Preprint; final version forthcoming in *The Philosophical Quarterly*

# Laws of Nature as Results of a Trade-Off — Rethinking the Humean trade-off conception

Niels Linnemann* and Robert Michels†

March 5, 2025

**Abstract**

According to the standard Humean account of laws of nature, laws are selected partly as a result of an optimal trade-off between the scientific virtues of simplicity and strength. Roberts and Woodward have recently objected that such trade-offs play no role in how laws are chosen in science. In this paper, we first discuss an example from the field of automated scientific discovery which provides concrete support for Roberts and Woodward's point that scientific theories are chosen based on a single-virtue threshold. However, we then use this very same example as a starting point to argue that i) by insisting on a single best theory, Humeans rely on an overly simplistic conception of trade-offs, that ii) this conception should give way to one which allows for a Pareto front of equally optimal theories, and iii) that given this new conception, threshold behaviour for a virtue like strength is a) compatible with the existence of a genuine trade-off and b) can even play the positive role of a selection criterion.

**Keywords**: Laws of nature, Humeanism, Best Systems Analysis, Trade-off between theoretical virtues, Automated Scientific Discovery, Symbolic Regression

## 1 Introduction

Humeans claim that laws of nature gain their status as such through a pragmatic trade-off based on theoretical virtues of true theories which cover all non-modal facts about our universe. According to the classical version of the theory, the Best Systems Analysis, *BSA* for short, true generalizations about our universe express laws if they are axioms or theorems of the best systematic theory about it, where the best such theory offers the best balance between simplicity and descriptive strength (cf. Lewis (1973, 1994)). (We will focus on non-probabilistic laws in the following. If probabilistic laws are taken into account, the Humean best system furthermore takes a theory's fit with the actual distribution of chances related to probabilistic processes into account.)

Unlike rival, anti-Humean theories—most notably Dispositional Essentialism and the DTA-view—BSA-style Humeanism does not rely on purely philosophical posits

*Department of Philosophy, University of Geneva, Switzerland, e-mail: niels.linnemann@unige.ch
†Lancog, Centre of Philosophy, University of Lisbon, Portugal, e-mail: mail@robert-michels.de

like essences, dispositions, or universals to explain what a law of nature is. Rather, it promises to build on pragmatic factors which are, so at least the standard story goes, drawn from science. The alignment with (supposedly) central pragmatic scientific virtues gives the Humean theory of laws a certain philosophically conservative naturalistic flair which also many philosophers of science find appealing. A further feature which *supposedly* appeals to naturalistically minded philosophers is that Humean accounts are congenial to a particular view about how laws are established in science; namely that scientists produce a number of alternative scientific theories to explain the fundamental facts and processes of the natural world, and then proceed to choose one of these theories based on pragmatic considerations, leading to a trade-off between the virtues of these theories.

In general, one might take the Humean analysis of lawhood to have a naturalistic side (whether a true universal claim about the natural world is a law directly depends on whether that claim is an axiom of one of several equally true best systems, where these are usually taken to correspond to actual scientific theories, or idealized 'final' theories depending on what kind of naturalism one subscribes to; we will not discuss general worries about naturalism in this paper); and a pragmatic side (which one of these scientific theories ultimately provides the laws depends on a pragmatic choice between these theories based on the mentioned trade-off).

However, the match between the Humean picture and actual science has recently been questioned. Woodward (2014) and Roberts (2008) have argued that there are often no genuine trade-offs between simplicity and strength in science, since the latter is clearly prioritized until a certain minimal threshold of strength is reached. This argument then appears to reveal a tension between the two sides of the BSA, since it suggests that the pragmatic side has no naturalistic basis.

In this paper, we first present recent findings from the field of automated scientific discovery which back up Woodward's and Roberts's point that in scientific theory choice, what matters, first of all, is achieving a certain amount of descriptive strength (provided that strength is understood as accuracy—a conception of strength that the traditional proponent of the BSA might take issue with). We thereby corroborate their point, which has so far been supported through introspective reflections by authorities such as Newton and Einstein on their research methodology (Woodward, 2014) and under recourse to a toy model drawn from scientific practice (Roberts, 2008).

Secondly, we use this example to argue that the conception of a trade-off involved in the BSA is too simplistic, independently of which notion of strength is ultimately accepted as salient for the trade-off. Trade-offs in science are operationalized through a Pareto front of equally optimal theories, rather than a single optimal theory as proponents of the BSA have it.

Third and finally, we show that a notion of strength which exhibits threshold behaviour is not just fully compatible with a coherent trade-off conception of laws (contra Roberts and Woodward), but also that it provides us with a way to single out theories from other theories which would usually count as offering equally good trade-offs. We argue that this latter point can be generalized beyond the context of the BSA: threshold behaviour with respect to a theoretical virtue does not rule out a trade-off between it and other theoretical virtues.

We proceed as follows. In the second section, we briefly present the naturalistic adequacy challenge posed by Roberts and Woodward for the BSA. In the third section, we survey recent work in automated scientific discovery on the identification of laws based on physical data, focusing in particular on very recent work on symbolic regression. In the fourth section, we discuss the fit of this example, in particular the notion of a trade-off involved, with the BSA and then elaborate on how it supports the case for a strength-threshold. In the fifth section, we argue first, that the notion of a trade-off involved in discussions of the BSA is overly simplistic (subsection 5.1). Secondly, we show, that the existence of a threshold with respect to one among a number of virtues of a scientific theory does in general not undermine the idea that the choice between these theories results from a trade-off between these virtues (subsection 5.2). Finally, we reconsider Roberts's and Woodward's objections in light of our findings (subsection 5.3). In the sixth section, we wrap up the discussion and suggest a challenge for alternative theories of the laws of nature which Humeans could raise once their conception of a trade-off is in good standing.

