Invariance, Intrinsicality and Perspicuity

Caspar Jacobs*

March 26, 2022
Forthcoming in Synthese

Abstract

It is now standard to interpret symmetry-related models of physical theories as representing the same state of affairs. Recently, a debate has sprung up around the question when this interpretational move is warranted. In particular, Moller-Nielsen (2017) has argued that one is only allowed to interpret symmetry-related models as physically equivalent when one has a characterisation of their common content. I disambiguate two versions of this claim. On the first, a perspicuous *interpretation* is required: an account of the models' common ontology. On the second, stricter, version of this claim, a perspicuous *formalism* is required in addition: one whose mathematical structures 'intrinsically' represent the physical world, in the sense of Field (1980). Using Dewar (2019)'s distinction between internal and external sophistication as a case study, I argue that the second requirement is decisive. This clarifies the conditions under which it is warranted to interpret symmetry-related models as physically equivalent.

Acknowledgements: I would like to thank Marta Bielińska, Adam Caulton, Neil Dewar, Henrique Gomes, Oliver Pooley and James Read for their insightful comments and discussions. I would also like to thank an audience at the Philosophy of Physics Graduate Lunch Seminar in Oxford and a helpful anonymous reviewer.

1 Introduction

The presence of symmetries in physical theories throws into stark relief the conflicted relationship between a theory's formalism and its interpretation.

*University of Pittsburgh, 1101 Cathedral of Learning, Pittsburgh, PA 15260, United States, caspar.jacobs@pitt.edu

On the most literal-minded view, there is a one-to-one map between elements of the theory's formal apparatus on the one hand, and elements of the physical world on the other. But symmetries seem to suggest that some elements of the formalism are representationally redundant, so that the proper mapping is in fact many-to-one. On this consensus view, symmetry-related models are physically equivalent: they represent the same state of affairs. The differences between such models are 'distinctions without a difference'.[1]

This raises the question: *under which conditions* are we warranted to declare symmetry-related models as physically equivalent? Recently, a lively debate has sprung up around this question.[2] In particular, Møller-Nielsen (2017) has distinguished between an *interpretational* and a *motivational* approach to symmetries.[3] In brief, on the interpretational view we are warranted to consider symmetry-related models as physically equivalent even in the absence of a 'perspicuous metaphysical characterisation' of their common physical content, while on the motivational view such a characterisation is required.

However, it remains somewhat unclear what counts as a perspicuous metaphysical picture: the phrase is ambiguous between two distinct notions. On the first notion, one must be able to tell a coherent story about the ontology (and ideology) common to classes of symmetry-related models; only then can they be interpreted as physically equivalent. But on the second, stricter notion, it is additionally required that one reformulates the theory in terms of a mathematical structure which more perspicuously *represents* this common ontology. The difference is easily illustrated. Consider a relationist who interprets shift-related models of Newtonian Gravitation—models which differ just over the absolute location of material bodies—as physically equivalent. Despite the fact that these models describe trajectories on a differentiable manifold, the relationist claims that their physically relevant content consists just of the distances between particles. This is a perspicuous metaphysical picture in the sense that (i) it offers a coherent account of the ontology which the theory's models are supposed to represent, and (ii) this ontology is invariant under the theory's symmetries. But there is another sense in which the theory so interpreted is not perspicuous, as the

[1] For various statements to this effect, see Ismael and van Fraassen (2003); Saunders (2003); Baker (2010); Caulton (2015); Dasgupta (2016).

[2] See, *inter alia*, Dewar (2019), Martens and Read (2020), Møller-Nielsen (2017), Russell (2018), Sider (2020) and references therein.

[3] This is not to say that the debate started with Møller-Nielsen's distinction. For instance, Earman (1989, 127) already called an approach similar to interpretationalism a 'cheap instrumentalist rip-off'.

fundamental quantities of the theory represent (absolute) positions rather than distances. The theory's formalism does not match the metaphysical picture provided. That is still a bit vague, but a first approximation would be that the 'components' of the theory's models (predicates, relations, functions, etc.) do not directly correspond to its metaphysical posits.[4]

In brief, then, the first notion concerns whether the theory has a perspicuous *interpretation*, the second whether it has a perspicuous *formalism*. The former criterion has received most attention, but I will argue that it is often the latter which is decisive. I will defend this claim a case study of Dewar's (2019) distinction between 'internal' and 'external' sophistication. Both are approaches to the interpretation of symmetry-related models, but Martens and Read (2020) have argued that while internal sophistication can provide a perspicuous metaphysical characterisation, external sophistication cannot. I will show that this is false if a perspicuous characterisation concerns the theory's interpretation, but true if it concerns the theory's formalism.

