Unification and Surprise:

On the Confirmatory Reach of Unification

Elena Castellani[*], Radin Dardashti[†] and Richard Dawid[‡]

Abstract

There is no doubt that a theory that is unified has a certain appeal. Scientific practice in fundamental physics relies heavily on it. But is a unified theory more likely to be empirically adequate than a non-unified theory? Myrvold has pointed out that, on a Bayesian account, only a specific form of unification, which he calls mutual information unification, can have confirmatory value. In this paper, we argue that Myrvold's analysis suffers from an overly narrow understanding of what counts as evidence. If one frames evidence in a way that includes observations beyond the theory's intended domain, one finds a much richer and more interesting perspective on the connection between unification and theory confirmation. By adopting this strategy, we give a Bayesian account of unification that (i) goes beyond mutual information unification to include other cases of unification, and (ii) gives a crucial role to the element of surprise in the discovery of a unified theory. We illustrate the explanatory strength of this account with some cases from fundamental physics and other disciplines.

[*]Department of Humanities and Philosophy, University of Florence. E-mail: elena.castellani@unifi.it
[†]Department of Philosophy, Johannes Gutenberg University Mainz. E-mail: dardashti@uni-mainz.de
[‡]Department of Philosophy, Stockholm University. E-mail: richard.dawid@philosophy.su.se

1 Introduction

Unification plays an important motivational role in physics. It is a core question in this context whether the fact that a theory unifies should be taken to increase the theory's probability of being viable.[1] Wayne Myrvold (2003; 2017) has defended the position that only one specific form of unification, which he calls *mutual information unification*, carries epistemic weight. If a theory is unifying in this sense, its probability of being viable is increased because the theory renders seemingly distinct phenomena informationally relevant to each other. Cases of unification that do not fall into this category carry no epistemic weight on that account. Myrvold's analysis plays out within the framework of Bayesian confirmation theory. Cases of unification that cannot be modelled in terms of mutual information unification do not, in his understanding, generate confirmation in terms of Bayesian updating.

Myrvold's view has been criticised or amended by a number of authors. Lange (2004) points at scenarios where, on his account, the attribution of confirmation value along the lines of Myrvold's criterion seems implausible. Blanchard (2018) argues that Myrovld's account is in need of being amended by some specific guidelines regarding the selection of priors to fully account for the confirmatory role of unification. Earlier, Janssen (2002) pointed out that explanatory unification, by tracing different "classes of facts" to a common origin, has provided substantial grounds for trusting a theory in historical cases of theory assessment. Those modes of reasoning, in Myrvold's terminology, amount to "common origin unification" and, therefore, would extend the range of forms of unification that are epistemically relevant beyond what Myrvold's criterion allows for (Myrvold 2017).

In this paper, we argue that Myrvold's account does not exhaust the types of unification that are epistemically significant in a Bayesian framework. Myrvold's conclusion that the

[1] We will mostly rely on the concept of viability rather than the difficult concept of truth. A theory is viable in a given empirical domain if it is in agreement with all the empirical data that can be collected in that domain.

confirmation value of unification is limited to mutual information confirmation only holds if one constrains what counts as evidence for a theory to observations that lie within the intended domain of that theory (that is, to observations of the kind that can be predicted or ruled out by the theory). This constrained understanding of evidence is not implied by Bayesian reasoning and has been shown to be extendable in important ways (Dawid et al. 2015). We will show that, once one lifts the stated constraint, one can establish, within a Bayesian framework, confirmation value for cases of unification that reach beyond what is covered by Myrvold's mutual information unification.

Our analysis supports the considerations of Lange, Blanchard and Janssen that point to a wider confirmatory role of unification than Myrvold's account allows. We will demonstrate, however, that the confirmation value of unification does not hinge on explanatory power per se, as was suggested by Lange or, more implicitly, by Janssen. Unification becomes epistemically significant, on our account, if the existence of a unified theory is *surprising* on the basis of previous expectations. Historically, this element of surprise was already recognized as an important aspect of the research process by Whewell (1863) in the context of his consilience of induction.

In Section 2, we present Myrvold's canonical Bayesian account of the confirmation value of unification and discuss some criticisms of his approach. Section 3 introduces the two key elements of our approach: Sect. 3.1 presents the theory space approach that provides the basis for an extended understanding of Bayesian confirmation. Sect. 3.2 highlights the importance of the element of surprise through Whewell's analysis of consilience of induction. Section 4 presents the Bayesian model of our account of unification. Section 5 discusses four case studies that illustrate the importance of our presented approach for understanding the epistemic role of unification.

2 The debate on the confirmation value of unification

Myrvold adopts a thoroughly Bayesian perspective on unification. He assumes that, by definition, confirmation hinges on updating credence in a hypothesis H under empirical evidence E: evidence E confirms H iff $P(H|E) > P(H)$. Myrvold's strategy of identifying confirmatory elements of unification therefore relies on the standard procedure of identifying confirmation. That is, he looks for a theory's successful empirical predictions and establishes that updating on the predicted evidence increases credence in the theory. To establish a link to unification, Myrvold distinguishes two types of predictions. First, a theory may predict characteristics of a given physical phenomenon unconditionally or solely dependent on previous data on the same phenomenon. This kind of prediction, if successful, leads to confirmation but is not related to unification on his view. Second, a theory may predict evidence E conditionally on a different set of evidence E' that would seem independent from E in the absence of theory H. Linking the two sets of evidence for Myrvold amounts to a

unificatory step based on conditional prediction. In Myrvold's words "a hypothesis can unify disparate phenomena: the hypothesis can make two phenomena that in the absence of the hypothesis seem to be independent phenomena yield information about each other." (Myrvold 2003, 408). Myrvold calls this form of unification *mutual information unification*. He then proceeds to show that mutual information unification is confirmatory.

In a second paper, Myrvold (2017) provides more extensive reasoning to establish that mutual information unification is the only form of unification that has confirmation value. He contrasts the former with what he calls "common origin confirmation", which "[has] to do with hypotheses that posit a common origin for the phenomena in question, be it a common cause or some other type of explanation." [92]. The latter, according to Myrvold, provides no basis for increasing credence through updating on new data, and therefore does not have confirmation value.

Myrvold's account has been criticized or amended by a number of authors. So it has been suggested that Myrvold's account fails to identify those cases of unification that would intuitively be considered confirmatory. An argument along those lines has been presented by Lange (2004): mutual information is no good indicator of confirmation value since it makes the generation of confirmation value too easy in some contexts and too difficult in others. Let's focus on the "too easy part" for the moment. Lange argues that stapling together two entirely unrelated theories would generate confirmation value on Myrvold's account, though it is counter-intuitive to attribute any confirmation value to the unified nature of a theory if its two "unified" parts are entirely disjoint.

This criticism has been responded to by Schupbach (2005) and Myrvold (2017). They agree that the cases pointed at by Lange indeed provide confirmation on Myrvold's account. They argue, however, that this fact looks problematic only as long as one is begging the question by assuming from the start that the confirmation value of unification must be based on explanatory value. Myrvold's account, they argue, provides a perspective that acknowledges confirmation value of non-explanatory unification. It does so not by following any specific intuition, but by rendering the confirmation value of unification consistent with a probabilistic understanding of theory assessment.

In a more recent paper, Blanchard (2018) shifts the problem Lange was pointing at to a more general conceptual level sand proposes a solution that remains in line with Myrvold's general approach. In line with Lange's stapling case, Blanchard argues that there is always the possibility to establish mutual information by hand without providing any deeper explanation. For example, if one observes a correlation between measurement outcomes in different regimes, one might add the hypothesis that this correlation will hold in the future just by chance, without any deeper explanation. Blanchard strengthens Lange's case against this perspective by directly comparing a case of mere mutual information unification to a case of common origin unification. He argues that mutual information unification, on Myrvold's account, would need to attribute the same credence to the described "just by chance"

hypothesis as to a hypothesis that offers a genuine explanation of the correlation based on finding a common origin. Blanchard considers this equivocation implausible and misguided.

