23

Symmetry as a guide to superfluous theoretical structure

JENANN ISMAEL AND BAS C. VAN FRAASSEN

Symmetries can be a potent guide for identifying superfluous theoretical structure. This topic provides a revealing illustration of the power of formal methods for illuminating the *contents* of our theories, and bears potentially on some very old philosophical problems. The philosophical and scientific literature contains a good many discussions of individual cases, but the treatment is rarely general and tends to be technically involved in a way that may bury the basic physical insight as well as making it inaccessible to philosophers. We wish to identify the sorts of symmetry that signal the presence of excess structure, and do so in a completely general way, applicable to all theories and all genres of theory.

1 What is superfluous structure?

For any entity whether concrete or abstract we distinguish its elements and its structure; the latter is specified by listing relations between the elements (equivalently, features of sets or sequences of elements). Whether or not some of its structure is superfluous is clearly an interest-relative question. A sowing machine has superfluous structure if some features of or relations between its elements are dispensable for sowing, although these may be quite relevant to it from an aesthetic or antique collectors' point of view. Each of two features may be dispensable for the given purpose, but they may not be both dispensable at once, namely if the machine has multiple features which can play each other's roles. Obviously then any machine at all – classified in terms of intended function and design – has superfluous structure. In the case of an abstract entity or intellectual product, classified in the same way, it may not be absurd to think of discarding all superfluous structure.

A physical theory provides us with descriptions and models that can be used to represent physical situations. We say that a theory has superfluous structure if it provides multiple representations for the same physical situation. Unfortunately, the way we describe and identify or distinguish physical situations tends itself to

https://doi.org/10.1017/CBO9780511535369.024 Published online by Cambridge University Press

be quite theory-laden. So it is not usually possible to have also at hand an authoritative, adequate, theory-independent account of nature against which we can test for superfluous structure. What we can do is inspect the theoretical representations and look for internal evidence to suggest that several of them are so alike that to distinguish between them makes a distinction without a difference.

'So alike': thereby hangs a tale. If they are alike to that extent, they must differ only in ways that amount to physically superfluous structure. Logically, of course, there could be physical differences that correspond to no measurable or observable difference. Logic alone will not decide what is physically superfluous in a model. Metaphysical views, physical intuition, and empiricist preferences for empirical content over metaphysical distinctions will all play a role in what is identified as really superfluous. But symmetry is a guide to this identification in all cases, and we wish to display this in a sufficiently general framework for the study of theory structure.

Outside the theoretician's study our clues to physically significant differences are the observable or measurable differences. But these provide no easy guide, and as we will see, the distinction between 'observable' and 'measurable' itself turns out to matter. A theory is usually around long before we know whether it contains unmeasurable quantities, and which quantities those might be. The process of identifying unmeasurable quantities, i.e. of smoking them out of their hiding places in the theoretical apparatus, is long, hard, and highly non-trivial. It is a discovery, then, and not a happy one, that a theory contains unmeasurable quantities.

Why is there something amiss with a theory that contains unmeasurable quantities? It is not because we have any *a priori* guarantee that there are no such quantities, i.e. that our senses see right through to the bottom of things, but because they are not a proper part of the subject matter of theorizing. To isolate an unmeasurable quantity mathematically is to demonstrate that the theory to which it belongs has idle parts, wheels which turn without turning anything. But it requires a clear view of theories as well as of theory–phenomena relations to clarify this.

2 What is a theory?

A theory has two main ingredients: a *theoretical ontology* which specifies its initial (*metaphysical*) *possibility space*, and a set of *laws* which selects therefrom the *physical possibilities*. There is for any historically developed theory a great deal of leeway in how it is conceived and presented. In the case of classical mechanics, for example, we need only recall the names of Newton himself, Lagrange, Hamilton, Kirchhoff, Mach, Birkhoff, and Mackey to see the diversity possible in foundational reconstruction. Our own form of presentation is meant to help bring out, as

https://doi.org/10.1017/CBO9780511535369.024 Published online by Cambridge University Press

perspicuously as possible, the roles of symmetry in conceptions of physical theory structure.

2.1 Theoretical ontology

The initial framework of a theory – its (*theoretical*) *ontology* – serves to delineate in very broad terms what will count as its initial (*metaphysical*) *possibility space*. Following common usage we will refer to the points in this space as (*possible*) *worlds*.

We will think of the ontology as specified by means of a *catalogue*, listing classes of entities, quantities, and relations which together determine the theory's parameters of representation. These serve as the 'supervenience base' for the description of nature: every possibility conceivable in this theoretical context can be conceived of entirely in terms of parameters that are either among or derivative from items in the catalogue. The items in the catalogue must be 'typed' (in the way that sets in type theory are, for example). Thus if A is one of the entities and Q a quantity, the specification that $Q(A) = r$ will make sense only if Q, A, and r are of the right type. All complete specifications of this sort identify points in the initial possibility space: the (*metaphysically*) *possible worlds*. This space is, in effect, the set of worlds obtained from the catalogue by means of any arrangement whatsoever of the theory's basic building blocks.[1]

As a first example think of the revival of ancient atomism in the sixteenth and seventeenth centuries. The world is conceived of as made up of atoms, whose number is the first basic quantity; each atom is characterized by means of the fixed list of primary qualities. Those primary qualities are actually quantities; their quantification was a crucial step for the new sciences. Initially at least, the 'mechanical philosophy' of the seventeenth century saw no need for more; any more clearly qualitative aspects of the world were merely derivative. What David Lewis and others have called a Principle of Recombination is clearly held in this context.[2] For if we take any part of the class of atoms of one world, combine it with some of those of another world, keeping in each case their primary qualities, then the result is a third (metaphysically) possible world.

If a theory is more holistic in its world picture, it is not equally straightforward to see it in this way. One way would be to think of the supervenience base as containing more complex quantities with the simpler ones as derivative. Another way, perhaps equally 'formal', would be to think of holistic properties of complex systems as

[1] In what follows, when we say 'possibility', we will always mean *metaphysical* possibility. Physical possibilities will always be explicitly identified as such.

[2] There are questions, here, about the metaphysical possibility of worlds that contain quantities or entities of kinds that are nowhere instantiated at our world (see Lewis, 1983), but these play no role in physical contexts.

external relations (i.e. relations that don't supervene on the intrinsic properties of their relata) between their parts.