## 2 A central adequacy challenge to Classical Humeanism

Can friends of the BSA rightfully claim that the way laws are chosen in science corroborates their lawhood-criterion since this process involves a trade-off between simplicity and strength? One argument against this claim relies on the observation that theory choice in science often involves a strength-threshold as opposed to a free trade-off between the two virtues.

Roberts raises this point in the context of a discussion of a toy example, in which a scientist has to decide whether to adopt a theory consisting only of Kepler's laws, or one which in addition specifies the positions of all planets throughout the history of the universe:

> ...scientists don't ever face a choice like the choice between the two systems in the toy example. ...Scientists might have to face the question of whether they have sufficient evidence to justify accepting the stronger theory, or whether they should be more conservative and merely accept the weaker theory. *But this is a judgment about the strength of the available evidence, and not a judgment about the competing theoretical virtues of the two systems.* Scientists just don't ever need to make the kinds of trade-offs that it must be possible to make if the best-system account of laws is correct. ((Roberts, 2008, 9); emphasis added)

Woodward has furthermore pointed out that at least some scientists, including Newton and Einstein, appear to have endorsed a view according to which there is no genuine trade-off, in the sense that simplicity only begins to matter once a sufficient threshold of explanatory strength is reached by a theory. Einstein is, for instance, quoted by Woodward in support of this claim with the following passage:

> It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible *without having*

> *to surrender the adequate representation of a single datum of experience.* ((Einstein, 1934, 165); emphasis added.)

The claim in particular entails that, pace what the BSA suggests, a loss of strength can not always be outweighed by a gain in simplicity. (Cf. Woodward (2014), 101-102.)

Both Roberts and Woodward then provide variants of an argument that poses a serious problem for the BSA by undermining an important part of its motivation, namely its fit with deliberations about theory-choice in actual science. We will respond to this argument in section 5. Our approach will not be to deny Roberts' and Woodward's claim that there is a strength-threshold. Rather, we will in the following argue that the existence of such a threshold concerning e.g. strength does not in principle preclude a trade-off with another theoretical virtue. The former can instead be seen as an extremal case of the latter. Applied to the classical BSA: what looks like an undeniable strength-threshold, may just amount to a (very) strong preference for strength over simplicity *within a strength-simplicity trade-off*. Ultimately such cases can thus be subsumed by the trade-off view. More than that: we will see that these cases actually have a special significance, since they provide us with a criterion to choose between theories which are otherwise on a par with respect to a given set of theoretical virtues.) Roberts' and Woodward's arguments can hence be taken to be based on a wrong presupposition, namely that strength-thresholds exclude trade-offs with other theoretical virtues.

## 3 Symbolic regression

In this section, we introduce symbolic regression and highlight some recent achievements using this method in the recovery of natural laws. The main point of the section is that symbolic regression gives us a concrete example of how a trade-off between scientific virtues effectively contributes to the identification of laws. In the next section, we will then first discuss the analogy between this trade-off and the one envisaged in the context of the BSA. Based on this discussion, we will afterwards assess to which extent symbolic regression bears on the adequacy challenge raised by Roberts and Woodward.

Symbolic regression is a type of regression, i.e., an operation of finding a relation between a dependent variable, and one or several independent variables based on a training data set; the characteristic goal of symbolic regression is to find an explicit (typically differentiable) function composed from a wide set of presupposed functions that manages to express the relation of interest in an intelligible manner. Proponents of symbolic regression, Udrescu et al. (2020) for instance, describe symbolic regression as "discovering a symbolic expression that accurately matches a given dataset. More specifically, we are given a table of numbers, whose rows are of the form $\{x_1, \ldots, x_n, y\}$ where $y = f(x_1, \ldots, x_n)$, and our task is to discover the correct symbolic expression for the unknown mystery function $f$." (p. 1)

Contrast symbolic regression with a more familiar type of functional regression such as linear regression where the functional relation is presupposed to be captured by a certain type of function and where the specific parameters that pin down a

specific instance of that function type are determined by means of the training data (in linear regression, the functional relation is presupposed to be that of a linear function and only slope and offset are then to be determined). Or compare it to deep neural networks where the resulting functional relation is simply not transparent, although it can formally be seen as a composition of functions (namely, the activation functions associated with each layer).

Traditionally, the algorithmic core strategy behind symbolic regression has been to cleverly explore the space of possible functions—restricted by a fixed set of elementary functions out of which all other functions have to be composed—by mutating and crossing a set of initial sample functions; for this, one typically represents functions as tree structures (see figure 1 in Vyas et al. (2018)). Several libraries for symbolic regression in this sense exist by now, both open-source (PYSR, see Cranmer (2023)) and proprietary (eurequa®, QLattice®).

Now, the core computational problem behind such a strategy is NP-hard; it thus easily becomes practically unsolvable for large input data spaces. It is important to note though that the mere categorization of a problem as NP-hard does not render it unfeasible to solve in practical scenarios (see, for instance, Gamarnik et al. (2022) for a physics-based account of why this is so). In fact, pre-processing the training data with a neural net has recently led to significant advances in the context of symbolic regression: one fits the data with a neural network first and only then aims to obtain an explicit symbolic expression relative to the black box function associated with the neural net. Generally, this two-step procedure makes symbolic regression computationally (much more) feasible, as it is only applied to a specific functional relation within the training data (namely, that given by the black box function of the neural network). This functional relation already includes effective simplifications in the sense that the neural network model has itself been set up with an inductive bias (for technical details, see, for instance, Cranmer et al. (2020b)).