Blanchard thus concludes that Myrvold's Bayesian analysis, as long as it merely focuses on updating under evidence, does not succeed to fully represent the confirmation value of unification. He agrees with Myrvold, however, that there is no other way than the one proposed by Myrvold to represent confirmation value of unification in terms of Bayesian updating. Blanchard therefore proposes to acknowledge explanatory unification as an argument that supports the selection of a higher prior, while there is no formal way of representing this increase in terms of Bayesian updating. Common origin unification, on this view, justifies substantially higher credence in the unifying hypothesis through the specification of priors.

3 A new take on the epistemic significance of unification

3.1 A wider understanding of Bayesian confirmation: A theory space approach

In the following, we will demonstrate that, under specific circumstances, the epistemic significance of other types of unification beyond mutual information unification can be established as a Bayesian updating effect. We agree with Myrvold and Blanchard that a Bayesian analysis that restricts the evidence used for updating to evidence within the intended domain of the theory allows the confirmation value of unification only within the limits set out in (Myrvold 2003; 2017). We further agree with Blanchard that Myrvold's account misses out on a crucial element of confirmation that can be found in unification. However, as our discussion will demonstrate, merely alluding to the selection of priors is not sufficient to fully represent the confirmation value of unification either. We will demonstrate that a wider form of Bayesian analysis can represent the at times substantial confirmation value of common origin unification in terms of a genuine updating effect.

The proposed approach is based on the concepts of theory space and empirical horizons.[2] An empirical horizon specifies limits within which we consider empirical tests (Dawid 2018, pp. 493, 499). Theory space counts the scientific theories that can be formulated, are in agreement with the available empirical data, and can be distinguished from each other by empirical testing within the specified empirical horizon. If we have credence in a scientific theory, we attribute a certain probability to the hypothesis that the theory is empirically viable within a certain empirical horizon. Credence in a theory is then understood to be informed by certain implicit views on constraints on theory space: to have high credence in a theory's viability, one implicitly assumes that there are probably just very few, if any, alternative

[2]We will use the intuitively appealing word "theory space" (Dardashti 2019) in what follows. Strictly speaking, in view of the lack of clear structural or metric characteristics of "theory space", it might be more accurately described as a "spectrum of theories".

scientific theories that are in agreement with the available empirical evidence and differ empirically from the given theory within the considered empirical horizon.

The crucial point is that assumptions of such constraints on theory space are not just ad hoc and arbitrary but can be supported or disfavored both by actively investigating theory space and by collecting *meta-empirical evidence*: observations that don't lie in the theory's intended domain but serve as indicators of the size of theory space. Based on specific arguments of meta-empirical theory assessment (MEA), (Dawid 2013), they play an essential role in empirical theory confirmation (Dawid 2018). The confirmatory character of meta-empirical evidence and the significance of confirmation on its basis has been established under plausible conditions in (Dawid et al. 2015, Dawid forthcoming).

In this paper we present a mechanism of theory confirmation that, like the proposed arguments of meta-empirical assessment, plays out in the context of assessments of theory space. While elements of meta-empirical assessment play an important role in our proposal, the core of the proposed argument works at a different level and is based on a different type of evidence: the surprising discovery of a unified theory for a given set of observations.

3.2 The role of surprise: Whewell on Newton's Universal Law of Gravity

Scientists can be surprised in a variety of situations and for a variety of reasons. While there can be a subjective element involved in a surprising observation, it has been widely recognized that the element of surprise plays a crucial *epistemic role* in scientific practice (see e.g. (Morgan 2005, Currie 2018, French and Murphy 2023)) and more generally in confirmation theory (Horwich 1982). We will focus on a specific kind of surprise that was emphasised early on by William Whewell when discussing Newton's universal law of gravity. The element of surprise is an aspect of Whewell's account of Newton's achievement in terms of consilience that is of eminent importance to Whewell himself. According to a famous passage by Whewell, consilience "takes place when an Induction, obtained from one class of facts, coincides with an Induction, obtained from another different class." This consilience, for Whewell, "is a test of the truth of the Theory in which it occurs." In this regard, it is important to note that what makes such a "coincidence" a test of truth for the theory, beside the multiplicity and independence of the evidence ("classes of facts altogether different"), is also its *unexpectedness* – in Whewell (1840)'s own terms, an agreement "unforeseen and uncontemplated" (p. 65), and where "the unexpected coincidences of results drawn from distant parts of the subject" (p. 67).[3]

The role of surprise is well evident in Whewell's account of Newton's Theory of Universal Gravitation, his most known example of consilience. As Whewell himself underlines, the fact that Newton's inverse square law, which explained Kepler's Third Law, explained also

[3]Note also that what is unexpected – and therefore *surprising* – is the coincidence, not a new fact or prediction (Castellani 2024, Sect. 2.2).

Kepler's First and Second Laws "although no connexion of these laws had been visible before, ... is a most striking and surprising coincidence, which gave to the theory a stamp of truth beyond the power of ingenuity to counterfeit" (1840, pp. 65-66). In other words, according to Whewell the "striking and surprising" fact that the unifying theory can explain unrelated phenomena or laws is an essential part of the argument for increasing the trust in the theory's truth. It was, at the time of Newton, surprising that the disparate sets of observed regularities on planetary motions were such that a unified description existed.

Relying on the concepts introduced above, we can represent the stated surprise in terms of expectations regarding the corresponding theory space. If one expected that a large number of unified theories could be constructed that would fit any pattern of observed regularities in the given context, there would be nothing surprising or coincidental about these data sets. Surprise translates into the expectation that, given the available data, credence in the existence of even one unified theory that fits the data is much lower than one. Note that this in turn implies that an element of surprise that is based on an inadequate understanding of theory space will not add a "stamp of truth" to the theory. A scientist who misjudges the situation or is simply uninformed about the conceptual space for theory development may be surprised, but their surprise is a consequence of ignorance, which makes it irrelevant for theory assessment.

Janssen (2002) in his discussion of Whewell, disregards Whewell's emphasis on surprise and focuses on the given theory's explanatory value based on a common origin. But his disregard for the surprise factor makes his proposal overly inclusive. For example, it would grant conspiracy theories epistemic value, as they heavily rely on allusions to a common origin explanation (Keeley 1999). Conspiracy theories point at the unificatory power of their theories as a main reason to believe their theory: if that theory can integrate so many seemingly independent observations into a coherent whole, they argue, there must be something to it. To the extent a conspiracy theory is well crafted and not obviously ridiculous, Janssen's account of the confirmation value of common origin unification would suggest that the conspiracy theorist is right to make that claim. In the next section, we will show that linking the confirmation value of unification to a well-founded surprise factor rather than to an explanation avoids this unattractive and unintuitive consequence.

4 Modelling the argument

4.1 The general idea

So we are now concerned with the following question: Does the element of surprise generate confirmation value of unification in Bayesian confirmation theory when modeled as a feature of scientists' understanding of theory space? Here is the reasoning we propose to explain why it does. It is by no means to be expected that scientific theories that unify two seemingly different sets of phenomena can always be built. In fact, the world abounds with sets of phenomena where a unified scientific description seems hopeless. Finding a unified theory that

agrees with the available data on two or several different phenomena amounts to finding out that the data collected on the said phenomena is of a kind that allows for a unified description. This can be, under certain conditions, a surprising discovery. The very fact that a unified theory has been found in such cases increases, or so we argue, the credence in the given theory.

More specifically, the mechanism that leads to an increase of credence in a theory due to the theory's unified character would be the following.

Condition 1: Scientists face data[4] from two or more distinct domains.

Condition 2: Due to the substantial difference between the characteristics of those sets of data and the complexity of observed phenomena, they do not expect that a unified theory that covers all these domains exists.

Condition 3: They find a unified theory.

If those conditions are fulfilled, scientists are willing to make the following two inferences:

Inference I: On the scientists' understanding, the best reason why the data in all those domains is of a kind that allows for a unified theory is that the viable theory (not necessarily the one found) is unified. Therefore, they infer that the viable theory is probably indeed unified.

Inference II: The lack of alternative unified theories increases our credence regarding the unified theory that has been found.

Inference I increases our confidence that the viable theory is indeed unified. Inference II allows us to increase our confidence in the concrete unified theory that has been found. It is important to distinguish between these two inferences, because finding a unified theory may provide techniques for developing further unified accounts of these domains and therefore generate credence in the existence of a higher number of unified theories. We will illustrate this feature with concrete examples in Sect. 5.