Consider, for example, an *N*-body system in elementary quantum mechanics. The state of the whole is not determined by the states of its parts. In fact, as Schrödinger already pointed out, if the whole system is in a pure state, states ascribed to its parts can in general not be pure; and different pure states of the whole system are compatible with the same (mixed) states for its parts. Thus, one way to think of this is to take the state of the whole as basic and the rest as derivative. We can think here for example of taking all the states in the Schrödinger picture of *N*-body systems, for each *N*, as basic. With time then as an independent parameter, represented by the real number continuum, each possible world is a trajectory in one of the relevant state spaces. But it is also possible, in a formal sense, to think of the state of the whole as encoding the states of the parts plus certain non-supervening, non-spatial relations between its constituents that are not captured in their individual states.[3]

Even so, before any laws are introduced, and depending a bit on how much or how little we specify in the catalogue, many of these (metaphysically) possible worlds will be very strange and nothing like what we can think of as a quantum mechanically (physically) possible world. But what we will have at this point is a basic framework for which the laws can be formulated.

2.2 Laws and physical possibilities

Thus we can represent a theory as a structured set of possibilities, and this turns out to be a nice way of thinking of theories.[4] These structured sets of possibilities (heretofore, 'possibility spaces') relate in a straightforward way to spaces that physicists are accustomed to dealing with (phase spaces and configuration spaces, for instance, can be identified with subspaces of them), and they are familiar sorts of objects to philosophers.

We can think of the laws of theory as given in many different ways, but they have one simple role. They select a subset of points of the metaphysical possibility space: that selection is (or represents) the set of *physical possibilities*; we speak of these also as the *physically possible worlds*. If the laws are specified separately as conditions expressed by propositions or equations, for example, then the selected

[3] The case is complicated by the fact that the relations in question are relations between probabilities, i.e. correlations. See Mermin (1998, appendix A).

[4] Features that play an important role in some conceptions of theories (e.g. the language in which they are formulated or the mathematical form of their equations) are treated as incidental on this approach, relegated to the subsidiary role of picking out the worlds that are the real locus of interest. We don't set much store on debates about the 'right' way to represent theories; any way that contains enough information to reconstruct ontology and laws is adequate. It also doesn't much matter how we mark the physical possibilities, but it's convenient to keep the markings separate from the intrinsic structure of the space, so that structures intrinsic to the space represent only internal relations between possibilities.

worlds are precisely those points in the metaphysical possibility space that satisfy those conditions.

3 Empirical content

We must now introduce the ingredient that is external to theory structure, i.e. to the ontology and laws, namely the theory–phenomena relation. This enters in two contexts: when we are trying to decide what kind of epistemic attitude to take towards a theory, and when we are trying to interpret it.

3.1 Qualitative structure

To begin, we introduce the interpretative predicate ‘qualitative’. A world has qualitative features as well as structure, but this is a distinction we impose from without, by relating its features to us, the epistemic community. Thus the distinction is not to be found inside the theory or its models; it pertains to the theory’s (and its models’) use. Qualitative aspects of a physical situation correspond in our terminology to parameters which characterize that situation, and are directly accessible to us through perception. When we specify implications of a theory that pertain to qualitative aspects of a situation we are linking the theory to the content of possible perceptions, hence specifying empirical content.

We realize that the equation of qualitative with perceptible features is contentious, since something more than a choice of terms may be at issue. So we will expand on this by relating these terms to each other and to further terms typically used in this context.

Traditionally the terms ‘quality’ and ‘qualitative’ did indeed have a different meaning. Thus Kant’s Table of Categories lists *quality*, *quantity*, and *relation*, and distinguishes them in that traditional way. But for us today quality cannot contrast with quantity. Take any quality Q which an entity A may or may not have; we can think of Q as a quantity (a function that takes numbers or similar mathematical objects as values) such that $Q(A)$ equals 1 if A has Q and equals zero otherwise. Nor need quantities be real-number valued: a partially ordered set of determinations under a determinant can be represented by a function that maps entities into a partially ordered mathematical structure for its values. The distinction between metric and non-metric aspects of even a geometric structure is closer to the traditional quantity/quality distinction, but equally soft.[5]

[5] Coordinates can be introduced in synthetic projective geometry by construction, and a metric can be defined from these. This point was central to Russell’s (1897) diatribe against those who in his view betrayed the real theory of space.

3.2 Qualitative vs. measurable

But the distinction between directly accessible and non-accessible quantities, with the former's link to observation and experience, is crucial to any evaluation of how well or poorly a theory manages to represent nature. This distinction often appears disguised in terms such as 'physical intuition', 'physically meaningful', 'physically significant', and the like; indeed, its name is legion. Thus in a recent paper by Belot (2001), which we shall discuss further in a later section, it is easy to find telling examples. In his discussion of Hertz's programme, for example, he writes (p. 72):

> in fact, it is all too easy to implement it technically. The problem is rather that known implementations have little discernible *physical* interest. The kinetic energy in question bears no straightforward relation to any physical notion.

It would be hard to find any reading for 'physical' in this context that does not imply at least 'measurable' or even 'observable' .

Like so many in the philosophical lexicon, 'observable' is an accordion term all too easily squeezed or stretched out. One use applies only to *entities* – objects, events, processes which may or may not be small enough, large enough, massive enough, etc., to be observable by us.[6] A second use, which we will adopt here, applies to quantities that can characterize a situation, distinguishable by even a gross discrimination of colour, texture, smell, and so on. These alone we shall here count as *qualitative*. They are accordingly to be distinguished from *measurable* quantities. The latter include only quantities whose values make some discernible impact on gross discrimination of colour, texture, smell, and so on, but it doesn't matter how attenuated the connection is, how esoteric the impact, or how special the conditions under which it can be discerned. It will do no harm to count qualitative quantities as measurable, but many measurable quantities will be non-qualitative. We observe therefore a tripartite distinction between qualitative, measurable but non-qualitative, and unmeasurable quantities.