In some cases, the pre-processing even allows for applying recursive divide-and-conquer programming (rather than just the prima facie less directed genetic/evolutionary programming techniques) to find explicit symbolic expressions for the black box function. For instance, in Udrescu and Tegmark (2020); Udrescu et al. (2020), the differentiable black box function $f$, which has been set up by training a neural network, is explicitly guessed from various graph modularity heuristics, which are applied randomly and recursively. Part of the 'inductive bias' of the neural network that gets exploited by the graph modularity heuristics is the black box's function smoothness (which is entailed by the construction of the neural net). Take the graph modularity heuristic of 'compositionality': by testing whether the derivative of the function $\nabla f$ is proportional to that of another function $\nabla h$ one can check for compositionality of $f(x) = g(h(x))$ where $g$ is some other function. (If the test is successful, the problem of finding $f(x)$ becomes that of finding $g(y)$ with $y = h(x)$. Once this has been solved, one can obtain $f$ by composition again, i.e. $f = g \circ h$.) Other such modular strategies include multiple separability, additive separability, generalized symmetry.

Using this two-step procedure, impressive achievements which by far surpass early-day discovery automation algorithms à la Bacon have been made in the automated discovery of physical formulae and laws through symbolic regression. For instance,

Lemos et al. (2022) rediscover the specific force law behind orbital mechanics—that is, the Newtonian law of gravitation—and the masses of the involved bodies from orbital trajectories—under the presumption of a surrogate to a general Newtonian framework. The two-step procedure concretely takes the following form here: first, a black box neural network is trained to output accelerations of the celestial body given the relational distances between them, using data from over 30 years with temporal resolution of data points every 30 min. Distances (the input) and interactions (the output) are represented as edges of graphs; the graph structure is generally chosen such that it effectively implements a Newtonian framework: each node carries the mass of a body, and the rules for interaction, edges are analogous to rules for forces (*actio* = *reactio*, additivity of forces, ...). Note also that the masses of the bodies (values associated to nodes of graphs), as well as further internal weights of the net, are obtained through training. Then, in the second major step, PYSR is used to re-interpret the black box function of the network into interpretable mathematical expressions.

To name some other examples, Greydanus et al. (2019) demonstrate the discoverability of the specific Hamiltonian for mechanical systems under presumption of a general Hamiltonian framework, and Cranmer et al. (2020a) that of the specific Lagrangian under presumption of a general Lagrangian framework. (Just as one pre-encodes a Newtonian-type interaction structure through a graph in the work of Lemos et al. (2022), these two works respectively pre-encode the general Hamiltonian and Lagrangian frameworks into the net structure.) Operating at a more general level, Udrescu and Tegmark (2020); Udrescu et al. (2020) have managed to rediscover all sorts of physical formulae from respectively structured observation data sets.

Notably (and as already highlighted), in all these cases laws/formulae are discovered from data within a pre-given/structured context. One might even call for some stronger form of symbolic regression that itself suggests the relevant variables for the data at hand. But leaving this point aside, it is still far from clear whether more complex laws (say the field equations of general relativity, or Lagrangians in quantum field theory) can ever actually be found this way. One remaining reason to be sceptical about a straightforward recovery of more complex physical laws is that the formal structures used in physics become ever more mathematically and conceptually contrived, making it more difficult to come up with formalisations that can be fed into the neural network. In fact, proponents of symbolic regression in a physical context are careful to warn explicitly that their equation-finding systems require (partly significant) physical background assumptions.

## 4 Trade-offs in symbolic regression and in the BSA

Symbolic regression gives us a concrete example of the nature of and role played by trade-offs in scientific practice. Despite differences in algorithms and contexts of interests, researchers in the field of automated scientific discovery (Udrescu and Tegmark, 2020; Udrescu et al., 2020; Cranmer et al., 2020a) seem to effectively arrive at two insights: (i) simplicity and strength (in the sense of accuracy) for significant laws/formulas in physics stand in a trade-off relation in the sense of a 'Pareto front':

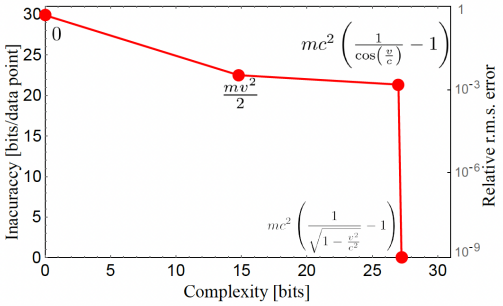


Figure 1: Our symbolic regression of data on how kinetic energy depends on mass, velocity and the speed of light discovers a Pareto-frontier of four formulas that are each the most accurate given their complexity. Convex corners reveal particularly useful formulas, in this case Einstein's formula and the classical approximation $mv^2/2$.

Figure 1: Taken from Udrescu et al. (2020) (with permission)

the significant laws/formulas are part of a subset of formulas that can neither be improved in accuracy without a loss of simplicity nor in simplicity without a loss of accuracy, i.e., they are all 'Pareto-optimal', and (ii) significant laws/formulas in physics correspond to the simplest formulae possible just before the level of accuracy *drops drastically* with a further increase in simplicity.

For a concrete illustration, consider figure 1 which shows the findings for kinetic energy for a data collection of physical trajectories by Feynman AI 2.0 by Udrescu et al. (2020): in the region demarcated by the first and second red dot from the left, there is a straightforward trade-off behaviour between accuracy and simplicity in the found formulas—the more accurate the formula, the less simple, and vice versa. Complexity/simplicity is quantified by the description length of the fitting expression (thus being a form of descriptive complexity/simplicity),[1] and strength in the sense of accuracy by the description length of the mean error with which the data is fitted. In the region demarcated by the second and the third red dot, the accuracy of the formulas found via symbolic regression stays approximately the same (i.e. they are on an accuracy-plateau) even when their simplicity decreases; in the region marked by the third and the fourth red dot, there is a rapid and sudden jump in the accuracy of the formulas after their simplicity has only decreased a bit.