4.2 A Bayesian reconstruction

4.2.1 The variables

In the following, we will reconstruct the above line of reasoning in Bayesian terms. To be able to update under the observation that a given theory H is unifying, we define H as the theory developed by some particular scientists at some particular time, rather than in terms of the content of the theory. On this basis we can treat unifying/not unifying as values of a variable to be discovered. If we were to define H in terms of the content of the theory, the value of this variable would be implied by the theory.

[4]Here data can also be encoded in terms of phenomenological laws.

We introduce the variable F_U, which can take the two values: The developed theory H does (does not) unify the available evidence E in domain D.[5] Note that we distinguish two sets of empirical data. E is the empirical data known at the given point, which theory H was developed to account for. Domain D, to the contrary, denotes the much wider set of all empirical data that could be collected in principle within some specified "empirical horizon". If a theory agrees with all possible data D, it is called *viable* in D.[6] We introduce the variable $\exists_U$, which can take the two values: There is at least one (is no) unified theory that accounts for the available evidence E in domain D. Obviously, whether or not there is a unified theory affects whether or not we find the developed theory to be unified. In particular, if we knew that there are no unified theories, the probability that we would find the developed theory to be unified would be zero. We introduce the variable T_U, which takes on the two values: The viable theory of the domain D is (is not) a unified theory. Further we introduce the variable T_H, which takes the values: H is (is not) the viable theory of domain D. Recall that H refers to the theory that has been discovered, which leaves open whether or not the theory is unified. Note that while F_U and $\exists_U$ are about theories in agreement with the available evidence E, T_U and T_H are about a theory's viability in D. The question of whether a unified theory accounts for E is distinct from the question whether the viable theory of the domain D is actually unified.

4.2.2 The overall setup

Our analysis will be based on three core dependencies between variables. Since F_U and T_U imply $\exists_U$, we have:

$$P(\exists_U|T_U) = 1, \quad P(T_U|\overline{\exists_U}) = 0 \tag{4.1}$$

$$P(\exists_U|F_U) = 1, \quad P(F_U|\overline{\exists_U}) = 0 \tag{4.2}$$

Moreover, since F_U and T_H together imply T_U, we have

$$P(T_H, F_U) = P(T_H, F_U, T_U) \tag{4.3}$$

[5]Theories can be more or less unified. This is evident in Myrvold's account of mutual information unification. However, we use a binary variable to depict the evidence for unification because we are only concerned with whether the theory unifies at all, rather than by how much. Although we would indeed expect a more unified theory to have a stronger evidential impact on the viability of a theory than a less unified theory, this is not modelled explicitly; however, it will impact the analysis via the surprise element that will be introduced later on.

[6]We individuate theories in a way that only allows for one viable theory of a given domain. Schemes that may differ in their predictions outside D but are empirically equivalent in D are treated as one "effective" theory in D that has several possible fundamental theories. See (Dawid 2016, Section 3)

Our aim is to demonstrate that there are constellations under which $P(T_H|F_U)$ is substantially larger than $P(T_H)$. Using Equation (4.3) and the definition of conditional probability we get:

$$P(T_H|F_U) = \frac{P(T_H, F_U)}{P(F_U)} = \frac{P(T_H, F_U, T_U)}{P(F_U)} = P(T_H|F_U, T_U)P(T_U|F_U) \tag{4.4}$$

Inference I , as spelled out in Section 4.1, amounts to extracting a high value of $P(T_U|F_U)$. Inference II then leads from there to the extraction of a high value for $P(T_H|F_U)$.

4.2.3 Modeling Inference I

Let us first analyse the factor $P(T_U|F_U)$. Using Eq. (4.1) and the law of total probability for $P(T_U)$, we can write

$$P(T_U|F_U) = \frac{P(T_U)}{P(\exists_U)} \frac{P(T_U|F_U)}{P(T_U|\exists_U)} \tag{4.5}$$

This allows us to assess $\frac{P(T_U|F_U)}{P(T_U|\exists_U)}$. $\frac{P(T_U|F_U)}{P(T_U|\exists_U)} > 1$, which corresponds to the scenario that finding a unified theory increases the credence that the viable theory is unified compared to knowing that there exists a unified theory without finding it, may have some plausibility. For example, one might assume that simpler theories are in some sense more likely to be viable and also more likely to be found. However, even if such an argument had some basis, we are not interested in it. In order to focus on the surprise effect, we will neglect any such effect. The scenario $\frac{P(T_U|F_U)}{P(T_U|\exists_U)} < 1$, where finding a unified theory lowers credence that the viable theory is unified compared to knowing that there exists a unified theory without finding one, seems somewhat pathological and shall be discarded. We will therefore assume for the rest of this paper that $P(T_U|F_U) = P(T_U|\exists_U)$, so that we have

$$P(T_U|F_U) = \frac{P(T_U)}{P(\exists_U)} \geq P(T_U) \tag{4.6}$$

Using the total probability for $P(\exists_U)$ and Eqn. (4.1), we can write

$$P(T_U|F_U) = \frac{P(T_U)}{P(T_U) + P(\overline{T_U})(P(\exists_U|\overline{T_U})} \tag{4.7}$$

Equation (4.7) shows that the surprise effect has its source in $P(\exists_U|\overline{T_U})$. If that value is small compared to T_U, the viability of a unified theory is the most probable reason for the existence of a unified theory, which boosts the posterior credence in T_U. F_U thus confirms T_U due to the surprise factor.

4.2.4 Inference II and Theory Space

While we have established confirmation of the viability of some unified theory, we have not yet learned much about the viability of the specific unified theory H that has been found. How

can T_U affect the viability of H? From the previous section we obtained:

$$P(T_H|F_U) \quad = \quad P(T_H|F_U, T_U)\frac{P(T_U)}{P(T_U) + P(\overline{T_U})(P(\exists_U|P(\overline{T_U}))} \tag{4.8}$$

The first term in (4.8) is where the theory space concept, introduced in Section 3.1, becomes crucial. Treating theory space as an epistemically significant concept changes the status of $P(T_H|F_U)$. The method of MEA (Dawid 2013, Dawid et al. 2015) allows for epistemically significant evaluations of the size of theory space based on various forms of meta-empirical data and turns assessments of theory space into an integral part of Bayesian updating. MEA is therefore essential to our discussion by providing the basis for the epistemic significance of theory space in general, and of the surprise effect in particular. Once embedded in theory space reasoning, F_U can be treated as data on which hypotheses about the size of theory space can be updated. To account for unification as a concept of potential confirmation value, we need to introduce two separate theory spaces, one for unified and one for non-unified theories. In line with the Bayesian model of MEA provided in Dawid et al. (2015), we therefore introduce two variables Y_U and $Y_{\overline{U}}$, which take the values Y^i_U and $\mathrm{Y}^{\mathrm{j}}_{\overline{\mathrm{U}}}$, respectively: there exist $i \in I_U$ ($j \in I_{\overline{U}}$) possible unified (not unified) theories of domain D.

We now have all nodes necessary to establish the dependence of T_U on T_H. To extract the probability $P(T_H|F_U, T_U)$, we need to consider the structure of theory space. The idea of theory space is that the numbers i of possible alternatives determine the probability that theory H is viable, given that it is unified and the viable theory is unified as well. Therefore, we need to consider all $\mathrm{Y}^{\mathrm{i}}_{\mathrm{U}}$ weighted by their probabilities, which gives

$$P(T_H|F_U, T_U) = \quad \sum_{i \in I_U} P(T_H|F_U, T_U, Y^i_U)\, P(Y^i_U). \tag{4.9}$$

Similarly, for the non-unified case, we have

$$P(T_H|\overline{F_U}, \overline{T_U}) = \sum_{j \in I_{\overline{U}}} P(T_H|\overline{F_U}, \overline{T_U}, Y^j_{\overline{U}})\, P(Y^j_{\overline{U}}) \tag{4.10}$$

The default view on the probabilitsic import of the $\mathrm{Y}^{\mathrm{i}}_{\mathrm{U}}$ and $\mathrm{Y}^{\mathrm{j}}_{\overline{\mathrm{U}}}$ is to view a theory as a random pick from the pool of all theories that are in agreement with the available data and empirically distinguishable within the considered empirical horizon. If i theories satisfy the stated conditions, the probability of developing (picking) the viable theory would be $1/i$.[7]

At first glance, one might think that an extensive assessment of theory space for possible unified theories is not necessary in case of a very big surprise factor: the latter would imply that we take the chances of even one non-viable unified theory to exist to be very small. But if

[7]Theory space analysis does not require this simplest quantitative model, but we use it in our quantitative examples for the sake of simplicity.

we basically assume that there is no other unified theory than H, then knowing that T_U and F_U are true would directly imply the viability of the discovered theory H. However, one needs to take the following possibility into account: it might happen that the discovery of one unified theory opens the methodological path to the discovery of many more. In this case, we would not expect much impact on the viability of that particular discovered H. Thus, in order to understand the epistemic significance of finding a unified theory, we need to carry out an assessment of the unified theory space in light of the newly discovered unified theory.