Which quantities are qualitative we take to be part of our resources for fixing the interpretation of a theory. But then we may take that either to be a fact of nature, or to be contextual, with the qualitative/non-qualitative line drawn differently in different theoretical contexts. While we have views on this, the choice does not seem to affect our main discussion. The measurable/unmeasurable distinction, on the other hand, is certainly not such a brute given line, but must be determined using those resources, for each theory to be interpreted. This point relates directly to our main topic: superfluous structure will align with the presence of unmeasurable quantities in the theory's world picture.

[6] This is the preferred exclusive use of one of us (BvF), but is relaxed in the present context.

https://doi.org/10.1017/CBO9780511535369.024 Published online by Cambridge University Press

3.3 Measurable but non-qualitative features

The history of this subject was ably told and given a remarkable continuation in Glymour's *Theory and Evidence* (1980). Imagine a discussion between Newtonians and their rivals the Cartesians around 1700. All will consider the quantities of extension as directly given to observation, for observable objects. But the measurement of some of these quantities, measurable by means of rulers and clocks, already presupposes certain substantive assumptions – in the case of planetary motions, for example, the immediate observations do not suffice without some assumption about how light travels. So far, however, with respect to the kinematic quantities, the two parties share the minimal needed presuppositions. It is different for the Newtonians' new quantities of mass and force. Certainly they offer operational procedures for determining these, but the measurement of the masses and forces in a system is a calculation on the assumption that this system is a system of Newton's mechanics.

This was the main point of Poincaré's discussion of quantities in classical mechanics in his *Science and Hypothesis* (1952), but it followed on the nineteenth century's attempts to define the dynamical quantities in kinematic terms. These attempts cannot be successful (given our current precise notion of definition) but it is possible to state theoretical assumptions presupposed in the measurement of those dynamical quantities using only kinematic terms. The best known example is still Mach's 'definition' of mass in terms of mutually induced accelerations under special circumstances. Since the special circumstances usually do not obtain, the challenge is to see whether the kinematic behaviour of the system implies the values of the masses and forces. The trivial case of a body never accelerated refutes this as a general claim, even if we amend 'implies' to 'implies relative to Newtonian mechanics'. Mach's procedure is inadequate for less trivial cases also.[7] But in the next century Pendse proved that if the complete kinematic data are given for the elements of a point particle system at a sufficient number of times then the forces and masses will be uniquely determined for a large class of cases. The main point is of course that this large class of cases is certainly not exhaustive – not only the general validity of the theory (such as constancy of mass) but also dynamical isolation (with conservation of linear and angular momentum) are assumed in the measurement calculations.

Subsequently Sneed (1971) gave a very precise form to Poincaré's conclusion; Glymour, while severely critical of Sneed, actually shows how this point emerges also in such other writers as Hans Reichenbach, and relocates it in his concepts of testing and evidence. A theory is tested by means of measurements of quantities on the assumption that the theory is satisfied; for certain quantities nothing more basic is possible. These are quantities which are measurable and definitely not qualitative,

[7] For a clear discussion of this and also the following, see Jammer (1961, chapter 8).

https://doi.org/10.1017/CBO9780511535369.024 Published online by Cambridge University Press

in our terms. From an epistemological point of view, therefore, measurement is in effect a procedure conditioned on ambient theories to generate new observable phenomena set in a theoretical context. Hence our emphasis here on observation rather than measurement for the theory–phenomena relation.

To sum up then: we are going to connect superfluous structure with the presence of unmeasurable quantities. However, what is measurable/unmeasurable cannot be read off directly from a theory. We need to make use of what is observable in order to make this distinction, but what is observable/unobservable does not align with what is measurable/unmeasurable. How the two distinctions are related to each other is spelled out to some extent in the above capsule history, but there is clearly more work to be done on this topic.

4 Symmetries: kinds of structure-preservation

We will adopt here the following precise terminology. A *transformation* is a one-to-one mapping of its domain onto its range. In the present context the domain and range are the initial possibility space onto itself. A transformation may be definable by means of instructions for adding or removing objects from a world, changing their properties, or shifting them around in space and time, but in general it may not even be computable or definable in any sense.

Intuitively, *symmetries* are transformations that preserve structure, and are therefore to be classified in terms of the (kind of) structure that they preserve. The symmetries *of a world* have to be distinguished from the symmetries *of a theory*. The symmetries of a world are transformations that map the world to itself. Thus if M is a possible world and h is a transformation pertaining to it, h is a symmetry of M if and only if $hM = M$. Symmetries of a theory, by contrast, are transformations of the initial possibility space that preserve satisfaction of its laws, and also preserve non-satisfaction. (These definitions are to be kept clearly in mind when we discuss various symmetry groups that can be associated with a theory.) In the terminology adopted above, the symmetries of a theory take physically possible worlds into physically possible ones and physically impossible worlds into physically impossible ones. Hence we can equally well call them the symmetries of its laws. (Indeed, one way to specify the laws is simply to specify that symmetry group – not to be confused with narrower symmetry groups that may be under discussion in a specific context!) They need not preserve any particular characteristic of these worlds.

It would in general make no sense to suggest that only features preserved by (invariant under) the symmetries of a given theory are real. For to identify all possible worlds related by such symmetries (as representations of the same physical possibility) would be to claim that there is only one physically possible world. Note

https://doi.org/10.1017/CBO9780511535369.024 Published online by Cambridge University Press

that *any* mapping of worlds to worlds that leads from 'legal' to 'illegal' ones are included; hence each physically possible world is the image of any other such world under some such mapping. In the case of Newtonian particle mechanics, for example, these mappings include ones that relate one system to another one with the same number of particles (but different trajectories), but also mappings that relate systems with different cardinalities.

We must look therefore to other symmetries, transformations that preserve not only satisfaction and non-satisfaction of the laws but other significant features as well. For example, in the construction of a theory one could begin with transformations that preserve some spatiotemporal structure, and perhaps some other quantities, and then ask under which among these the laws are preserved. Weyl (1952, pp. 26–7) made this general point, which can be illustrated with the example of Galilean transformations in classical mechanics. They are intimately related to the laws. Yet we cannot say that a feature invariant under all Galilean transformations is always a matter of law. For number – the number of particles in the system, say – is invariant, but the laws of classical mechanics do not impose any constraint on number. The laws constrain *relations between* quantities, but not, typically, the particular values they take. The nomologically contingent, but nonetheless real, features of a situation are precisely those that distinguish worlds related by a symmetry of the laws.