Moreover, I will use Field's (1980) notion of 'intrinsic' theories to defend the strict version of motivationalism, which requires a perspicuous formalism in addition to a perspicuous interpretation. This lesson extends to theories more broadly. If I am correct, the proper demand of motivationalism is a demand for intrinsic theories, in Field's sense.

2 Symmetries, Motivationalism, Sophistication

In this section I briefly discuss some important concepts and distinctions: symmetries (§2.1), motivationalism (§2.2), and the distinction between internal and external sophistication (§2.3).

2.1 Symmetries

In order to define symmetries, it is useful to distinguish between the 'kinematically possible models' (KPMs) and 'dynamically possible models' (DPMs) of a theory. The former are mathematical structures that are of the right form. The KPMs that satisfy the dynamical equations of the theory are called DPMs. DPMs represent ways the world could be if the theory were true. Put in these terms, symmetries are transformations of the KPMs that map the space of DPMs onto itself. In other words, symmetries preserve the

[4] This point is related to Arntzenius' (2012, §5.7) objection that relationism 'piggy-back' on the substantivalist formalism. Arntzenius objects to piggy-back relationism on the account that it is not simple. I will offer a different reason against this practice, namely that it is not 'intrinsic' in a Fieldian sense.

dynamics. When a theory contains symmetries, the space of KPMs is thus partitioned into equivalence classes of symmetry-related models (SRMs).

It is usually thought that SRMs are empirically indistinguishable, although it is contested whether this is the case for all symmetries.[5] But even if this is only true for some symmetries, it *is* the case for the symmetries of our most successful physical theories, such as general relativity and (classical) gauge theory. If it turns out that, on some reasonable definition of symmetries, not all SRMs are empirically equivalent, then we can just restrict our attention to those symmetries which in addition satisfy a condition of empirical equivalence. From this, the conclusion is usually drawn that symmetries are, in the words of Ismael and van Fraassen (2003), a guide to superfluous theoretical structure. Consequently, SRMs are physically equivalent.

It is helpful to have an example of a symmetry. Consider Newtonian Gravitation set in Galilean spacetime. The KPMs of this theory are of the form $\langle M, t_{ab}, h^{ab}, \nabla, \varphi, \rho, \xi^a \rangle$, where M is a four-dimensional smooth manifold diffeomorphic to $\mathbb{R}^4$; t_{ab} and h^{ab} are compatible temporal and spatial metrics ($t_{an}h^{nb} = 0$); ∇ is a covariant derivative encoding a standard of uniform motion compatible with both metrics ($\nabla_a t_{bc} = \nabla_a h^{bc} = 0$); φ and ρ are scalar fields that represent the gravitational potential and the matter distribution respectively; and ξ^a is a time-like vector field that represents the four-velocity of a test particle.[6] The DPMs of Newtonian Gravitation satisfy the following dynamical equations:

$$R^a{}_{bcd} = 0 \tag{1}$$

$$h^{ab}\nabla_a\nabla_b\varphi = 4\pi\rho \tag{2}$$

$$-\nabla^a\varphi = \xi^b\nabla_b\xi^a \tag{3}$$

where $R^a_{bcn}\xi^n = \nabla_{[b}\nabla_{c]}\xi^a$. Here, (1.1) imposes flatness on ∇; (1.2) is the Newton-Poisson equation; and (1.3) is Newton's second law for a test particle. The DPMs represent worlds in which Newtonian Gravitation is true.

The so-called *kinematic shifts* are symmetries of this theory. Let d denote a diffeomorphism which enacts a velocity boost: expressed in inertial coordinates it acts as the map $\vec{x} \to \vec{x}+\vec{v}t$ for some constant vector $\vec{v}$. The fact that such transformations are symmetries means that $\langle M, t_{ab}, h^{ab}, \nabla, \varphi, \rho, \xi^a \rangle$ is a

[5] Belot (2013) and Dasgupta (2016) express scepticism, but see Wallace (2019a) for a response.

[6] For more details, see Malament (2012, §4.2).

DPM whenever $\langle M, t_{ab}, h^{ab}, \nabla, d^*\varphi, d^*\rho, d^*\xi^a \rangle$ is, where d^* denotes the push-forward map induced by d.[7] The latter model represents a world in which all matter fields are boosted. The question that we are concerned with is under which circumstances we are warranted to regard such pairs of models as physically equivalent.