These findings suggest the following interpretation: In the region demarcated by the first and third dot, one discovers a single relevant physical formula ($mv^2/2$). It can be characterised relative to a certain range of simplicity as the simplest formula with significantly higher accuracy than any simpler formula. In the whole diagram, there are two relevant physical formulas ($mv^2/2$ and $mc^2\left(\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}-1\right)$). Each is marked by being the simplest formula with a significantly higher accuracy than any simpler formula (for a certain range of simplicity).

To make things clearer, let us distinguish between two conjectures which we can formulate based on the presented findings in symbolic regression in the previous section:

**(CONVEX)** Physical formulae/laws correspond to the convex points in the trade-

[1] See also (Wu and Tegmark, 2019, 3-4)

off relationship between accuracy and simplicity, i.e. points at which accuracy has drastically increased after a mild increase in complexity; and

**(MULTIPLE)** Multiple physical formulae/laws corresponding to such convex points can be found relative to a sufficiently rich set of data, (in figure 1, one sees two such convex points—one associated with the non-relativistic mass formula, and another one with the relativistic mass formula).

We should of course note that these conjectures are tentative, since they are based on a snapshot of current work on symbolic regression and thus in need of further corroboration.

In sum, we take it that automated scientific discovery has the potential—in particular, subject to further developments in the field of symbolic regression—to support a picture of laws where strength and simplicity stand in a trade-off relation, but where, importantly, this relation is at the same time compatible with the existence of a strength-threshold. We will come back to this crucial observation in the next section.

Before we do so, we have to address a point which threatens to break the analogy between the way laws are rediscovered using symbolic regression and how the best system is selected according to the BSA. The point is that the trade-off between notions of strength (in the guise of accuracy) and simplicity (understood as descriptive simplicity) used in symbolic regression appear to be mismatched to the trade-off between homonymous notions Humeans rely on in their postulated trade-off.

This problem, to which extent the simplicity/strength trade-off in the BSA can be understood to be analogous to the simplicity/accuracy trade-off in regression scenarios (as we have it here with symbolic regression), has previously been discussed by Woodward. Concretely, he looks at the Akaike Information Criterion (AIC), a utility function which weighs accuracy (in terms of the log of the maximised value of the likelihood function, $L$), against simplicity (in terms of numbers of free parameters in the regression function, denoted by $k$) for a regression function $f$ with $k, L$ as $AIC(f(k, L)) = 2k - \log(L)$. Through minimisation relative to the regression function, the AIC is supposed to reveal the regression with the best predictive value.

Unpacking Woodward's criticism then, there are two problems for transferring this treatment of trade-offs from regression scenarios to the context of the BSA: (1) The reason why simplicity is considered in addition to accuracy in a regression scenario is to avoid overfitting. In contrast, the problem of overfitting theories to a restricted data set is not a concern for the classical BSA. The BSA posits strength and simplicity as factors in a trade-off between theories which are supposed to already apply to the Humean mosaic as a whole, i.e. to what one might think of as a complete data set about the actual universe (and in particular not a mere training data set). As Woodward writes:

> it is hard to see how the particular rationale that motivates the use of AIC—the need to avoid overfitting—applies in the case of the BSA. In the case of ordinary curve-fitting on noisy and incomplete data, it is clear how overfitting leads to less empirically accurate predictions on new data.

> Again, it is unclear what is analogous to this in the case of the BSA. (Woodward, 2014, 115)

(2) The trade-off between accuracy and simplicity in the regression context and that between strength and simplicity in the context of the standard BSA seems to serve two fundamentally different purposes: in the former, it is supposed to select one among a number of theories which maximises predictivity, while in the latter, it serves to single out one among a number of different theories solely for the sake of a good trade-off itself (Woodward, 2014, 116). To put it differently, in the first context, the result of the trade-off merely has instrumental value, since it is supposed to indicate which theory has the highest predictive power, in the second, the result of the trade-off itself constitutes an intrinsic value (it indicates the best theory, where 'best' is non-instrumental, i.e. just 'best' full stop, not 'best for . . . ').

There are different (and somewhat obvious) ways to address point (1) which all depart from the observation that the idea that scientific theories flawlessly describe the universe as a whole is, from a naturalistic perspective, an unrealistic idealization. We assume here that the BSA has to anyway address this mismatch between its conception of laws and actual scientific laws, which are, as far as contemporary science goes, limited in scope. (Recall that the challenge to the BSA which we are concerned with here questions its naturalistic credentials!) We will briefly outline two ways here.

First, one might hang on to the (still to a certain degree idealizing) idea that the scientific theories whose strength and simplicity is at issue in the trade-off posited by the BSA cover the whole Humean mosaic, but reject the idea that these theories are based on perfect data. There are good reasons for doing so: if one thinks of the data as empirical (thus paying tribute to the empiricist roots of BSA), then one of course has to leave room for errors in the data due to errors in observation or in other means to gather the data. Overfitting to one particular data set, even if complete, could then be a problem. Even just the possibility for errors would leave room for multiple possible completions of the data set throughout all of which one might reasonably expect a law to hold.

The second way embraces the need to take into account data sets which fall short of covering the whole Humean mosaic: independently of how laws are discovered, a plausible structural precondition for being a law about a certain system is that it has to be applicable to different subsystems of that target system. Transferred to the BSA, this would mean that laws should not only hold for the whole Humean mosaic, but also for smaller subdivisions of it. That this perspective on the laws indeed makes a difference in the context of a Humean approach has been argued for by Hicks (2018), who proposes a pragmatic version of the BSA called the Epistemic Role Account (ERA). This account is, among other things, sensitive to the fact that initial conditions may hold in some subsystems but not in others. This fact in conjunction with the fact that initial conditions locally have the status of laws within the subsystem in which they obtain, can be taken to show that there is a need to avoid (a sort of) 'overfitting' to the subsystem, even in the context of a variant of the BSA. After all, the initial conditions of a local subsystem are not those of the whole target system.