So in a diagnosis of what is real and what is superfluous structure the symmetries to focus on may belong to some intermediate class(es). At least one springs to mind: the class of *symmetries of the theory that also preserve all qualitative features of every model*. This is a class that can be defined in the same way for all theories, and therefore has the same generality as the notions of symmetry of a world and symmetry of a theory.[8]

5 The signs of superfluous structure

Consider a symmetry of theory T which maps each physically possible world to one that is qualitatively indistinguishable.[9] It will always be possible, without affecting

[8] The Galilean transformations do not preserve position or velocity, yet these are directly measurable quantities. That is an objection at first blush, but perhaps not on second thoughts. Position and velocity in a given frame of reference are not preserved, while distances and relative velocities are. But it is precisely the latter that are measured; the former result from a measurement plus some convention or paper-and-pencil operation. Recall from above that, as we draw the lines, the qualitative features are those directly observationally accessible to the observer, while measurement in general involves the deliverances of observation plus some theoretically conditioned calculation. Note that it may also be necessary to add an indexical element into this conception. A navigator on a ship may, for example, record the presence of an iceberg ten miles to the north-east. It is clear that such a report is to be understood as carrying the indexical 'from here' Such a report can be made in ignorance of the navigator's location on the globe, and its indexical content is independent of that. But we leave this aspect of observation to other occasions.

[9] This is essential; the transformations in question have to preserve the qualitative structure of all physical possibilities. See below for a further apparent disagreement with Belot.

https://doi.org/10.1017/CBO9780511535369.024 Published online by Cambridge University Press

the empirical content of theory, to interpret such a transformation as a symmetry of each of the worlds it relates. It will always be possible, that is to say, to interpret it as acting as the identity on the worlds in question, notwithstanding that it may not act as the identity on our *representations* of those worlds. If h is a qualitative-structure-preserving symmetry of the laws, and M_T and M_T^* are theoretical representations of worlds (descriptions or mathematical world-representing structures), such that $M_T^* = hM_T$ and $M_T^* \neq M_T$, it will always be possible to identify the worlds represented by M_T and M_T^* respectively, and to think of the structures that distinguish their representations as superfluous.

5.1 Trivial and non-trivial transformations

To interpret a transformation as *trivial* is to interpret it as permuting (at most) physically insignificant features of theoretical structure. As physically significant we count precisely (i) those features that are (jointly or individually) constrained by the laws, plus (ii) the qualitative features, regardless of whether they are constrained by the laws. *Hence we submit that it is precisely the qualitative-structure-preserving symmetries of the laws that are indicative of the presence of superfluous theoretical structure and should always be interpreted as trivial.*

We are distinguishing here three classes of transformations, with corresponding interpretational claims:

(a) Transformations that preserve qualitative structure but are not symmetries of the laws. These will turn at least some physically possible worlds into physically impossible ones, though there is no observable difference.
(b) Symmetries of the laws that are not qualitative structure preserving. These produce distinct physically possible worlds.
(c) Qualitative-structure-preserving symmetries. We submit that it is precisely these which suggest the presence of superfluous structure.

The symmetries in class (b) include non-geometric transformations that permute the values of dynamically relevant unobservable quantities (i.e. transformations that induce unobservable changes with potential but non-actual observable effects), and geometric transformations that are not symmetries of the laws (i.e. transformations that can change the state of motion of a system in a way relevant to their dynamical behaviour).[10]

5.2 A sophistical argument

Here is an argument that all theories will necessarily have superfluous structure by the above criteria. Consider even the very simple theory that says:

[10] We will discuss an example below, in connection with a claim made in Belot (2001).

https://doi.org/10.1017/CBO9780511535369.024 Published online by Cambridge University Press

(i) There are exactly two objects.
(ii) There is a red object and a black object.

Take a world that satisfies these laws; it has two members in its domain. Call the red one **a** and the black one **b**. There is another world, produced by the permutation of the two (a transformation that does not preserve colour, but certainly legality) in which **a** is the black one and **b** the red one. These two worlds are qualitatively indiscernible. But they are clearly distinct, for in the first one the property $(\lambda x)(x = \mathbf{a}\ \&\ \text{Red}(x))$ is instantiated and in the other world it is not instantiated. (Notice that the Principle of the Identity of Indiscernibles does not get any purchase here: in each world each individual has a uniquely defining description.)

Various gambits are familiar from the philosophical literature to try and block this conclusion. Could we insist that non-trivial transformations must preserve the individuals' essential properties, and that the *identity properties* (such as $(\lambda x)(x = \mathbf{a})$) are essential? Bad move, for then only the identity transformation $f(x) = x$ will count as non-trivial. Could we suggest that '=' be dropped from our admissible language for constructing theories? Sounds like a bad move too, if we construe (i) in the usual way as 'There are x, y such that not $[x = y]$ but for all z, $z = x$ or $z = y$'. We should be able to express counting. However, it is possible to go higher order and have primitive number properties: 'in these worlds Red and Black are singly instantiated' and 'the property of being Red or not Red is doubly instantiated'. Then '=' can perhaps be dropped.

We see this as a sophistical problem, artificially generated. There are two obvious choices for philosophical bookkeeping. We can say that the argument succeeds in displaying superfluous structure for every theory, but that this is *trivial superfluous structure*, defining that notion precisely as 'superfluous structure that every theory has'. Or alternatively we can say that the displayed worlds are identical, that we only have two descriptions differing in an arbitrary labelling that has no physical significance. Our own preference is for the latter, but we do not see it as a substantive issue.[11]

In objection it may be suggested that each individual has its own haecceity, which we thereby ignore, and that this makes the issue substantive. But introducing haecceity, or any other hidden individuating factor, will just push the sophistical argument one step further. For if all non-trivial transformations must preserve haecceity, then only the identity transformation is non-trivial. If on the other hand haecceity need not be preserved, then there will be trivial but distinct variants on worlds once more produced by permutation of the individuals. This dialectic is then merely a useless replay of the same argument. In what follows we will sometimes just say 'ignoring the individuals' identity' to remind ourselves of this bit of

[11] Cf. van Fraassen (1991, chapter 12, section 4.2) and Huggett (1999).

philosophical book-keeping. But let us add a few more remarks to locate this issue properly.