2.2 Motivationalism

I will now discuss Møller-Nielsen's (2017) distinction between interpretationalism and motivationalism in more detail. According to interpretationalism, SRMs can always be interpreted as physically equivalent, even in the absence of a 'perspicuous metaphysical characterisation' of their common content. According to motivationalism, more is needed. But as I mentioned, it is still unclear what that 'more' consists of. In this subsection, I will consider two relevant precisifications: the requirement of a perspicuous *interpretation*, and of a perspicuous *formalism*.

Firstly, what may be required is an account of the metaphysical commitments of the theory—a statement of its ontology and ideology—which explains the physical equivalence of SRMs. This view is suggested by Møller-Nielsen himself, who writes of "a metaphysically perspicuous characterization of the reality that is alleged to underlie symmetry-related models" (1256). In the same vein, Martens and Read (2020, 6) speak of "a coherent metaphysical picture of the common ontology underpinning their [SRMs] equivalence". Once one specifies a metaphysics which is common to SRMs, one has thereby explained their equivalence. For instance, once one posits that distances and relative velocities between particles are fundamental, the fact that these quantities are invariant under kinematic shifts explains the equivalence of shift-related models.

However, the examples these authors use to illustrate their position suggest a different account of perspicuity. Møller-Nielsen explicitly argues that motivationalism often requires a novel mathematical formalism. For example, consider his discussion of electrodynamics. This theory admits of (at least!) two formulations: one in terms of the Faraday tensor, F_{ab}, and one in terms of the vector potential, A_a. The behaviour of A_a compared to F_{ab} under symmetries parallels the behaviour of absolute locations compared to distances: the former vary under symmetry transformations while the latter remain the same. It is therefore natural to claim that the Faraday tensor represents a physical field, whereas the vector potential contains redundant

[7] This follows Earman's (1989) definition of what he calls *dynamical symmetries*.

degrees of freedom. Since one can define the Faraday tensor in terms of the vector potential, it would seem that one can easily offer a perspicuous metaphysical characterisation of electrodynamics *as formulated in terms of the vector potential*. In particular, for any model of that theory one can indirectly calculate the Faraday tensor and declare that it represents the physical degrees of freedom of the model in question. This is analogous to declaring that only the distances represented in models of Newtonian Gravitation are physically real. Of course, in this case it is almost trivial to formulate a reduced theory directly in terms of the Faraday tensor, which is not the case for Newtonian Gravitation in terms of distances. But consider an alternate history in which the laws of electrodynamics were *first* formulated in terms of A_a; in that world a reformulation in terms of F_{ab} would not have been trivial. Since F_{ab} is invariant under symmetries, this immediately explains the physical equivalence of SRMs of electrodynamics.

But Møller-Nielsen argues that these facts motivate not just a different interpretation of electrodynamics, but a different formalism: "it is [the tensor formulation] that, I take it, constitutes the metaphysically perspicuous characterization of this theory [...] the vector potential A_a does not directly represent a genuinely real field: rather, it is merely a mathematically convenient 'shorthand' way of characterizing and determining the values of the Faraday tensor, which is taken to represent the genuine material ontology of the theory" (1258). This suggests a second account of perspicuity. On this account, an appropriate new formalism (rather than just an interpretation of the old formalism) is required. In particular, one which is formulated in terms of mathematical entities which 'directly' represent physical fields. Thus, the tensor-formulation of electrodynamics succeeds because F_{ab} directly represents the EM field, while the potential-formulation fails because A_a only indirectly (that is, redundantly) represents the same field.

I believe that both forms of perspicuity play a role in the interpretation of symmetries. But whereas Møller-Nielsen, Martens and Read have either displayed ambivalence between them or emphasised the former, I believe that the latter can also be decisive.[8] I will show this with a case study of sophis-

[8] As the discussion above illustrates, Møller-Nielsen is at least somewhat sensitive to the need for a perscipuous formalism, although his official statement of motivationalism as the demand for a *metaphysical* chracterisation suggests that his main concern still the theory's interpretation. Martens and Read also focus on the need for a perscipuous interpretation, as is clear from their criticism of external sophistication discussed below. In a more recent article, Read explicitly states that he is "no longer convinced that mathematical reformulation is necessary, even for models which are not isomorphic, in order to secure what the motivationalist calls a 'metaphysically perspicuous characterisation'" (Read,

tication in the next two sections. Martens and Read (2020) attack Dewar's 'external' approach to sophistication on the basis that it does not offer a perspicuous interpretation. I will argue that this is not the full story. There is a sense in which external sophistication does offer an interpretation of the theory, that is, an account of the theory's fundamental metaphysical commitments, as I will show in §3. However, these commitments are presented *indirectly*, and it it this which poses a problem for external sophistication. I thus agree with Martens and Read that external sophistication fails. My criticism differs from theirs in that I emphasise that external sophistication fails as a consequence of its non-perspicuous formalism.

