

Lipton (Lipton et al. 2004, pp. 57-62, 142-148, 192-198, 207-208) has provided an alternative approach of adequacy in his *Inference to the Best Explanation* book. He defends that one should subscribe to the *loveliest*, as opposed to the likeliest, explanation. He argues that likeliness relates to truth, but that very link might require the explanation to be overly vague or noncommittal, whereas loveliness correlates with "potential understanding." This however, as mentioned by Douglas (Douglas 2009, pp. 16-18), does not justify why lovely explanations should be considered likely at all. And if an explanation is unlikely, there does not appear to be strong support behind it, and it may even be hard to disprove, if, for example, it resorts to the metaphysical.

It is possible to resort to a different approach, based on the predictive power of each model being evaluated. As part of the *Predictor framework*, multiple metrics have been designed, which can also be adapted to specific contexts. These metrics are the main tools by which predictors are evaluated, and their diversity aims at capturing different aspects of the correctness of prediction of each predictor. As it has been previously highlighted, these metrics can give importance to different things.[17]

Mirroring the diversity of categories of explanation that can be treated equally as predictors, the flexibility of the focus of these metrics allows to capture several facets of a predictors adequacy. As will be detailed in section 4.1, the best set of metrics to be used is always first constrained by the domain, and secondly by the purpose with which one is evaluating these predictors.

[17] An easy example is *accuracy*, which values correctness (finding only needles in the haystack, or as few stalks of hay as possible, even if some needles remain in the haystack), as opposed to *recall*, which values exhaustiveness (finding all, or as many as possible, needles in the haystack, even if some hay comes through as well).

4.1 Comparing the adequacy of different predictors

In this section, the general procedure for comparing two predictors is presented, while section 4.2 provides a concrete example of its application. To accomplish the comparison between different predictors, the evaluation metrics outlined in section 3.4 are employed.

1. Firstly, the relevant domain must be established. For example, if the domain is Earth's gravity, then there is a negligible difference between Newtonian and General relativity. This is no longer true if the domain encompasses the universe, or even just the solar system, since the Newtonian predictions would start losing accuracy the closer the body being studied was to the Sun, due its mass. This highlights the extreme importance of domain definition. Section 4.2 presents a more complete discussion, using the same example.

2. After establishing the evaluation domain, the same metric (or group of metrics) should be evaluated on each of the different models. The metrics should be tailored to the problem being evaluated, but if there is no expert-level knowledge of the topic, then the metrics presented in section 3.4 will serve as a good baseline. While it is true the more knowledgeable one is of an area, the more fine-grained verifications they can make, the general metrics provide much power already. This is because of two main reasons: they incorporate much flexibility already, and when evaluating a model, even one who is not knowledgeable of the area will know what about their model's performance is the most important;[18] and the fact that almost all more specialized metrics are derived or otherwise obtain from these general metrics, or combinations of them.

3. Finally, the results should be compared (i.e. the results are compared to check which model has the higher accuracy or the lower loss). Only if the difference between the obtained results is negligible (for example, the difference between the two models' results is an order of magnitude smaller than the metrics themselves), should one employ a secondary criterion to judge which forms the better explanation.[19] Otherwise, the predictor with the better results is the best model, since it has the strongest explanatory power for that domain.

[18] For example, using our *2.i* example from section 3.3, what matters the most? Overall accuracy? Correctly identifying humorous headlines, even if the results of critical and neutral suffer as a consequence? Identifying all critical headlines, even if some humorous or neutral ones are filtered in as well? That will inevitably depend on the problem this model is being used to explain.

[19] This secondary criteria should be the second most relevant metric, for the given domain.

4.2 Comparing Newtonian Gravity and General Relativity

All throughout the essay, Newtonian and Relativistic physics have been referenced as an application example for the *Predictor Framework*. In section 3.3, the predictors for both theories were defined. Let us now apply the procedure defined in 3.1 in order to compare them.

As previously stated, the choice of the better model is constrained first and foremost by the domain of application. Let us consider two possibilities:

1. if this domain is unlimited, applying to the observable universe, or if captures the Solar System at least (thus including the previously-mentioned perihelion of Mercury), then General relativity will emerge as the better theory, because its accuracy is better by a non-negligible amount.

2. if however, the domain does not include any instance for which classical gravity fails,[20] the differences between the model predictions are insignificant for that domain, and which of the two models is the better one is conditioned on the secondary criteria.

Secondary criteria are dependent on the application for the models, in order the choose the most appropriate secondary metric that will be used in order to break the tie, such as the ones presented in (Cherkassky and Yunqian Ma 2004 pp. 2-3), for this particular case. Alternatively, if one wishes to break the tie not in favor of performance, but on a specialized criterion, non-general metrics can be applied. If there is only desire for understanding, then a metric evaluating the theory's simplicity behind each predictor can be applied, for example by evaluating the computation cost of generating a new prediction.[21] A better motivation for picking simplicity as metric can be based on the real-life example of Apollo 11's lunar mission. In such a case, there is a concrete objective for the application of these predictors. In Apollo 11's lunar mission, the model needed to be programmed into the spaceship's guidance computers, and computer memory was a severe constraining factor, due to its high cost in both monetary and spatial senses. As such, in this case, simplicity should be chosen as the secondary criterion, yielding Newtonian gravity as the better model, given both domain and application. Indeed, Apollo 11 took mankind to the moon using Newtonian gravity, despite the scientists knowing its faults![22]

[20] Classical gravity is not accurate, for example, when Mercury's perihelion is included in the domain, due to its close proximity to the Sun's enormous mass.