There are two ways one can think about transformations. One can describe transformations in an interpreted metalanguage, with as much structure as one likes. In that case, for the purposes of describing their action on the class of structures under consideration, they should be grouped into equivalence classes insofar as they are indistinguishable by their effects on structures in the relevant class. Alternatively, one can describe the transformations in the same language in which the structures are described. In that case they are individuated more coarsely, so that there are no distinct transformations that are indistinguishable by their effects on the structures in question. This has the effect of tying the individuation of transformations to the structures in the class in such a way that when one identifies superfluous structure in the models, it is identified simultaneously in the class of applicable transformations. The second alternative is more natural, because it ascends a level, applying the same reasoning to the individuation of transformations as to the individuation of physical possibilities. Transformations that act in the same way on all structures in the class are identical, notwithstanding notational variation in their presentation. In application to the example, it has the consequence that talk of permuting colours over individual elements while leaving global colour distribution the same is nonsense.

6 Symmetries of the laws

The claim we have now introduced we have not so far supported, but by examining other accounts of the matter and examples that illustrate the differences we mean to make a plausible case. Our claim is in at least apparent disagreement with Belot, who writes (2001, p. 55):

> if there were a class of possible worlds whose shared geometry and laws were invariant under some set of symmetries, then . . . there would exist distinct worlds sharing all of their qualitative and relational properties.[12]

Belot's article, framed as a discussion of Leibniz's Principle of Sufficient Reason (PSR), is devoted to an exploration of the possibility of a trivial interpretation of certain symmetries of classical mechanics. Instead of focusing on spaces of possible *worlds*, Belot speaks in terms of spaces of possible *states*, and he focuses on spaces whose geometric structure determines the physically possible trajectories through it, effectively writing a theory's dynamical laws into the space. So what we call worlds are the trajectories in his discussion. That is just a different way of speaking, but there are reasons (good reasons, we think, that discourage otherwise seductive

[12] We readily admit that we are using this excellent and insightful article as a stalking horse, and that the apparent disagreements may not go very deep; they will still clearly bring out the contents of our claims.

https://doi.org/10.1017/CBO9780511535369.024 Published online by Cambridge University Press

confusions) to prefer talking, as we have, in terms of spaces of worlds with no more intrinsic structure than is given by the internal relations among their elements, and we will continue to do so.

Returning then to the quoted thesis, we note that Belot in effect omits the restriction to symmetries that preserve qualitative structure, allowing for simple counterexamples. Consider, for instance, the very simple theory that says that there are two types of thing in the world, triangles and squares, and that there are two laws:

(i) There are only three objects in the world.
(ii) Either everything is a black triangle, or everything is an orange square.

Ignoring the identity of the individuals (see above!) there are here only four physically possible worlds, distinguished by having 0, 1, 2, or 3 black triangles. Consider now the transformation $\mathbf{h}^*$ that changes triangles into squares and squares into triangles, and changes black things into orange ones and orange into black. This transformation $\mathbf{h}^*$ is a symmetry of (i) and (ii), and maps the set of physically possible worlds onto itself. But these worlds are all distinct in structure.[13] Physically realistic examples are easy to find. Whenever you've got a theory with only qualitatively distinguishable models, any automorphism of the set of physical possibilities will constitute an example.

For a contrasting example in which qualities are preserved and in which there are distinct, indiscernible worlds, take any theory, add to it a hidden, unmeasurable quantity, Q, and let $\mathbf{j}$ be the transformation that leaves everything else intact but permutes the value of Q. Since Q is unmeasurable, it is causally isolated from the values of all other quantities, and so permuting Q-values at a world is not going to affect the values of observable or measurable quantities. As long as we don't add new laws explicitly mentioning Q and relating them to observable or measurable quantities, $\mathbf{j}$ will be a symmetry of the theory, and among the theory's models will be worlds that share a geometry, and all qualitative and relational properties.

A more interesting example is provided by Glymour's example of the revised Newtonian theory that replaces Newtonian force with the complex quantity morce + gorce (1980, pp. 356–62). Let $\mathbf{k}$ be the transformation that adds to morce what it takes from gorce, leaving force unaffected (suppose morce and gorce can take negative values). The revised theory will be invariant under $\mathbf{k}$, and world $\mathbf{w}$ will be (except in degenerate examples) distinct but indistinguishable from world $\mathbf{kw}$.

Why would Belot have ignored such examples? Perhaps if we focus on geometric symmetries the distinction between qualitative-structure-preserving and non-qualitative-structure-preserving symmetries of the laws will not spring to mind. What distinguishes the geometric symmetries as a class is that they preserve

[13] To fill out the example we could add that (i) and (ii) are invariant under Galilean transformations, and that $\mathbf{h}^*$ leaves the metrical structure of a world intact.

qualitative structure; but it is not in general true that symmetries of a theory preserve the qualitative structure of its models. It is only by contrast with geometric transformations that are *not* symmetries and symmetries that do *not* preserve qualitative structure, that what is special about geometric symmetries emerges. By considering other symmetries that figure prominently in the literature (e.g. gauge symmetries and permutation symmetry) we can get a handle on the general physical significance of symmetries. Transformations that preserve qualitative structure are not all symmetries of the laws, and transformations that are symmetries of the laws do not all preserve qualitative structure. Only those that have *both* features suggest the presence of superfluous theoretical structure. Only those that have *both* features permit a trivial interpretation.[14]

6.1 Qualitative indiscernibility with dynamic differences

Belot sometimes speaks as though Leibniz's Principle of Sufficient Reason (PSR) enjoins the identification of *any* pair of qualitatively indistinguishable models of a theory. He introduces the article, for instance, with the claim (2001, p. 55) that:

> a description of a set of possible worlds which includes pairs of worlds with identical qualitative structures can [and, according to PSR, *should*] always be taken to correspond to the sparser set of possibilities which arises when qualitatively identical worlds are identified.