2.3 Sophistication

Dewar (2019) introduces sophistication as an approach to the interpretation of SRMs. He contrasts sophistication with the standard account, *reduction*: "this account says that we should seek a *reduced* theory: a theory which deals only in quantities which are invariant under the relevant symmetry" (486). For example, a theory of gravitation formulated in terms of distances is an instance of reduction, since distances are invariant under shifts. In contrast, Dewar argues that "we need not insist on finding a theory whose models are invariant under the application of the symmetry transformation, but can rest content with a theory whose models are isomorphic under that transformation" (498). This latter approach he calls *sophistication*, after sophisticated substantivalism in the context of spacetime theories.[9]

For an example, consider the move from Newtonian to Galilean spacetime. The models of Newtonian Gravitation set on the former kind of spacetime are of the form $\langle M, t_{ab}, h^{ab}, \sigma_a, \varphi, \rho, \xi^a \rangle$. Here, σ_a provides a standard of identity for points in space across time. From this object, one can derive a standard of absolute rest. This means that Newtonian spacetime has more structure than Galilean spacetime, which only contains a standard of absolute acceleration in the form of the covariant derivative ∇. The structure σ_a is not invariant under boosts ($\sigma_a \neq d^*\sigma_a$), and hence boost-related models of Newtonian Gravitation set on Newtonian spacetime are non-isomorphic. The covariant derivative, on the other hand, *is* invariant under boosts ($\nabla = d^*\nabla$), and hence boost-related models of the same theory set on Galilean spacetime *are* isomorphic. This allows us to interpret the theory anti-haecceitistically: spacetime points are qualitatively individuated. Since isomorphic models agree on all qualitative features, anti-haecceitism

2021, 15).

[9] For more on the latter, see Pooley (2013) and references therein.

entails that they represent the same state of affairs. In this way, the physical equivalence of SRMs is explained by the doctrine of anti-haecceitism. The general procedure for sophistication is to restructure the theory's models such that SRMs are isomorphic, and then declare anti-haecceitism to interpret them as physically equivalent.

Next, Dewar draws a distinction between what he calls the *internal* and the *external* approach to sophistication. Both approaches share the aim of redefining the theory's models such that SRMs become isomorphic. But on the internal approach one does so by specifying a particular Tarski-style interpretation of the theory's formalism, while on the external approach one gives an 'extrinsic' characterisation via a group of transformations.[10] In Dewar's words: "rather than trying to define the objects of the new semantics 'internally', as mathematical structures of such-and-such a kind (paradigmatically, as sets equipped with certain relations or operations), we instead define them 'externally': as mathematical structures of a given kind, but with certain operations stipulated to be homomorphisms (even if they're not 'really' homomorphisms of the given kind)" (502).

I will now elaborate on these approaches below.

2.3.1 Internal Sophistication

The procedure for internal sophistication is to (i) define mathematical objects which represent physical quantities; (ii) write down a dynamics in terms of these objects; and (iii) ensure that the symmetry-related models of these dynamics are isomorphic. Whether (iii) holds depends on a judicious choice for (ii) and especially (i). For instance, whether the models of Newtonian Gravitation are isomorphic depends on the mathematical objects one uses to express the dynamics: if one writes down the dynamics in terms of ∇ they are, but if one employs σ_a they are not.

Internal sophistication provides both a perspicuous metaphysics and a perspicuous formalism. In fact, the latter leads to the former. The formalism is such that each of the objects defined at (i) represents part of the theory's ontology/ideology. Hence, one can 'read off' the theory's metaphysics from the formalism: the theory wears its interpretation on its sleeves. However, this also means that it is often difficult to construct an internally sophisticated theory, for one first has to find a satisfactory set of objects within which to couch the dynamics, such that SRMs are isomorphic. It is for this reason that Dewar offers an alternative, namely external sophistication.

[10] A similar distinction was in fact already drawn in Suppes (1967).