How can we assess the size of theory space? The general mechanism has been discussed in detail in Dawid et al. (2015). For our purpose we can consider a simplified version. The variable for the no alternatives argument (NAA) F^H_{NAA} takes the values: the scientific community has not (has) found a unified theory of domain D that is not H, despite serious search for such theories.[8] One can then show (see Dawid et al. (2015) for details) that $P(T_H|F^H_{NAA}) > P(T_H)$ under the reasonable assumptions that increasing the number of alternatives does not make it less likely to find an alternative and less likely to have found the empirically adequate theory. Depending on whether a substantial no alternatives argument exists or not, we will have an argument in favour of the discovered unified theory in some situations and won't in some others. We will come back to this and discuss examples in the next section.

4.2.5 A measure for the confirmation value of unification

Let us now get a grasp of the confirmation value of unification on our account. To emphasize the difference between finding a unified theory with the case of finding it to be non-unified, we use the confirmation measure

$$\Delta = P(T_H|F_U) - P(T_H|\overline{F_U}), \tag{4.11}$$

We have spelled out the first term already in Equation 4.8. The second term yields:

$$\begin{aligned} P(T_H|\overline{F_U}) &= P(T_H|\overline{F_U}, T_U)P(T_U|\overline{F_U}) + P(T_H|\overline{F_U}, \overline{T_U})P(\overline{T_U}|\overline{F_U}) && (4.12)\\ &= P(T_H|\overline{F_U}, \overline{T_U})P(\overline{T_U}|\overline{F_U}). && (4.13) \end{aligned}$$

Note that T_H now refers to the viability of a non-unified H. So we have used that $P(T_H|\overline{F_U}, T_U) = 0$: if H is not unified but the viable theory is unified, then H can not be viable. As we have already mentioned, the impact of the variable T_U on T_H happens via the theory space node: if T_U is the case, the posterior of T_H depends on the space of all viable unified

[8] Our knowledge of alternatives may depend on certain contextual issues, such as how difficult it is to find a theory. These issues are discussed in detail in Dawid et al. (2015) and will not be repeated here.

theories (i.e. Y_U); if $\overline{T_U}$ is the case, the posterior of T_H depends on the space of all viable non-unified theories (i.e. $Y_{\overline{U}}$). Turning back to Equation (4.11), and using Equations (4.9) and (4.10) we get:

$$\begin{aligned}
\Delta &= P(T_H|F_U) - P(T_H|\overline{F_U}) && (4.14)\\
&= P(T_H|F_U, T_U)\frac{P(T_U)}{P(\exists_U)} - P(T_H|\overline{F_U}, \overline{T_U})P(\overline{T_U}|\overline{F_U}) && (4.15)\\
&= \sum_{i\in I_U} P(T_H|F_U, T_U, Y_U^i)\; P(Y_U^i)\frac{P(T_U)}{P(\exists_U)} - \sum_{i\in I_{\overline{U}}} P(T_H|\overline{F_U}, \overline{T_U}, Y_{\overline{U}}^i)P(Y_{\overline{U}}^i)P(\overline{T_U}|\overline{F_U}) && (4.16)
\end{aligned}$$

4.3 Results

We are now in the position to discuss the implications of the model we developed. The main point, as we will see, is the fact that the introduction of theory space assessment, in conjunction with our way of specifying hypothesis H "externally" without determining its unifcation characteristics, introduces an epistemically relevant new degree of freedom that allows us to model the surprise effect.

The point of departure: the general random pick principle. Under specific conditions, the confirmation effect of unification is indeed zero. As described in Section 3.3., theory space controls the probability of a theory's viability. Let us assume that a theory is a random pick from the pool of all theories that are in agreement with the available data and empirically distinguishable within the considered empirical horizon. If i theories satisfy the stated conditions, the probability of developing (picking) the viable theory would be $1/i$. If we apply this *general random pick principle*, $P(T_U)$ cannot be freely chosen but follows directly from the expected overall number of theories (unified and non-unified).

What drives the confirmation effect of unification is the partition of theory space into two subspaces for unified and non-unified theories that serve as independent pools for the random picks in case of T_U and in case of $\overline{T_U}$ respectively. The random pick principle is then only applied to the two subspaces individually. $P(T_U)$ therefore becomes a free parameter that can be specified based on a priori considerations on the probability that the viable theory is unified. Those considerations are independent from the priors for the sizes of the two theory spaces of unified and non-unified theories. Still, even once we have introduced the two theory subspaces, we are free to give $P(T_U)$ exactly the value that would be implied by the general random pick principle. Let us call $P_{RP}(T_U)$ the value of $P(T_U)$ that would have been implied by applying the genral random pick principle to one overall theory space. If we now set our prior as

$$P(T_U) = P_{RP}(T_U)\,, \tag{4.17}$$

we represent, within the framework of separate theory spaces for unified and non-unfied theories, the situation that no significance is attributed to the property of unification. In that case, the confirmation value of unification goes to zero. We provide a formal proof of this fact in appendix A. The described scenario corresponds to the straightforward Myrvold account without higher priors for unified theories. From this starting point, we now switch on two effects that increase credence in a unified theory. First, we increase the prior for unified theories above the general random pick level. This will resemble Blanchard's proposal to favor unification based on prior specification. Second, we will switch on the surprise effect.

Effect 1: Raised Priors By lifting $P(T_U)$ above $P_{RP}(T_U)$, we by hand assign extra credence to a theory if it unifies. Before looking at this point formally, let us ask the question: why can it make sense to raise $P(T_U)$ above the general random pick value? There are three answers to that question. First, scientists do it, as we will see when discussing the case studies in Section 5. Second, philosophers such as Blanchard have given general philosophical reasons why one should do it. The general approach we propose suggests a third answer which, due to lack of space, we can only flag in the present paper and will discuss in detail in follow up work in progress. The rough idea is the following. Any Bayesian analysis of induction must rely on prior preferences for structured theorizing to get off the ground. Specific forms of those preferences get strengthened over time or get flattened out depending on their general success records on the grounds of meta-inductive reasoning. If successful, they can lead to priors substantially above the random pick values in actual scientific contexts. A priori preference for unification on this understanding is a necessary starting point for having an evidence based investigation as to whether being unified is a helpful indicator of being viable.

Let us now formally discuss the pure effect of increased priors in the absence of an element of surprise. In our setup, this corresponds to considering the confirmation measure (4.14) for the case where we are certain that a unified theory exists ($P(\exists_U) = 1$). The second term of Equation (4.16) then gives:

$$P(\overline{T_U}|\overline{F_U}) = \frac{P(\overline{T_U}, \overline{F_U})}{P(\overline{F_U})} \tag{4.18}$$

$$= \frac{P(\exists_U)P(\overline{T_U}|\exists_U)P(\overline{F_U}|\exists_U) + P(\overline{\exists_U})P(\overline{T_U}|\overline{\exists_U})P(\overline{F_U}|\overline{\exists_U})}{P(\overline{F_U})} \tag{4.19}$$

$$= \frac{P(\overline{T_U})P(\overline{F_U}|\exists_U))}{P(\overline{F_U})} = P(\overline{T_U}) \tag{4.20}$$

In the last step we have used the law of total probability for $P(\overline{T_U})$ and $P(\overline{F_U})$ and inserted $P(\exists_U) = 1$. Inserting (4.20) into (4.14) we get:

$$\Delta^{\text{no surprise}} = \sum_{i \in I_U} P(T_H|F_U, T_U, Y_U^i)\, P(Y_U^i)\, P(T_U) - \sum_{i \in I_{\overline{U}}} P(T_H|\overline{F_U}, \overline{T_U}, Y_{\overline{U}}^i)\, P(Y_{\overline{U}}^i)\, P(\overline{T_U}) \tag{4.21}$$

What our model provides in this context is a theory space based representation of the fact that we can a priori assume that a unified theory has a higher probability of being viable, which means that we get a credence boost for a theory by finding that it is unified. This mirrors the proposal of Blanchard (2018) to establish epistemic significance of unification by incorporating it into the priors. In our framework, Blanchard specification of a higher prior for a unified than for a non-unified theory is represented as an updating process under F_U. Since Blanchard defines H in a way that already determines whether it is unified, for him this specification amounts to fixing a prior. In our framework, confirmation value of unification is generated by raising $P(T_U)$ above $P_{RP}(T_U)$. The connection between $P(T_U)$ and $P(T_H|F_U)$ is then given by equation (4.21). Since our discussion up to this point is just a translation of Blanchard's selection of priors for unified theories into a theory space-based representation, it is still consistent with Myrvold's analysis.