[21] This metric would even allow one to somewhat follow Lipton's "loveliness" criterion, should they so wish, on the premise that the less convoluted a model is, the less likely it is for it to be correct, since it has less freedom to adapt to the entry data, than a similar, but more complex, alternative model.

[22] It may be worthy of emphasis that the domain comprising the space around Earth and the Moon is

5 Conclusion

In this essay, a framework which can encompass the most popular categories of explanation in philosophy of scientific explanation has been presented. This is done by abstracting away from the underlying mechanisms of the theory, and focusing on its model or predictor, that is, the concrete set of predictions it can generate, given all the input it requires to do so. This opens a myriad of possibilities for constructing new predictors, evaluate them and compare their explanatory power amongst several candidates.

Despite the existence and differences between the several categories of explanation models, it has been shown how they can come together under a single predictive umbrella. This predictive approach was further elaborated into the concept of a predictor, composed of three main components: the theory, the domain of applicability, and the predictive mechanism. This concept was more deeply expanded, incorporating different possible types of predictors - continuous or discrete - as well as the construction of one given a scientific theory, as demonstrated using the classical and relativistic theories of gravity.

Through a predictor, it then becomes possible to define the *Predictor Framework*, which incorporates not only the aforementioned predictions but also several evaluation and comparison metrics and procedures, including those of probabilistic predictors.

Finally, the topic of explanatory adequacy is discussed, leveraging the new set of tools the *Predictor Framework* enables, the domain of applicability and an appropriate set of evaluation metrics being of particular importance. This is demonstrated in a final comparison between the Newtonian classical mechanics and Einstein's general relativity theories of gravity.

This framework allows a prediction-based approach to scientific explanation, which is advantageous because it allows the grouping of different categories of explanation into a comparable whole.

such that the difference between the accuracy of two theories is so minimal that it can be disregarded.

References

Salmon, Wesley C. (1990). "Scientific Explanation: Causation and Unification". In: *Crítica: Revista Hispanoamericana de Filosofía* 22.66, pp. 3–23. ISSN: 00111503. URL: http://www.jstor.org/stable/40104633.

— (1982). "Comets, Pollen and Dreams: Some Reflections on Scientific Explanation". In: *What? Where? When? Why? Essays on Induction, Space and Time, Explanation*. Ed. by Robert McLaughlin. Dordrecht: Springer Netherlands, pp. 155–178. ISBN: 978-94-009-7731-0. DOI: 10.1007/978-94-009-7731-0_7. URL: https://doi.org/10.1007/978-94-009-7731-0_7.

Woodward, James (2017). "Scientific Explanation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2017. Metaphysics Research Lab, Stanford University.

Hempel, Carl (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press.

Douglas, Heather E. (2009). "Reintroducing Prediction to Explanation". In: *Philosophy of Science* 76.4, pp. 444–463. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/10.1086/648111.

Kitcher, Philip (1989). "Explanatory Unification and the Causal Structure of the World". In: *Scientific Explanation*. Ed. by Philip Kitcher and Wesley Salmon. Minneapolis: University of Minnesota Press, pp. 410–505.

Hitchcock, Christopher Read (1995). "Salmon on Explanatory Relevance". In: *Philosophy of Science* 62.2, pp. 304–320. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/188436.

Rosenbaum, Paul R. (2019). *Observation and Experiment: an Introduction to Causal Inference*. Harvard University Press.

Lipton, P. et al. (2004). *Inference to the Best Explanation*. International library of philosophy and scientific method. Routledge/Taylor and Francis Group. ISBN: 9780415242028. URL: https://books.google.ch/books?id=WIfYNExpSC0C.

Machamer, Peter K., Lindley Darden, and Carl F. Craver (2000). "Thinking About Mechanisms". In: *Philosophy of Science* 67.1, pp. 1–25. DOI: 10.1086/392759.

Misner, Charles W., Kip S. Thorne, and John Archibald Wheeler (1973). *Gravitation*. New York: W. H. Freeman and Company, p. 1049. ISBN: 9780716703440.

Grøn, Ø. and S. Hervik (2007). *Einstein's General Theory of Relativity: With Modern Applications in Cosmology*. Springer New York. ISBN: 9780387692005. URL: https://books.google.ch/books?id=IyJhCHAryuUC.

Weinberg, S. (1972). *Gravitation and cosmology: principles and applications of the general theory of relativity*. Wiley. ISBN: 9780471925675. URL: https://books.google.ch/books?id=0LrZkgPsZR0C.

Schwarz, John H. (2007). "String Theory: Progress and Problems". In: *Progress of Theoretical Physics Supplement* 170, pp. 214–226. ISSN: 0375-9687. DOI: 10.1143/ptps.170.214. URL: http://dx.doi.org/10.1143/PTPS.170.214.

Ashtekar, Abhay (Sept. 2008). "Loop Quantum Gravity: Four Recent Advances and a Dozen Frequently Asked Questions". In: *The Eleventh Marcel Grossmann Meeting*. DOI: 10.1142/9789812834300_0008. URL: http://dx.doi.org/10.1142/9789812834300_0008.

Trout, J. D. (2002). "Scientific Explanation And The Sense Of Understanding*". In: *Philosophy of Science* 69.2, pp. 212–233. ISSN: 00318248, 1539767X. URL: https://www.jstor.org/stable/10.1086/341050.

Cherkassky, V. and Yunqian Ma (2004). "Comparison of loss functions for linear regression". In: *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*. Vol. 1, pp. 395–400.