According to Hicks, overfitting has to be avoided to get at a universal notion of lawhood. This notion only admits laws which not only hold in one or more subsystems

of, but also in the overarching global system, which, following orthodox Humean metaphysics, corresponds to the full Humean mosaic. If this sort of universality is a central desideratum to be met by a theory of laws, then a Humean theory has to also treat descriptive simplicity and strength (in the sense of accuracy) as central proxies for overfitting avoidance. After all, keeping descriptive strength constant, a more complex model for a subsystem is less likely than a simpler one to generalise well to other subsystems. This is because a complex model for a subsystem is, in general, more adapted to the specifics of that subsystem. Hence, a Humean theory which gives up on the classical BSA's requirement that lawhood only ever means lawhood given complete data (i.e. the complete Humean mosaic) and instead adopts the more realistic, i.e. epistemically more apt view that lawhood must also be definable relative to proper subsystems of the data, also needs to avoid overfitting, so point (1) is addressed.

What about (2)? On the one hand, it can hardly be denied that the trade-off in symbolic regression and the one postulated in the context of the orthodox BSA are not the same; they involve different notions of strength (deductive strength versus strength of fit/accuracy) and serve different purposes. On the other hand, the application of symbolic regression establishes—or at least has the long-term potential to establish, acknowledging that the field is still somewhat in its infancy—a de facto trade-off conception of laws in physics in the context of real data, providing evidence for the aptness of the general idea that lawhood can be partially explained in terms of a trade-off between scientific virtues. This is something then, one would think, pragmatist philosophers cannot ignore.

We propose thus to adopt the view that a thoroughly pragmatist version of the BSA should indeed incorporate a symbolic regression-style trade-off. Dorst's recent pragmatist take on the Humean conception of laws gives us a good starting point for this view. According to Dorst (2018, 2019), Humeans qua reductionists should adopt a genuine pragmatic perspective on laws, since it is otherwise not clear why they should attribute relevance to law structures at all. In particular, then, laws (as scientists know and set them up) are not to be understood as 'Gods-eye efficient summaries' of what goes on in the Humean base (as Lewis putatively had it) but as predictively useful statements for 'navigating the world'; consequently, the search for the best system should be a search for the best 'predictive' system—as qualified by the optimal fulfilment of predictively relevant virtues.

Going beyond Dorst's view, we assume that choosing a systematisation for its predictive *strength* is a way of choosing it for its predictive *usefulness*. To be sure, there are other ways in which a theory can be predictively useful. Taking into account the perspective of agents who make use of laws to navigate the world, we, however, take it to be highly plausible that having predictive strength is the decisive way of being predictively useful for a law or theory. Of course, more could be said about this, but for our current purposes, we will simply assume it as a working hypothesis, to be confirmed or disconfirmed by further investigations, that pragmatist Humeans should embrace an accuracy/simplicity trade-off.

The general take-away of our discussion is that pragmatically-minded Humeans have good reasons to accept a close analogy between the simplicity-strength trade-off

in symbolic regression and the one they postulate in the context of their theory of laws. Admittedly, orthodox Humeans who closely stick to the letter of Lewis' BSA are free to simply deny the analogy, but doing so will certainly not help them bolster their naturalistic credentials. Furthermore, they would still have to address our more general point that (most) Humeans appear to assume a rather naive conception of trade-offs, which we will get to now.

# 5 The Humean trade-off revisited

Symbolic regression seems to corroborate Roberts's and Woodward's view that the choice of systematisations as laws involves a strength-threshold. At the same time, it provides a clear conception of laws as systematisations that lie on a trade-off line between (forms of) strength and descriptive simplicity, where this trade-off involves a Pareto front). In this section, we argue, first, that—pace Roberts and Woodward—threshold behaviour which favours strength over simplicity is *de facto* compatible with the idea of a trade-off and, second, that it may even act as an indicator which singles out some equally optimal systematizations as laws—independently of the particular notions of simplicity and strength (or more generally, independently of the particular theoretical virtues) involved. This means that even Humeans who do not accept a thoroughly pragmatic version of the BSA of the sort suggested in the previous section can potentially rely on the following rejoinder to Robert's and Woodward's arguments.

## 5.1 The compatibility of thresholds and trade-offs

Given how central the notion of a trade-off is to the BSA, it is surprising that it has not been discussed more systematically in the literature. In this subsection, we discuss a plausible proposal for operationalising the notion. This discussion notably reveals that the BSA's insistence on a single best system is problematic.

What is a trade-off between simplicity and strength in a space of theoretical systems? A straightforward approach to answering this question is to define a total utility function that puts strength and simplicity into a systematic relation (e.g. which increases in some way with strength and in some way with simplicity)—and to find those systems which maximise it. For example, in his pragmatic construal of the BSA (already mentioned in section 4), Dorst (2018, 2019) operationalises Lewis's idea of a trade-off between strength and simplicity via a *predictive* total utility function that takes into account various theoretical virtues.

The crucial problem with an approach of this sort, however, is that any choice of a particular total utility function encodes a priori unjustified biases about the relationship between simplicity and strength. To minimise the influence of such biases, a better approach might then be to consider the results of optimising a whole family of total utility functions, thereby taking into account a range of different ways of valuing strength relative to simplicity. For instance, one might consider the whole 1-parameter family of total utility functions $U(J, s) = JP(s) + (1 - J)S(s)$, where $P(s)$ is the simplicity of a system $s$, $S(s)$ its strength, and $J \in [0, 1]$ the characterising parameter for the family, rather than a specific instance of that family such as $U(0.5, s)$. In

particular, maximisation of $U(J, s)$ with respect to $s$ will lead to a whole bunch of systems rather than just a single or a few maximising systems.