Effect 2: The Element of Surprise We can now illustrate the impact of surprise by comparing a Δ that includes a surprise effect to the no-surprise scenario $\Delta^{\text{no surprise}}$ from above. Up to this point, theory space had not been essential for spelling out the described positions. This will change now. To formally represent surprise about the fact that there exists a unified theory, one needs to introduce a degree of freedom that characterises the expectation as to how many theories of a certain kind exist. That is, we need a perspective on our expectations regarding theories we have not yet found. As discussed in Section 4.2.4, this is exactly what the theory space approach provides.

On our theory-space based account, the surprise scenario and the no-surprise scenario amount to different systems of prior belief. A Myrvoldian analysis, to the contrary, does not take theory space considerations into account and thus is blind to the degree of freedom that controls the surprise factor. Therefore, it will in effect represent any belief system as a no-surprise system. What amounts to a comparison of two different scenarios on our account, therefore can also be viewed as a comparison between a Myrvoldian account and a theory-space based account of a scenario that does involve a surprise factor. From that angle, our comparison demonstrates that the Myrvoldian account disregards an epistemically important element of theory assessment. Equations (4.16) and (4.21), after simple rewriting, give:

$$\begin{aligned}\Delta - \Delta^{\text{no surprise}} \quad = \quad & \sum_{i\in I_U} P(T_H|F_U, T_U, Y^i_U)\, P(Y^i_U)\, P(T_U)\frac{P(\overline{\exists_U})}{P(\exists_U)} \\ & - \sum_{j\in I_{\overline{U}}} P(T_H|\overline{F_U}, \overline{T_U}, Y^j_{\overline{U}})\, P(Y^j_{\overline{U}}) \left(P(\overline{T_U}|\overline{F_U}) - P(\overline{T_U})\right) \qquad (4.22)\end{aligned}$$

Using (4.1), we can spell out

$$P(T_U)\frac{P(\overline{\exists_U})}{P(\exists_U)} = P(T_U)\frac{1 - P(T_U) - P(\overline{T_U})(P(\exists_U|\overline{T_U})}{P(T_U) + P(\overline{T_U})(P(\exists_U|\overline{T_U})} \tag{4.23}$$

Now we are in the position to understand the limits that maximize confirmation:

- **Small $P(\exists_U|\overline{T_U})$:** In the limit of vanishing $P(\exists_U|P(\overline{T_U}))$, we find

$$P(T_U)\frac{P(\overline{\exists_U})}{P(\exists_U)} \to 1 - P(T_U) \tag{4.24}$$

- **Small $P(T_U)$:** If we take the small $P(T_U)$ in addition, we get

$$P(T_U)\frac{P(\overline{\exists_U})}{P(\exists_U)} \to 1 \tag{4.25}$$

- **A small expected number of unified theories:** In the limit where MEA enforces $i = 1$, we find

$$\sum_{i\in I_U} P(T_H|F_U, T_U, Y_U^i)\, P(Y_U^i) \to 1 \tag{4.26}$$

  A small number of unified alternative theories is of course in line with a small $P(\exists_U|P(\overline{T_U}))$. However, as discussed above, it can happen that finding one unified theory substantially increases the probability that there exist others. We will point at one such case in the Section (5).

- **A large expected number of non-unified theories:** in the limit of a large number of non-unified theories, we find

$$\sum_{j\in I_{\overline{U}}} P(T_H|\overline{F_U}, \overline{T_U}, Y_{\overline{U}}^j)\, P(Y_{\overline{U}}^j) \to 0 \tag{4.27}$$

and the entire second line of Equation (4.22) vanishes.
If we take all of those limits, we find the maximal confirmation

$$\Delta - \Delta^{\text{no surprise}} \to 1 \tag{4.28}$$

The typical cases where our scenario applies are those where there is a significant degree of surprise, and a significantly smaller expected size of unified theory space than of non-unified theory space:

$$\sum_{i\in I_U} P(T_H|F_U, T_U, Y_U^i)\, P(Y_U^i) < \sum_{j\in I_{\overline{U}}} P(T_H|\overline{F_U}, \overline{T_U}, Y_{\overline{U}}^j) P(Y_{\overline{U}}^j) \tag{4.29}$$

In those cases, we get significant confirmation since the first line in (4.22) gets a significant boost due to the surprise effect while the second line is suppressed due to the larger theory space for non-unified theories (which also implies that $\overline{F_U}$ - finding a non-unified theory - will increase $P(\overline{T_U}|\overline{F_U})$ over $P(\overline{T_U})$ only to a modest extent). We have established that unification can have a significant confirmatory effect that is structurally different from setting priors. It is an effect that is *based* on setting the prior $P(T_U)$ but plays out as a genuine updating effect. Therefore, it is a genuine confirmation effect of unification.

4.4 A quantitative toy-example of the surprise-induced confirmation value of unification

In the following, we provide a quantitative toy example of what has been discussed. For the sake of simplicity, we do not introduce a full theory space account, which would sum over all possible hypotheses about numbers of alternatives, but assume a fixed number of theories. This simple model does not allow for updating under empirical or meta-empirical evidence. It therefore will only provide a rough assessment of credences, but it is sufficient for understanding the basic mechanism. We assume random pick of theory H from the overall set of theories and random pick of the viable theory from the subsets of i unified respectively j non-unified theories. Let us first model a scenario **Prior** with a prior preference of unified theories by a factor 2 without any surprise factor (i.e. $P(\exists_U|\overline{T_U}) = 1$). That is, we assume

$$P(T_{H_U}) = 2P(T_{H_{\overline{U}}}) \tag{4.30}$$

Let us further assume that scientists have credence $P(T_H) \simeq 1/10$ in the viability of a given hypotheses H. In terms of specific numbers of unified and non-unified theories, our quantitative assumptions translate into

$$\begin{aligned} i = 1, \quad j = 9, \quad P(T_U) = P(T_{H_U}) &= 2/11, \quad P(T_{H_{\overline{U}}}) = 1/11 \\ P(T_H) = \frac{1}{10}\frac{2}{11} + \frac{9}{10}\frac{1}{11} &= 0,1 \end{aligned}$$

Since we assumed $i = 1$, we have $P(T_H|F_U, T_U) = 1$. Equation (4.8) then gives

$$P(T_H|F_U) = 0.18 \tag{4.31}$$

Let us consider a different scenario **Surprise** where we have a substantial surprise factor, represented by $P(\exists_U|\overline{T_U}) = 1/10$. To get a more realistic account, we now implement "by hand" the requirement that credence in there being two unified theories after having found one is not different from what the credence in finding a non-viable unified theory had been before finding a unified theory. That is, we assume $P(Y_2^U|F_U) = P(Y_1^U|\overline{T_U})$, which gives a modified $P(T_H|F_U, T_U) = \frac{1}{1+(0.1)\frac{1}{2}}$. Equation (4.8) then gives

$$P(T_H|F_U) = \frac{1}{1.05}\frac{2/11}{2/11 + 0.1(9/11)} = 0.66 \tag{4.32}$$

In **Surprise**, the surprise effect generates very substantial confirmation value of unification. The high confirmation value of the surprise factor compared to the selection of priors demonstrates that the element of surprise is an important component of confirmation that is conceptually distinct from a mere specification of priors. The discussion also shows, however, that same degree of confirmation could in principle also be achieved in a no-surprise scenario by drastically increasing prior credence in unification. In the next section we will discuss scientific examples to demonstrate that the surprise factor does play an important role in science.