But he can't mean that.[15] Consider the transformation that maps an empty rotating bucket in a Newtonian world onto a qualitatively indistinguishable bucket at rest, and maps every other world onto itself. Both are models of Newton's theory, and if neither bucket is filled, they are qualitatively indistinguishable. There are undoubtedly differences between the two according to Newton's theory, which would become manifest if measuring instruments (or water) were introduced into the buckets. But by hypothesis there are only the buckets, and they are therefore the same in all their qualitative features. PSR, however, ought not to be understood as instructing us to identify these worlds. It ought to be understood, rather, as counselling aversion to recognition of dynamical distinctions that have no *potential* qualitative effects. By 'potential qualitative effects' we mean here qualitative effects that show up in some *physically possible* circumstance, according to the theory (i.e. in some of the theory's physically possible worlds).[16]

[14] This provides some insight also into the conditions under which we recognize intrinsic geometric structure, given that geometric transformations all preserve qualitative structure; viz., when they have potential qualitative effects.

[15] And some of the other things that he says suggest that he doesn't, though there are still others that make it clear that he is not *simply* misspeaking.

[16] Belot discloses (in personal correspondence) that he means 'qualitative' in a different sense from ours: 'I use "qualitative" in the metaphysicians' sense [to refer simply to the intrinsic properties of an object]; it is, I think,

https://doi.org/10.1017/CBO9780511535369.024 Published online by Cambridge University Press

The difference is far from trivial. We get the space of metaphysical possibilities, in physical contexts, by unconstrained recombination from the entities, quantities, and relations that we take to be the building blocks of the actual world. If we took the state space of Newtonian mechanics, as suggested by Belot, and simply identified any pair of qualitatively identical worlds, we would give up combinatorial structure (the space would contain, for instance, worlds in which there are spinning water-filled buckets, but no worlds in which there are empty spinning buckets), and that isn't something we can surrender just because it happens to suit our theoretical purposes. One of the reasons why we *care* about the space of possibilities, one of the things that makes it an object of physical interest, is that it is related, in the way expressed by the Principle of Recombination, to the structure of every world *in* it. We care about (metaphysical) possibilities, in physical contexts, at least in part, because they relate in a principled way to the structure of actuality, and we can't abandon the relating principle without relinquishing (this aspect of) their significance.[17]

The physical intuition to which Leibniz's principle answers is that if we are recognizing a set of possible worlds with a great deal of qualitative redundancy, we may be recognizing more structure in the actual world than there is good reason to suppose, and if we can find a way of trimming away some of the fat without cutting into the meat (i.e. if we can find a way of doing without some of the non-qualitative structure without losing any qualitative distinctions), we should do so. But this business of identifying qualitatively indistinguishable possibilities won't tell us anything about the structure of the actual world if we give up Recombination. We don't understand what the actual world is *like* according to a theory that removes some of the qualitative redundancy unless we can place it against the background of a combinatorially structured space. We don't really understand what a theory says the building blocks of the actual world are unless we know how to take them apart and put them back together; that is, unless we know what the theory says the real degrees of metaphysical freedom are.

much weaker than your usage; and it is not directly founded on considerations involving perception . . . '. We have not emphasized the difference because it does not affect the counterexamples. We can simply stipulate in the orange square/black triangle example that the only real, intrinsic properties of the worlds described are the colour and shape, and there are no more any real, intrinsic properties to distinguish the rotating from the stationary empty Newtonian bucket worlds than there are qualitative ones. The weakened sense of 'qualitative' does make Belot's discussion less ambitious than one might have hoped; since which properties are qualitative in the weakened sense is something that is not determined independently of a theory's interpretation, Belot is not recommending a general, theory-neutral criterion for identifying superfluous structure. His discussion applies only after interpretive decisions have been made. See further, below.

[17] There may be a new space, with a combinatorial structure, containing all and only the worlds in the modified Newtonian one, but the worlds in that space would have a non-Newtonian structure. The claim is that we wouldn't understand the physical situations depicted by the modified Newtonian worlds – we wouldn't understand what their constituents were, what kinds of entities, quantities, and relations they were really made of – until we saw them in the context of the new space, until we saw how to take them apart and put them back together.

To review, then: symmetries of a theory, T, are transformations that map the set of physically possible worlds onto itself. Transformations that *aren't* symmetries of T, by contrast, sometimes take you from a world that is physically possible by T's lights, to one that is not, and can be understood as changing the world in some dynamically relevant way. Some of these dynamically relevant changes will be visible, but some will have visible effects only under certain conditions (e.g. the difference between empty Newtonian buckets in rotating and non-rotating universes). The difference between such worlds is qualitatively potential because if we filled the buckets with water then, *ceteris paribus* and keeping the laws fixed, qualitative differences would emerge. The PSR should be understood as a prohibition on invisible, dynamically irrelevant differences. It should be thought of as a ban on the recognition of invisible, intrinsic differences that do not have visible manifestations, under any physically possible conditions, actual or counterfactual.

Geometric transformations are special in that they preserve qualitative structure, and so recognizing differences between worlds related by geometric symmetries is always a violation of the PSR. But symmetries do not in general preserve the qualitative structure of their models, neither in our sense of 'qualitative', nor in Belot's weaker sense. If they did, and if we identified all physically possible worlds related by symmetries, no theory would have more than a single model. Belot has suggested, in personal correspondence, a restricted notion of symmetry:

> I count as symmetries of a theory only those permutations of its space of possible worlds which preserve the structure defining the dynamics of the theory (thus my symmetries are diffeomorphisms preserving, say, the Hamiltonian and the symplectic structure or Hilbert space structure, in typical physics cases).

There is nothing illegitimate about building restrictions into your definitions. But there is a cost to this. If you *define* the symmetries of a theory as those that preserve the structure defining the dynamics, you cannot then use considerations involving symmetry to see how much dynamical structure is needed to reproduce the empirical content of a theory, i.e. to reveal dynamical structure that is not really doing any empirical work. Symmetry is a mathematical notion, and we think it best to keep our definition of the symmetries of a theory uncontaminated by physics. We also consider it important that any notion of symmetry used in a special context should derive from general notions defined for any theory in the same way. Combine a purely mathematical notion of symmetry with a theory-neutral distinction (as indispensable precursor to interpretation) between qualitative and non-qualitative structure in the models, and you have a good guide to identifying superfluous structures, one whose epistemic motivation is plain, and whose application doesn't wait on the very interpretive decisions we want to use it to make.