Nevertheless, the choice of a family of total utility functions might already introduce some bias into the relationship between strength and simplicity. To stay completely neutral, one should acknowledge that there is no natural total order on the two-dimensional space spanned by strength and simplicity. Accordingly, one should then discard commitments to any specific family of total utility functions other than for the instrumental purpose of formalising the preference structure between strength on the one hand, and simplicity on the other hand. Doing so will lead us to a conception of trade-offs in terms of a Pareto front (i.e. the conception put into practice in symbolic regression, see section 3).

In this spirit then, let $S$ be the set of systems (in the sense of the BSA). Let $v^i, i \in \{1, ..., m\}$ be functions that assign to each system $s \in S$ a virtue value with $v^1$ assigning a simplicity value, and $v^2$ assigning a strength value ($v^i, i \in \{3, ..., m\}$ stand for other virtue values). Hereby, we only require the co-domains of the functions $v^i$ (call them $C_i$) to be sufficiently ordered (one might want to just think of $C_i$ as $\mathbb{N}$ in the following and of each virtue as measured by natural numbers). Now, define $\vec{J}_m : S \to C_1 \times ...C_m, s \mapsto (v^1(s), ..., v^m(s))$. For two systems $s_1, s_2 \in S$, we can say that

- $s_1$ weakly dominates $s_2$ iff $J^i_m(s_1) \geq J^i_m(s_2)\ \forall i \in \{1, ..., m\}$ and $J^i_m(s_1) > J^i_m(s_2)$ for at least one $i \in \{1, ..., m\}$.
- $s_1$ strongly dominates $s_2$ iff $J^i_m(s_1) > J^i_m(s_2)\ \forall i \in \{1, ..., m\}$.

The Pareto front $P^m$ (of systems) is then to be defined as that set of systems which are not weakly dominated by any other, i.e.

$$P^m := \{s \in S | \{s' \in S | s' \text{ weakly dominates } s\} = \emptyset\}.$$

In particular, requiring an optimal trade-off between strength and simplicity for systems is first of all just tantamount to having a system $s \in P^2$. However, the formalisation easily allows one to take into account other virtues as well—the need for this is ultimately an empirical question.[2]

The result of the strength-simplicity trade-off initially corresponds to this set of points—and thus generally not at all to a single point. Only by introducing additional structure (expressing specific evaluations of how strength should be viewed relative to simplicity) could we ever arrive at a single optimal trade-off. Again, to stress, talk of a *single* trade-off is, without further justified commitments, inappropriate.[3]

This conception of a trade-off (set) which admits a plurality of equally optimal points is a standard conception that is widely applied in economics and the sciences. This makes the stark contrast between it and the picture of a trade-off between theories envisaged by proponents of the BSA, which admits only a single optimal trade-off point, all the more surprising. According to Lewis: "*The* best system is *the*

[2] Wilhelm (2022) has for instance argued for an extension of the BSA which integrates a virtue of computational tractability into its central trade-off.

[3] See for example De Weck (2004).

*one* that strikes as good a balance as truth will allow between simplicity and strength. [...] A regularity is a law iff it is a theorem of *the* best system." ((Lewis, 1994, 478), emphasis added.) Furthermore, there seems to be a consensus about this uniqueness assumption in the contemporary discussion about the BSA.[4] (Interestingly, Lewis did not commit to a single best trade-off in his earliest, tentative discussion of the regularity account of laws in Lewis (1973, 73ff.).)

It is not entirely clear why this consensus exists in the literature, especially given that there is no more systematic discussion of the nature of this trade-off. Perhaps it is due to the fact that the trade-off squarely belongs to what we in section 1 called the pragmatic side of the BSA: From a naive point of view, one might think that since the trade-off os 'merely' pragmatic, resolutions of the trade-off are reached by arbitrary deliberation. But this way of thinking is mistaken; pragmatic does not mean arbitrary. In any case, there will not be a single solution to the sort of two-dimensional optimisation problem Humenas have in mind, but rather a whole solution set as given by the points on the Pareto front. This suggests that the common Humean conception of laws as resulting from *the single "best"/"proper"/"optimal"* balance in a trade-off relation between simplicity and strength is without further qualifications highly questionable.

Note that the possibility that distinct systems could be tied for best in certain special cases has been discussed by a number of authors. (See Loew and Jaag (2020) for discussion and further references.) Our point here is that such situations are not just outliers, but systematically arise from the very nature of the trade-off to begin with.

## 5.2 Thresholds as tie-breakers

If the above argument is correct, then proponents of the BSA have to adapt their theory to cases in which there is not a single 'best' system which dictates what the laws are. Instead, they have to take into account the possibility that there can be a potentially large number of equally good, i.e. Pareto-optimal, systems. This might seem problematic. Whatever naturalistic reasons there are to admit more than one equally optimal system, it seems that we should not be forced to admit a potentially large multitude of optimal laws, where a smaller set could also do. We will now suggest a way to avoid cases of massive overdetermination of this sort and bring the suggested view of trade-offs more in line with the classical Humean's preference for a single best system.

Now, it is true that all systems that are represented by points on a Pareto front correspond to optimal trade-offs. Does this mean that every point on the Pareto front represents a different 'best' system? Or should the notion of a 'best' system rather be reserved for an elite subclass of all the systems which minimize the cost function? It strikes us as evident that proponents of the BSA, if presented with a choice, will pick the latter option. To align this preference with the operationalized notion of a trade-off introduced above, one can charitably construe the BSA's notion of a distinguished trade-off point—Lewis's "properly balanced combination of simplicity and strength"

[4] See e.g. Hicks (2018, 987) or Loewer (2007, 319).

(Lewis, 1973, 73)—in terms of a distinguished point on the Pareto front.