5 Examples

We discuss two kinds of examples in this section. First, we are interested in cases of unification which provide epistemic support for the theory in question that does rely on the surprise factor. Such cases would be at variance with Myrvold's view that only mutual information unification can be a basis for confirmation. The most clear-cut examples would be those where mutual information confirmation does not apply at all. Most often, however, there will be an element of mutual information confirmation. What we aim to illustrate is that it is helpful to acknowledge that confirmation value is generated in other ways as well. It thus suffices for our purposes to find examples where the element of mutual information is insufficient for explaining the overall confirmation value of unification in the given case. Second, we discuss two cases where common origin unification lacks confirmation value because the element of surprise is missing. Such examples are most convincing if parts of a scientific community can be argued to have taken common origin unification as an argument in the unifying theory's favor, but were later disappointed because a non-unified theory turned out to be viable.

5.1 Cases with confirmation through unification

5.1.1 Newton and Kepler according to Whewell

Let us return to Whewell's perspective on the confirmation of Newton's theory. In line with Janssen, our account identifies confirmation due to the unificatory character of Newton's theory. This distinguishes Newton's theory from the less unified account by Kepler, despite the fact that both accounts are empirically equally well confirmed at the time of Newton. In line with Blanchard's reasoning, we take Newton and Kepler to be on equal footing with respect to mutual information confirmation. The difference between the two accounts can be understood in terms of common origin unification: while Newton provides the latter, Kepler does not. In this light, we agree with Janssen that the Newton case demonstrates confirmation

value of unification that reaches beyond what Myrvold is willing to accept. The confirmation value of unification in the given case is based on the surprise factor.

As discussed in Section 3.2, the role of the surprise factor in the given case is emphasized by Whewell himself. On his account, the unification provided by Newton's theory had confirmation value specifically because it was by no means clear that such a unified theory could be constructed at all. The fact that the data available at the time was of a kind that allowed for such a unified theory could be explained by the hypothesis that the true theory was unified. If that unified theory ended up not being viable beyond the set of observed planets, it would have been difficult to explain why it worked with respect to the data that had initially been available. It would just seem overly improbable that the planets moved in agreement with a unified theory in so many per se independent respects if the true theory were not unified. This consideration increased the credence in the hypothesis that the true theory of planetary movements was unified, which in turn increased credence in the unified hypothesis Newton had developed.

Representing the argument in terms of our model of Section 4, all three criteria for finding confirmation value of unification are fulfilled. The surprise factor that corresponds to a low $P(\exists_U|F_U^H)$ is explicitly emphasized by Whewell. $P(T_U)$ is not very high since it is by no means clear that a unified theory exists. But setting $P(T_U)$ above what would be implied by random pick is justified because Newton as well as many observers did favor unification. Finally, Newton's theory did not provide a recipe for developing other promising unfied theories. Therefore, chances of finding further promising unified theories were not much increased by finding Newton's theory. Altogether, the scenario thus provides a convincing case of confirmation value of common origin unification on our model.

5.1.2 A counterfactual retelling of the discovery of special relativity

Let us construe a slightly counterfactual retelling of Einstein's development of the special theory of relativity. The counterfactual element in our story does not strengthen the non-mutual information element of the confirmation value. It just reduces the mutual information element in order to make the former more visible.

Before Einstein's theory of special relativity, Maxwell's electromagnetism and Newtonian mechanics were understood to be compatible theories about different sets of phenomena that were based on entirely different principles. Einstein's motivation for searching for a theory of relativity was in part based on his dissatisfaction with this state of the art. His theory of special relativity provided a fully unified perspective on the two theories that allowed to understand them as the low velocity and high velocity limits of the full theory. In the actual historical process that led up to the development and testing of special relativity, the Michelson-Morley experiment did have an empirical bearing on the status of special relativity. One might nevertheless ask the question whether, in the absence of this empirical support for special

relativity, the theory would still have been taken to be substantially supported by the mere fact that it provided a full unification of formerly conceptually independent theories. It is plausible to assume that the theory would have had such support in the eyes of Einstein himself, who de-emphasized the role Michelson-Morley had played in his thinking (Van Dongen 2010).

How could confirmation value of unification be generated in described scenario? When it was developed, the theory of special relativity reproduced tested predictions of Newtonian mechanics and electromagnetism, but predictions going beyond these two theories had not yet been tested. Special relativity therefore would not have gained any new empirical support at the time. Nor would there have been any potential for mutual information confirmation based on correlations that remained unexplained by the original set of theories, but found an explanation based on the unified theory. No measured parameter values of Newtonian mechanics or Maxwell's electromagnetism were explained by special relativity on the basis of observed parameter values of the other theory. Nor were any empirically non-equivalent alternatives to Newtonian mechanics or Maxwell's electromagnetism considered at the time that could be ruled out on the basis of endorsing special relativity.

What would have spoken strongly in favor of the theory's viability was the surprising existence of a unified theory. Since Newtonian mechanics and Maxwell's electrodynamics were based on seemingly entirely different and mutually inconsistent principles, it was by no means to be expected that the two theories could be covered by one unified description. Einstein's discovery that there indeed was a way to unify classical mechanics and electrodynamics therefore did have the surprise factor needed to increase the credence in special relativity.

In terms of our model, all conditions for having confirmation value due to the surprise effect are fulfilled. As argued above, there was a significant surprise factor associated with finding a unified theory that covers Newtonian mechanics and Maxwell's theory. Still, unification was considered a desirable feature, which licenses putting $P(T_U)$ significantly above what would be suggested by random theory pick from the entire theory space. And, even more clearly than in the Newton case, special relativity does not increase credence in the existence of alternative unified theories of Newtonian mechanics and electrodynamics. Another aspect is of specific importance in the given case: Before Einstein, there was a very high degree of trust both in Newton's and in Maxwell's theory. In other words, prior theory space was strongly constrained for non-unified theories. The increase in credence in the given case therefore works based on the strong surprise factor rather than through the size of non-unified theory space.

5.1.3 GUT Theories

Our third example plays out in the context of high energy physics. The standard model of particle physics introduces the non-simple gauge group SU(3) × SU(2) × U(1) to characterize

the gauge structure of strong and electro-weak interaction. In 1974, it was understood by Georgi and Glashow that the standard model particle spectrum fell into representations of larger simple gauge groups, such as SU(5) and SO(10). If one assumed that one of those large simple gauge groups actually is physical in the sense that the corresponding gauge fields and matter particles exist, one is led to a theory of grand unification (GUT). If the theory is correct, the GUT group is a gauge symmetry group of our world.

In our example, we will focus on this one piece of evidence in favor of GUT. The case of GUT unification plays out at a different level than the cases of Newtonian gravity or special relativity. It is not about building a new theory but about model building within a given theoretical framework. Therefore, unlike in the case of full-fledged theory building, the degree of surprise can be quantified based on the theory's structure. On the one hand, this means that we don't need to frame confirmation in terms of unification. Since the specific form of unification is determined by the understanding of what it means to have a unified theory in the given sense, we can understand the confirmation value of the data in terms of the constraints grand unification puts on the observable data. But, given that there is an intuitive way of grasping the situation in terms of unification, we *can* frame confirmation value in terms of unification. And for the very reason that the analysis is embedded in a specific theoretical framework, the example shows the structure of the surprise argument with particular clarity.

The prior probability for finding a particle spectrum that can be put in full GUT group representations, assuming plausible physical conditions, can be assessed to be around 3.2% (Herms and Ruhdorfer 2024, Dawid and Wells 2024). Mutual information unification does not seem be the right framework for understanding what is going on in this case. What is unified are electroweak U(1) and SU(2) and strong SU(3) interaction. But it seems akin to putting the horse before the cart to view the situation in terms of extracting information about strong interaction from what we know from electroweak interaction based on a GUT theory or vice versa. It is the fact that many particle spectra would not allow for a simple unified gauge group at all that makes the representation-based arguments for GUT convincing. It is not the idea that we can extract mutual information correlations between the non-unified standard model gauge groups from assuming GUT.