6.2 Symmetries of a world, and identity of indiscernibles

The symmetries of a given world, as distinct from the symmetries of a theory, are just those transformations that map it into itself (its *automorphisms*). That is the sense in which a 360° rotation is a symmetry of any letter, while a 180° rotation is a symmetry of the letter O but not of the letter P. For these symmetries it would certainly make no sense to suggest that they detect superfluous structure in a particular world. Since they map the world into itself, they cannot be used to support any claim to the effect that this possible world is really the same as (represents the same physical situation as) some other world.

However, if a transformation is a symmetry of some worlds but not others, we can *raise the question* whether it does not perhaps preserve *all* significant structure, and thus relates only worlds that represent the same way a real world could be. If physical situations S and S' are represented as mirror-images of each other, for example, are there really two distinct physical possibilities being represented, or only one? The question is: does the mathematical operation correspond to a physical operation? In applying this transformation are you really reorganizing a world, or mapping it onto a duplicate, or just permuting insignificant bits of the representation?

There should be a strong suspicion of superfluous structure if two distinct worlds are related by a transformation that has some world as a fixed point. As an intuitive example, familiar from much literature, imagine that worlds w_1, w_2, and w_3 have in them respectively only a left hand, a right hand, and two hands (one right and one left, which are each other's mirror-image reflected through a central plane). Reflection will turn w_1 into w_2 and vice versa but turns w_3 into itself. This is the clear danger sign that makes us think that w_1 and w_2 do not represent two really distinct possible physical configurations but only one.[18]

This historical example has much about it that is questionable and has made it the topic of a large and diverse literature. Consider a more abstract example, the four-element group known as Klein's *Viertelgruppe*.[19] This is a commutative group with four elements e, a, b, c, in which e is the identity element ($ex = x$ for all x), each element is its own inverse, and if x, y, z are distinct elements other than e then $xy = z$:

	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

18 Cf. Pooley, this volume.
19 This example is used to advantage in Rynasciewicz (2001).

Here the element *e* is uniquely definable. But we cannot construct a uniquely identifying description for any of the other elements; they are structurally and qualitatively indiscernible. The automorphisms of this structure are precisely the permutations of its set of non-identity elements, and only what is invariant under the automorphisms can be definable. It would however make no sense to suggest that those elements should therefore be identified. The result would be a two-element group, and we would have the strange consequence that the *Viertelgruppe* does not exist (though there are objects that have this form, plus additional structure).

There are two points to be made here. The first is that the group is invariant under permutations of the three non-identity elements. Replace *a* by *b* and conversely, and the table is just

	e	*b*	*a*	*c*
e	*e*	*b*	*a*	*c*
b	*b*	*e*	*c*	*a*
a	*a*	*c*	*e*	*b*
c	*c*	*a*	*b*	*e*

and that is quite obviously the same table as before, written in slightly different order. This should raise the suspicion that if two worlds are related by such a permutation, they do not represent two distinct possibilities. But secondly, this permutation invariance is no basis for suggesting that this world, the *Viertelgruppe*, has been redundantly depicted in the above table.[20]

6.3 A new sophistical argument

We emphatically used the word ‘suggest’ above: each case has to be examined separately. Thus we distinguish this topic (symmetries of worlds) very emphatically from that of qualitative-structure-preserving symmetries of the laws and their clear implication of superfluous structure.

But there have certainly been suggestions, typically connected with Leibniz’s Principle of the Identity of Indiscernibles (PII), to draw an exceedingly general moral. Consider world w_3 above consisting of two hands which are each other’s mirror-image. That world is left unaffected by reflection through its central plane – conclude then that actually it contains only a single hand! Black’s world consisting of two identical spheres provides the simplest (if not the most illuminating) example. Why not conclude that this world actually contains only a single sphere, and is here redundantly described, as in our own familiar duo of the Evening Star and the Morning Star?[21]

[20] Cf. French and Rickles, this volume.
[21] One of us did at best (on the most charitable reading) come exceedingly close to being taken in by this line of argument; see van Fraassen (1985, pp. 63–5).

These arguments are sophistical, trading on an untenable version of PII. There are undoubtedly cases in which a theory has distinct models that are related by a qualitative-structure-preserving symmetry of its laws and cannot represent distinct possibilities. That this is indeed so in a specific case may be suggested by the fact that this symmetry has some worlds as fixed points. But here, in this sophism, the inference goes in the opposite direction: to the conclusion that it is those fixed points which have superfluous structure in themselves, and hence are superfluous items in the theory's physical possibility space!

That it is a sophism the very examples of spheres and hands should already illustrate.[22] If they do not, it is because in such fanciful cases the cost of denying obvious possibilities may not seem so high. But first of all, the other examples we gave should make the cost quite clear, and secondly, we can put the matter quite abstractly.[23]

Suppose that for a certain two-place relation R the following is true in a given world:

> There are objects x and y such that Rxy and not Rxx.

In that case the world contains at least two objects. This follows regardless of anything else that may be true in this world, and therefore regardless of whether the objects in question are differentiated in any describable way.

This simple point defeats many a naive version of PII. There are more sophisticated versions that do not fall so easily, but we will not hold them sacrosanct if they are not tautologous.[24]

The main point is also well illustrated by an example Belot cites in his discussion of whether indiscernibles in a theory's models can be identified. The logical way to do that would be to reduce each world *modulo* the indiscernibility relation (2001, p. 60):

> The upshot: whenever we have a structure that admits non-trivial symmetries, we can factor these out, constructing a quotient structure . . .

The example that halts this suggestion in its tracks (suggested to Belot by Kit Fine) is simple enough (*ibid.*):

> Consider two structures for a given countable set of objects; in one structure that set of objects is given an ordering isomorphic to the integers, in the other an order isomorphic to the rationals. The quotient of each structure is just: a single object, related to itself.

Belot notes simply that it is 'necessary to examine the relation between a structure and its quotient on a case-by-case basis' (*ibid.*). Rather disappointing, if one was

[22] See the discussion in van Fraassen (1991, pp. 454–6 and 459–65).
[23] *Ibid.* (p. 456, last paragraph of section 3.2).
[24] For a thorough discussion of the issue, with carefully nuanced distinctions, see Saunders, this volume.

hoping for a general method for wholesale elimination of putative superfluous structure![25] We can see now, however, what crucial distinction tends to be ignored in this area. Symmetries *of worlds* – of single structures meant to represent nature – can at best offer a suggestion of, or clue to, possible superfluous structure in the theory. They can certainly not imply that – the world could after all be symmetric in any way it likes! So a theory must be allowed to have models that have any conceivable kind of symmetry. It is not the symmetry of any given world, but the qualitative-structure-preserving symmetries *of laws*, that can definitively reveal superfluous structure.