This raises the question of whether there is a criterion for singling out a small number of points, or perhaps in some cases also just a single distinguished trade-off point, on that Pareto front. We want to propose such a criterion which is based on an idea from economics, namely that of resolving opportunity costs.

Our discussion of symbolic regression illustrates that a Pareto front of equally Pareto-optimal points may contain certain points which are singled out geometrically: in figure 1, points located right next to a steep increase in accuracy coupled with slightly increased simplicity were associated with physical laws. This may be taken to suggest the following general threshold criterion for singling out systems on a Pareto front: best systems in a Humean sense are those Pareto-optimal system which are as simple as they can be without incurring 'too great' of a loss of strength. In economic terms, these are the systems that optimise simplicity while keeping the opportunity cost in losing out on strength sufficiently low. Note that 'too great'/'sufficiently low' are to be specified relative to a context; any assessment depends clearly on the range of simplicity and accuracy values considered and further contextual factors. (Generally, the opportunity cost can be formalised in terms of the slope of the strength-simplicity-trade-off line, allowing one to specify threshold values or ranges which fix which points on the line are designated in a given context.)

Now, one might think that the focus on distinguished points on the Pareto front supports Woodward's claim that simplicity only comes into play once a sufficient level of strength is achieved. After all, these points indicate that a local strength threshold is met. It needs to be kept in mind, however, that the other points on the Pareto front are still equally salient as genuine options in the trade-off. So there is indeed a genuine trade-off, but at the same time, extra structure (such as, in our case, a criterion which singles out some Pareto-optimal points based on strength opportunity costs) provides a way to validate an insistence on one, or perhaps a small elite class of preferred outcomes of the trade-off in the spirit of the BSA.

We should stress that the resulting picture of how trade-offs between scientific virtues allow multiple Pareto-optimal points and of how a number of these points may still be singled out as distinguished does not depend on what these values are. The strategy we suggest is indeed perfectly neutral regarding the theoretical virtues involved in the trade-off; it can generally be applied to two-dimensional trade-offs between arbitrary quantifiable virtues. In the current context, it is of course worth stressing that with respect to different variants of the Humean theory of laws, this in particular means that it can be applied to trade-offs involving different notions of simplicity and strength, including those embraced by orthodox Humeans who want to stick as closely to the classical version of the theory as they can.

## 5.3 Roberts's and Woodward's criticism of a trade-off reconsidered

Finally, let us reconsider Roberts's and Woodwards's criticisms of the relevance of a strength/simplicity trade-off in the formulation of physical theories in light of our operationalized conception of trade-offs.

The upshot of Woodward's criticism is that the BSA cannot appeal to science to justify its reliance on a trade-off between strength and simplicity, because scientists (including Newton and Einstein) seem to accept a strength-threshold which rules out the trade-off.

In general, our response is that, in the context of theory choice, a threshold with respect to a particular virtue—in the case of the pragmatic version of the BSA we prefer this is an accuracy-threshold—does not preclude a trade-off with another virtue. Rather, such a threshold can even serve as a useful selection criterion: According to our proposed criterion for choosing the best system(s) among all Pareto-optimal systems, one is to prefer a system with a higher simplicity as long as its explanatory strength is not significantly lower than that of a less simple one, i.e., as long as the opportunity cost of losing out on strength by caring about simplicity is not too high. In particular then, one will not commit to a (non-trivial) theory at all unless a significant gain of explanatory strength is thereby established.

Secondly, recall that Roberts argues against the relevance of trade-off considerations between strength and simplicity by giving us an example in which a less simple theory is supposedly clearly more acceptable than a simpler one solely on grounds of being more explanatory—no matter how much simpler the less explanatory theory is. The view of trade-offs which we propose in the previous section gives us another possible reading of Roberts's toy example: one could take it to illustrate a case in which one faces a choice between two Pareto-optimal theories and then relies on a particular strength threshold in order to resolve the tie between the two theories.

Roberts' claim that the stronger of the two theories is always chosen then amounts to the claim that scientists always assume a particularly high strength-threshold when choosing between Pareto-optimal theories. This argument can be contested based on the observation that such strength thresholds can vary between contexts. For instance, if one accepts the premise that Humeanism is an account of lawhood for both fundamental and non-fundamental contexts (as in the pragmatic version of the BSA we propose), what counts as a law and what does not simply has to be context-dependent. Simplicity can then effectively play no role vis-à-vis explanatory strength in some contexts (say, because one is concerned with the explanation of a more fundamental regime, so that the explanatory strength regarding *that* data is immensely higher in case of the less simple account) while it can very well play such a role in others (say, because in them, one is concerned with the explanation of a less fundamental regime, so that the explanatory strength regarding *that* data is not significantly higher in case of the less simple account). If this much is granted, then our response to Roberts's argument is that it is simply false that scientists assume the same strength-threshold in any context. His toy model then fails to make the intended point.

# 6 Conclusion

The BSA is often thought to be more naturalistic than rival theories of laws of nature. In particular, its reliance on a trade-off between strength and simplicity could be claimed to be in line with how laws are selected in science. We argued that this advantage is not only questionable in light of meta-scientific reflections by physical

authorities like Newton and Einstein (Woodward, 2014), or considerations about toy models (Roberts, 2008), but also in light of how trade-off relations are treated in simple scenarios of automated scientific discovery in which trade-offs actually play the law-selecting role Humeans assign to them (Udrescu and Tegmark, 2020; Udrescu et al., 2020; Cranmer et al., 2020a; Cranmer, 2023). This example also uncovers a further problem for the BSA, namely that the insistence on a single 'best' system appears to be based on a naive, overly simplistic view of how trade-offs are treated in scientific practice.