Pointing at the surprise factor thus is the more adequate way to understand why we get increased credence in GUT from the standard model particle spectrum. Only a small share of possible particle spectra can be fit into GUT representations. Therefore, it seems a priori unlikely that grand unification is possible. The observation of the particle spectrum we actually observe therefore comes at a surprise that, along the lines of our proposed scheme, generates credence in the GUT hypothesis. Applying our model to the GUT case provides the following story of GUT confirmation. In our model, H is the gauge field theory selected in light of collider data. $\exists_U$ denotes the statement that there exists a way to put the particles found experimentally into full representations of a GUT group. $P(\exists_U)$ is 3.5% . F_U denotes the observation that particles actually found in collider experiments do fall into full GUT

representations. T_U denotes the statement that the viable theory is a GUT theory.

Physicists do not a priori expect the known particle spectrum to fit into any GUT representations. They have a candidate GUT theory (SU(5)), and after checking the details, find that they actually do. There is a surprise factor that, as discussed, can be spelled out quantitatively. Setting $P(T_U)$ above the random pick value amounts to attributing, as any physicist would, a small but significant prior probability to GUT before checking whether representations support the hypothesis. (Random pick, to the contrary, would amount to assigning to any of the huge number of gauge structures that would be possible before checking the characteristics of the known elementary particles the same prior probability than a grand unified gauge structure. This would correspond to attributing an excessively small prior probability to GUT.) In the given case, finding SU(5) did lead towards other GUT groups that work as well for the observed particle spectrum. So there is an a posteriori unified theory space that is larger than what may have been assumed a priori. This reduces credence in a specified GUT group, but does not destroy the confirmatory effect of unification since the number of possible non-GUT group structures is much much higher. If we treat the general hypothesis of grand unification rather than the specification of an individual gauge group as the theory, we by definition set $P(H) = P(T_U)$, so that the process reduces to Inference I. To provide a quantitative example, setting $P(T_U) = 0.03$ would then, based on equation (4.7), lead to $P(T_U|F_U) = 0.49$. In the absence of surprise, we would have $P(T_U|F_U) = P(T_U) = 0.03$. We thus see a very substantial confirmation effect of surprise.[9]

5.2 Cases of unification without confirmation

5.2.1 Indoeuropean languages

Let us now discuss an example where unfication was considered an attractive goal but is seen more critically today in light of new evidence. In historical linguistics, it has long been an important goal to understand the mechanisms that led to the wide distribution of Indo-European languages. It was long considered an attractive feature of any such explanation if it related the said expansion of the Indo-European language to one of the other striking cultural expansion processes that occurred in a comparable time-frame. Two candidates of such unified theories gained considerable popularity. One approach, discussed for example in Bouckaert et al. (2012), linked the distribution of language to the expansion of agriculture that started from Anatolia around the 6th millenium b.c. Another hypothesis, presented for example in Anthony (2007), assumed that the spread of Indo-European languages was propelled by the emergence of horse riding in the Eurasian steppe around 3000 BC, which for the first time allowed peoples from the steppe to travel very long distances in very brief

[9]Note that out analysis of the GUT even more strongly than the others relies on lifting the omniscience assumption of the agent. The agent "observes" features of group structure that would be known from the start to an omniscient observer.

periods of time.

Both of these explanations gained appeal by offering a unified story that linked the spread of a dominant language group to another important pre-historical process of cultural spread. The proposed hypotheses qualify as common origin unification because they represent the dominance of one language family as implied by a cultural achievement that was capable of propelling its own spread (agriculture making communities better nourished or horse riding making communities militarily more effective). The hypotheses would have merited confirmation value due to their explanatory quality on Janssen's and Blanchard's account. The hypotheses also amounts to (modest) mutual information unification: based on any of the two unifying hypotheses, the fact that there was one cultural group from which horse-riding or agriculture started respectively, makes the emergence of a dominant language group more likely.

In recent years, advanced genetic analysis of human and horse genomes (Heggarty et al. (2023) has rendered both described hypotheses more doubtful. The results of genetic analysis has been argued to suggest that a more complex explanation that does not rely on a unified theory of one main prehistoric processes of cultural spread is needed. If so, the viable theory would be less unified at the given level of description than what many experts had expected earlier on.

This paper's analysis offers a meta-conceptual reason why one might have been a little wary of the proposed unified solutions from the start. There is an interest in and preference for unified theories, which would justify setting (T_U) above the random pick value. But arguably, the flexibility of theory building due to weak constraints by data and conceptual cogency rendered the existence of unified theories less than surprising in the given context. It was to be expected that, with sufficient care and diligence, one could come up with a unified solution that seemed to agree with the data. So there was no convincing surprise factor associated with finding a unified theory. Therefore, even though the unified theories did represent cases of mutual information and common origin unification, their unifying power did not generate significant confirmation value on our account. It did not lower credences in the viability of non-unified alternatives. Those non-unified alternatives then ended up being supported by more constraining empirical evidence.

5.2.2 Conspiracy theory

For our second example, we return to the case of conspiracy theories. Keeley (1999) in his classic treatment on conspiracy theories states that "[t]he first and foremost virtue which conspiracy theories exhibit, and which accounts for much of their apparent strength, is the virtue of unified explanation" (p. 119). While Keeley does not address the virtue of unification as problematic per se, his arguments against conspiracy theories focus on the general skeptical attitude of conspiracy theorists and in believing in an "ordered universe" (p. 123) at all costs.

We however would challenge already the virtuousness of unified explanation in the case of conspiracy theories.

Viewed from the perspective we have developed, we can understand the role of unification in that context in the following way. It is, on our account, part of scientific and non-scientific reasoning alike, to have some degree of preference for unification that translates into having $P(T_U)$ above the random pick value. The mistake attributable to the conspiracy theorist at that level merely amounts to making that preference too strong. In many cases, the conspiracy theorist underestimates the epistemic risk associated with connecting lots of different phenomena that might just as well be unrelated to each other. So, while the conspiracy theorist is justified in putting $P(T_U)$ above the random pick value, it should nevertheless be fairly low.

The conspiracy theorists' core mistake, however, is related to assessing the surprise factor. Indeed, there is some form of surprise involved in the endorsement of conspiracy theories. The willing recipient of conspiracy theories is surprised that the theory in question does unify so much, and accepts this as a reason to believe the theory: if all those aspects fit together so surprisingly well based on the conspiracy theory, there must be something to the theory. But the surprise the conspiracy theorist is pointing at does not rise above the level of seeing something one personally did not expect to see. On closer inspection, the conspiracy theorist's use of the surprise argument is based on a fallacy. Technically, this fallacy stems from an inadequate view of the spectrum of possible theories. To generate confirmation value, the surprise factor would need to be based on a low $P(\exists_U|\overline{T_U})$. But there is no basis for assuming a low $P(\exists_U|\overline{T_U})$ in the case of conspiracy theories. The tools for building such theories are so flexible that one should assume $P(\exists_U|\overline{T_U})$ to be very close to one. Therefore, the unifying power of the conspiracy theory, though representing common origin unification, has no confirmation value on our account that goes beyond the slightly above random pick prior attributable to a unified theory. Well crafted conspiracy theories therefore are a case where Jannsen and Blanchard are forced to concede substantial confirmation through common origin unification in a way comparable to unified *scientific* theories, while our account allows to grasp the epistemic difference between the two cases in terms of the surprise factor.

6 Conclusion

We have demonstrated in this paper that unification has confirmation value that goes beyond what is covered by Myrvold's mutual information confirmation. If the agent takes it to be a priori unlikely that a unified theory of a certain kind can be developed, finding such a unified theory suggests that the available data is distributed specifically in a way that allows for such a unified theory. This constellation can be explained by the hypothesis that the true theory about the subject matter is unified.

Our analysis does not merely show that confirmation value of the given kind can arise. It demonstrates that, if i) we consider unification a possibility that deserves attention and ii)

there is a priori doubt whether a unified theory exists at the given level, finding a unified theory must generate confirmation of the described kind.