7 The wider theoretical context

Formalisms with little superfluous structure are nice, of course, because they reflect cleanly the structure of what they represent; they have fewer extra mathematical hooks on which to hang the mental structures that we project onto the phenomena. But we want to conclude this general discussion of *theory structure* with a reflection on the *structure of theorizing*.

Methods for removing excess structure are much more than mopping up procedures. They are not something merely to be done *after* our representations have been crafted, like portraitists erasing stray pencil marks, or sculptors removing extra clay. Methods for removing excess structure are the very heart of theorizing; we figure out what the world is *like* by seeing what kinds of representations it supports. In theorizing one starts, that is to say, with the *representations*, and works one's way towards ideas about the intrinsic character of their common object by a kind of triangulation.

There are two stages in theory construction. The first is to generate a set of models rich enough to embed the phenomena, the second is to attempt to simplify those models by exposing and eliminating excess structure. Continuing in this way the structure of the models is pared down, being careful not to jeopardize their capacity to embed the phenomena. The whole class is thrown over only if a new significantly simpler set of models is found.[26] These inside-out procedures for identifying superfluous structure are indispensable, and the identification of qualitative-structure-preserving symmetries of the laws is paramount among them.

[25] Something Belot, in correspondence, disavows: 'You might wonder, after all these qualifications, what the project of my paper really is? A modest one: to point out that there will always be available in philosophy of physics a trick which allows you to pass from a formulation of a theory that admits symmetries to a related one which does not; and to make a very modest start on assessing the interpretative merits and demerits of making this move in some classical cases.' Even this modest project, however, cannot proceed without a distinction between symmetries that are, and symmetries that are *not*, candidates for reinterpretation. Transformations that do not preserve qualitative structure, transformations that map observationally distinguishable worlds onto one another, even if they are symmetries of the laws, are not candidates for reinterpretation.

[26] Or one whose models are demonstrably simpler than our good faith estimation of the potential for simplifying the ones we have.

If we interpret such transformations as trivial, we drain the structures that distinguish the representations they relate of significance, giving us simpler models at no empirical cost.[27]

This two-stage conception of theorizing has many historical illustrations. One familiar example is provided by the development of quantum mechanics.[28] At first one must be struck by the differences between Schrödinger's wave mechanics and Heisenberg's matrix mechanics. A first simplification came with von Neumann's insight into the shared Hilbert space structure, a second with Weyl's display of the still more basic group-theoretic structure behind the algebras of observables.[29]

A guide for identifying superfluous structure, however, is not a recipe for formulating a nice theory that does without it, that is to say, a local intrinsic description of the world that has all of the properties we like theories to have. There is no better illustration of this than the problems, amply chronicled in this volume, associated with the interpretation of gauge symmetry. Focusing on Yang–Mills theories, we have local symmetries of the generalized phases associated with the wave functions of the matter fields that show all the formal signs of revealing superfluous structure, but we can't simply excise the problematic structures without rendering the theory non-local. It seems that we need *something* in the region of space occupied by the gauge potentials to explain effects like the Aharanov–Bohm effect in a local manner. Redhead and Nounou (both this volume) explore the options.

8 Conclusion

We have been exploring the question of how symmetries can function, in the context of physical theory, as guides to the presence of superfluous structure. The philosophical lesson that can be taken away from the discussion is an insight into what has emerged as the most characteristic feature of modern physics. The ontologies of our most fundamental theories are not guided by physical intuition; they are not shaped by philosophical prejudices, but led, at their best, by the ideal of a kind of formal simplicity. The history of modern physics has been (to adapt a phrase from a recent book by Barbour)[30] 'a long, sustained effort to shed redundant concepts', and symmetries of the right sort, symmetries of the sort that we have been talking about, can act as beacons of redundancy.

[27] The flip-side is, of course, that the larger the set of geometric symmetries of a theory's laws, the less dynamically significant spatiotemporal structure it recognizes.
[28] We can be brief here; see for example Otavio Bueno's discussion of this development in the context of the heuristic value of symmetry oriented theorizing, in his 'Weyl and von Neumann: symmetry, group theory, and quantum mechanics', PITT-PHIL-SCI00000409.
[29] See Bub (1981).
[30] Barbour (1999). Barbour applies the phrase not to physics, but to the book itself.

References

Barbour, J. (1999). *The End of Time: The Next Revolution in Our Understanding of the Universe.* London: Weidenfeld & Nicholson.
Belot, G. (2001). 'The Principle of Sufficient Reason'. *Journal of Philosophy*, **98**, 55–74.
Bub, J. (1981). 'Hidden variables and quantum logic – a skeptical review'. *Erkenntnis*, **16**, 275–93.
Glymour, C. (1980). *Theory and Evidence*. Princeton: Princeton University Press.
Huggett, N. (1999). 'Atomic metaphysics'. *Journal of Philosophy*, **96**, 5–24.
Jammer, M. (1961). *Concepts of Mass in Classical and Modern Physics*. Cambridge, MA: Harvard University Press.
Lewis, D. K. (1983). 'New work for a theory of universals'. *Australasian Journal of Philosophy*, **61**, 343–77.
Mermin, D. (1998). 'What is quantum mechanics trying to tell us?' *American Journal of Physics*, **66**, 753–67.
Poincaré, H. (1952). *Science and Hypothesis*. New York: Dover.
Rynasciewicz, R. (2001). 'Definition, convention, and simultaneity: Malament's result and its alleged refutation by Sarkar and Stachel'. *Philosophy of Science*, **68**, S345–57.
Russell, B. (1897). *An Essay on the Foundations of Geometry*. Cambridge: Cambridge University Press.
Sneed, J. (1971). *The Logical Structure of Mathematical Physics*. Boston: Reidel.
van Fraassen, B. C. (1985). *An Introduction to the Philosophy of Time and Space.* 2nd edn. New York: Columbia University Press.
(1991). *Quantum Mechanics: An Empiricist View*. Oxford: Oxford University Press.
Weyl, H. (1952). *Symmetry.* Princeton, NJ: Princeton University Press.