However, we also suggested a way forward for the BSA. The BSA can be adjusted to the more nuanced view of trade-offs illustrated by our discussion of symbolic regression, allowing it to accommodate cases in which there is a Pareto front of equally optimal outcomes. Furthermore, the adjusted account of laws can still partly conserve the spirit of the BSA view of trade-offs by integrating the assumption that the 'best' systems are those indicated by distinguished trade-off points on this Pareto front. This gives the account a way to make precise its standard of goodness: the best theories are those which minimize costs by offering a massive gain in strength coupled with a small decrease in simplicity (or, more formally, with a certain threshold in opportunity cost).

Our argument is arguably not only of interest to proponents of the BSA and other variants of Humeanism which rely on a trade-off to account for lawhood. Humeans who successfully address the adequacy challenge in the way we suggest are in a position to formulate a new challenge to rival theories of the laws of nature. If there is indeed a reliable method for identifying laws which relies on a trade-off between theoretical virtues, mirroring the BSA's conception of laws, then this poses a threat to rival theories. Given the existence of such a method, proponents of these theories have to explain how they can, despite relying on a criterion for lawhood which has nothing at all to do with such a trade-off, still guarantee the extensional adequacy of their theories: it would for example be a rather curious coincidence if there were exactly those dispositional essences needed to give us the exact same set of laws which can be effectively identified based on a simplicity/strength trade-off. This challenge may not amount to a refutation of alternative theories, but it certainly puts pressure on proponents of dispositional essentialism (see e.g. Bird (2007)), the nomic necessitation-view (see e.g. Armstrong (2016)) and Lange's counterfactual-stability-based theory (see Lange (2009)), who rely on metaphysical notions like essence, necessity, and primitive counterfactual truth to explain lawhood in manners which leave no, or little, room for a pragmatic trade-off between theoretical virtues to play a role in making a law a law.

## Acknowledgements

We would like to thank Feraz Azhar for many discussions on the concept of Pareto fronts and trade-offs, Rasmus Jaksland for general discussions on the nature of the project, and Mariana Seabra for discussions on automated scientific discovery. We thank Shelly Shi and Chris Wüthrich for helpful written comments. We furthermore thank the audiences of the HPS Seminar Copenhagen and the seminar se-

ries in analytic philosophy at the Centre for Philosophy of the University of Lisbon for their helpful comments. Thanks also to Silviu-Marian Udrescu for permission to reproduce a graph from a publication. N.L. thanks the Swiss National Science Foundation for support (grant number: 105212_207951). R.M.'s work on this paper was supported by the FCT (CEEC IND4ed, DOI: https://doi.org/10.54499/2021.03171.CEECIND/CP1702/CT0015). N.L. and R.M. furthermore acknowledge the FCT's support of the project 'Indeterminacy in Science' (DOI: https://doi.org/10.54499/2023.15136.PEX).

## References

Armstrong, D. M. (2016). *What is a Law of Nature?* Cambridge University Press.

Bird, A. (2007). *Nature's Metaphysics*. Oxford University Press.

Cranmer, M. (2023). Interpretable machine learning for science with PySR and SymbolicRegression. *arXiv preprint arXiv:2305.01582*.

Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S. (2020a). Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*.

Cranmer, M., Sanchez Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S. (2020b). Discovering symbolic models from deep learning with inductive biases. *Advances in Neural Information Processing Systems*, 33:17429–17442.

De Weck, O. L. (2004). Multiobjective optimization: History and promise. In *Invited Keynote Paper, GL2-2, The Third China-Japan-Korea Joint Symposium on Optimization of Structural and Mechanical Systems, Kanazawa, Japan*, volume 2, page 34.

Dorst, C. (2018). *The Best Predictive System Account of Laws of Nature*. PhD thesis, The University of North Carolina at Chapel Hill.

Dorst, C. (2019). Towards a best predictive system account of laws of nature. *The British Journal for the Philosophy of Science*.

Einstein, A. (1934). On the method of theoretical physics. *Philosophy of science*, 1(2):163–169.

Gamarnik, D., Moore, C., and Zdeborová, L. (2022). Disordered systems insights on computational hardness. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114015.

Greydanus, S., Dzamba, M., and Yosinski, J. (2019). Hamiltonian neural networks. *Advances in neural information processing systems*, 32.

Hicks, M. T. (2018). Dynamic Humeanism. *The British Journal for the Philosophy of Science*.

Lange, M. (2009). *Laws and Lawmakers. Science, Metaphysics, and the Laws of Nature*. Oxford University Press.

Lemos, P., Jeffrey, N., Cranmer, M., Ho, S., and Battaglia, P. (2022). Rediscovering orbital mechanics with machine learning. *arXiv preprint arXiv:2202.02306*.

Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103(412):473–490.

Lewis, D. K. (1973). *Counterfactuals*. Blackwell.

Loew, C. and Jaag, S. (2020). What Humeans should say about tied best systems. *Analysis*, 80(2):273–282.

Loewer, B. (2007). Laws and natural properties. *Philosophical Topics*, 35(1-2):313–328.

Roberts, J. T. (2008). *The Law Governed Universe*. Oxford University Press, New York.

Udrescu, S.-M., Tan, A., Feng, J., Neto, O., Wu, T., and Tegmark, M. (2020). AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *Advances in Neural Information Processing Systems*, 33:4860–4871.

Udrescu, S.-M. and Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631.

Vyas, R., Bapat, S., Goel, P., Karthikeyan, M., Tambe, S. S., and Kulkarni, B. D. (2018). Application of genetic programming (GP) formalism for building disease predictive models from protein-protein interactions (PPI) data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(01):27–37.

Wilhelm, I. (2022). Tractability and laws. *Synthese*, 200(4):318.

Woodward, J. (2014). Simplicity in the best systems account of laws of nature. *British Journal for the Philosophy of Science*, 65(1):91–123.

Wu, T. and Tegmark, M. (2019). Toward an artificial intelligence physicist for unsupervised learning. *Physical Review E*, 100(3):033311.