The question remains: how plausible is it, generally speaking, to doubt that a unified scientific theory can be constructed about a given subject matter at a given level of description. The answer is: the observation that science often is descriptively successful forces us to have substantial doubts of this kind. The predictive success of science implies that very many imaginable regularity pattern of data cannot be reproduced by a scientific theory. If most of them were, there would be no reason to expect that the theory we have developed, rather than one of the alternative theories that cover most of the space of possible data patterns, gives the correct predictions of future data collection (Dawid 2013, Sect. 2.2). But if many regularity patterns cannot be represented by a scientific theory, one must expect that many attempts at unification of two disjoint data sets at a certain level of description cannot be realized either. In other words, our experience of the predictive success of science forces us to have the kind of doubt about unification that, in turn implies confirmation value of unification.

Still, not all contexts of reasoning generate surprise about the existence of a unified theory. If that surprise is absent or is not justified based on an adequate assessment of theory space, as in the two cases discussed in Section 5.2, no scientific confirmation value is generated. Surprising unification therefore constitutes a specific and important class of confirmatory unification scenarios.

As a final thought, let's take a step back for a more general look at the structure of our argument. The reader may have noticed that nothing in our argument hinges on unification per se. In order to generate confirmation based on the element of surprise, we just need some characteristic X of a theory so that i) an observer has legitimate reasons to set a low value for $P(\exists_X|\overline{T_X})$ and ii) an observer can plausibly posit X as a feature of the actual world. Both conditions are under certain conditions fulfilled for the characteristic of being unified, since: i) there are contexts where the existence of a unified description at the given level is by no means to be expected and ii) it is a plausible hypothesis that the world (or a given part of it) has a unified description. Therefore, unification is a suitable case for applying the argument of surprise.

However, one might find other characteristics of a theory that can be confirmatory along the same lines. What we have proposed is a general mechanism by which we can judge the conditions under which a characteristic of a theory can have epistemic significance, thereby impacting the theory's viability. We see this proposed mechanism as in line with the efforts involved in solving the new demarcation problem, which aims to distinguish legitimate from illegitimate influences of values in science (Holman and Wilholt 2022). It is beyond the scope of this paper to analyse whether other examples where the argument of surprise generates substantial confirmation value do exist. However, a superficial search suggests that convincing examples are more difficult to find than one might initially think.

Appendix A An analysis of the random pick case

We prove that the general random pick scenario generates no confirmation value for unification. The general random pick scenario amounts to giving $P(T_U)$ the value it would have if we took it to be determined based on picking the viable theory from the joint pool of unified and non-unified theories.

Let us call the overall number of theories l, Let us further call the number of those theories that are unified i. Modelling a belief structure regarding the number of theories would then involve individual hypotheses Y_l^i: "There are l possible scientific theories that are in agreement with the available data, out of which i theories are unified." We define $P(Y_l) := \sum_{i \in I_U} P(Y_l^i)$. Let us now assume $P(T_H) = 1/l$. On that assumption, T_U cannot be freely chosen anymore but is controlled by the random pick principle. Assuming that prior credence as to whether the theory developed by scientists does not favor unified theories, puts it under the control of random pick as well. We thus have:

$$P(T_U) = \sum_l \left(P(Y_l)(l^{-1}) \left(\sum_{i \in I_U} \frac{P(Y_l^i)}{P(Y_l)}(i) \right) \right) = P(F_U). \tag{A.1}$$

Equation (A.1) implies that the probability that a theory is unified, respectively that the developed theory is unified, (that is controlled by the number i), is not correlated with the probability that a theory is viable (that is controlled by the number l), since the probability of being a unified theory and the probability of being the viable respectively the developed theory are all represented by random picks from the same full theory space. Due to $P(T_U) = P(F_U)$, we have

$$P(T_H, F_U) = P(T_H)P(F_U) = P(T_H)P(T_U) = P(T_H, T_U) \tag{A.2}$$

We can now write equation (4.4) as

$$\begin{aligned} P(T_H|F_U) &= P(T_H|F_U, T_U)P(T_U|F_U) && \text{(A.3)} \\ &= \frac{P(T_H, F_U, T_U)}{P(T_U, F_U)} \frac{P(T_U, F_U)}{P(T_U)} = \frac{P(T_H, F_U)}{P(T_U)} = \frac{P(T_H)P(T_U)}{P(T_U)} && \text{(A.4)} \\ &= P(T_H) && \text{(A.5)} \end{aligned}$$

using equation (4.3) and equation (A.2) on the way. Thus, we see that the full random pick scenario does not allow for any confirmation value of unification.

References

Anthony, David W. *The Horse, the Wheel, and Language*. Princeton: Princeton University Press, 2007. https://doi.org/10.1515/9781400831104.

Blanchard, Thomas. "Bayesianism and Explanatory Unification: A Compatibilist Account." *Philosophy of Science* 85, 4 (2018): 682–703.

Bouckaert, Remco R., Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell David Gray, Marc A. Suchard, and Quentin Douglas Atkinson. "Mapping the Origins and Expansion of the Indo-European Language Family." *Science* 337 (2012): 957 – 960. https://api.semanticscholar.org/CorpusID:36512809.

Castellani, Elena. "Convergence strategies for theory assessment." *Studies in History and Philosophy of Science* 104 (2024): 78–87.

Currie, Adrian. "The argument from surprise." *Canadian Journal of Philosophy* 48, 5 (2018): 639–661.

Dardashti, Radin. "Physics without experiments?" In *Why Trust a Theory?*, edited by Dawid R. Dardashti, R., and K. Thébault. Cambridge University Press, 2019, 154–172.

Dawid, Richard. *String theory and the scientific method*. Cambridge University Press, 2013.

———. "Modelling non-empirical confirmation." *Models and inferences in science* 191–205.

———. "Delimiting the Unconceived." *Foundations of Physics* 48, 5 (2018): 492–506.

———. "Does the No Alternatives Argument Need Gerrymandering to Be Significant?" *British Journal for the Philosophy of Science* 717081.

Dawid, Richard, Stephan Hartmann, and Jan Sprenger. "The no alternatives argument." *The British Journal for the Philosophy of Science* 66, 1 (2015): 213–234.

Dawid, Richard and James Wells "A Bayesian Model of Credence in Low Energy Supersymmetry" *philsci-archive/24172*

French, Steven, and Alice Murphy. "The value of surprise in science." *Erkenntnis* 88, 4 (2023): 1447–1466.

Heggarty, Paul, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechoslaw Jocz, Martin Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard Strand, Geoffrey Haig, Martin Macák, Ronald Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Dewey-Findell, and Russell Gray. "Paul Heggarty, Cormac Anderson, Denise Kühnert, Russell D. Gray [18 co-authors, Matilde Serangeli, 9 co-authors]. Languages trees with sampled ancestors support an early origin of the Indo-European languages." *Science* 381 (2023): 414, eabg0818.

Herms, Johannes, and Maximilian Ruhdorfer. "GUTs–how common are they?" *arXiv preprint arXiv:2408.11089*.

Holman, Bennett and Wilholt, Torsten. "The new demarcation problem." *Studies in history and philosophy of science* 91, (2022): 211–220.

Horwich, Paul. *Probability and evidence*. Cambridge University Press, 1982.

Janssen, Michel. "COI stories: Explanation and evidence in the history of science." *Perspectives on Science* 10, 4 (2002): 457–522.

Keeley, Brian L. "Of Conspiracy Theories." *The Journal of Philosophy* 96, 3 (1999): 109–126.

Lange, Marc. "Bayesianism and unification: A reply to Wayne Myrvold." *Philosophy of Science* 71, 2 (2004): 205–215.

Morgan, Mary S. "Experiments versus models: New phenomena, inference and surprise." *Journal of Economic Methodology* 12, 2 (2005): 317–329.

Myrvold, Wayne C. "A Bayesian account of the virtue of unification." *Philosophy of Science* 70, 2 (2003): 399–423.

———. "On the evidential import of unification." *Philosophy of Science* 84, 1 (2017): 92–114.

Planck, Max. *Wege zur physikalischen Erkenntnis: Reden und Vortäge*. volume 1. Hirzel, 1944.

Schupbach, Jonah N. "On a Bayesian analysis of the virtue of unification." *Philosophy of Science* 72, 4 (2005): 594–607.

Van Dongen, Jeroen. *Einstein's unification*. Cambridge University Press, 2010.

Whewell, William. *History of the inductive sciences: from the earliest to the present time*. volume 1. D. Appleton, 1863.