2.3.2 External Sophistication

In discussing a related issue, Møller-Nielsen (2017, 1262) alleges that we cannot assume that "there will always be [an internally sophisticated theory] waiting in logical space to be discovered". This sounds true enough: the move from Newtonian to Neo-Newtonian spacetime was certainly non-trivial. But as Wallace (2019b, 16) points out, this "underestimates the powerful, general resources by which structure can be subtracted from mathematical theories". The external approach to sophistication aims to make full use of these resources and thereby 'brute force' an isomorphism. The aim of external sophistication is to define structures *in terms of their isomorphisms*, such that they are by definition invariant under the theory's symmetries. Dewar has put this somewhat puzzlingly as "declaring, by fiat, that the symmetry transformations are now going to 'count' as isomorphisms" (Dewar, 2019, 502-3). But this obfuscates what is really going on, namely that new structures are defined in terms of the relations of isomorphism between them.[11]

Wallace (2019b) contains a clear exposition of what these structures are like. Here I adapt this treatment to the case of kinematic shifts. We start with an arbitrary coordinatisation x of spacetime, i.e. an injective function from M into $\mathbb{R}^4$. Then, we construct an equivalence class $[x]$ from x, such that $x' \in [x]$ iff $\vec{x}' = \vec{x} + \vec{v}t$ for some constant vector $\vec{v}$ (where $\vec{x}$ are the spatial coordinates of x). In other words, $[x]$ specifies which transformations on M are 'postulated' as isomorphisms. This results in a spacetime structure $\langle M, [x] \rangle$, which is (by construction) invariant under kinematic shifts.[12] The models of Newtonian Gravitation are then of the form $\langle M, [x], \varphi, \rho, \xi^a \rangle$. Wallace (2019b, 128) shows that such a structure is unique up to isomorphism. Therefore, the external approach defines a more-or-less unique structure which is invariant under the relevant symmetry transformations.

The crucial difference between the internal and the external approach is that the latter does not wear its interpretation on its sleeves. The functions in $[x]$ represent arbitrary coordinate systems, rather than real spacetime

[11] The symmetry-first approach is thus closely associated with the category-theoretic approach to theoretical equivalence. On this approach, one defines a theory by specifying both a class of models and a set of arrows between these models. When there is an arrow between a pair of models, we can treat them 'as if' they are isomorphic. On this view, theories are physically equivalent iff they are equivalent as categories. I lack the space to consider this approach in detail, but I believe that many of my arguments against external sophistication carry over to the category-theoretic approach.

[12] Formally, $[x] = d^*[x]$, where $d^*[x]$ is defined such that $d^*x \in d^*[x]$ iff $x \in [x]$

structure. The 'components' of the theory's formalism therefore don't directly represent the theory's ontology/ideology. In §4, I will explicate this claim in terms of Field's (1980) notion of intrinsicality. But as I will show now, one can in fact extract genuine metaphysical commitments from an externally-sophiciated formalism.

3 Perspicuous Interpretation

The external approach to sophistication has been criticised on the grounds that it does not offer a perspicuous metaphysical characterisation. Møller-Nielsen (2017), for instance, argues that on interpretationalist approaches in general: "it is simply opaque what [...] the world is really like" (Møller-Nielsen, 2017, 1264). On the topic of external sophistication in particular, Martens and Read (2020) write:

> We take it that a complete and honest form of realism should not only take these questions [about metaphysical commitment] on board, but consider them to be crucial. Leaving them out is at best a dishonest form of realism, and at worst a form of anti-realism. (Martens and Read, 2020, 27)

In terms of the distinction drawn in §2.2, these authors claim that external sophistication is defective because it fails to offer a perspicuous *interpretation*. On the contrary, I argue that external sophistication *does* have definite metaphysical commitments. It is an overstatement to call the approach "a form of anti-realism". Neither is it correct to claim that external sophistication is a 'dishonest' form of realism: the symmetry-first approach does not wear its metaphysical commitments on its sleeves, but it does not *mis*represent them either.

In the below, I will show that (a) external sophistication has an effective *decision procedure* for which fundamental structure it is ontologically committed to, and (b) from this decision procedure, it follows that external sophistication has the same ontological commitments as internal sophistication. In the next section I will argue that it is not the lack of a perspicuous interpretation but of a perspicuous *formalism* that renders external sophistication unsatisfactory.

In order to formulate the decision procedure, I will assume the following criterion of ontological commitment: if a theory posits a certain structure, then the theory is committed to whatever one can define in terms of that structure. In a sense, all definable structure 'comes for free'. I will call this

'derivative' or 'conditional' commitment. The criterion follows the spirit of Butterfield's (2011) identification of supervenience with (implicit) definability. This identification is motivated by the idea that implicitly definable structure is *already there* in the theory: when God created the world's most basic structure, it didn't require any further act of creation to bring into existence any further structure which is definable from this basic structure. This criterion of conditional commitment is commonplace in philosophy of physics. It is often expressed as the thought that (only) definable structure is physically meaningful or 'objective'.[13] For example, Malament (1977) argued that because a unique standard of simultaneity is definable from the structure of Minkowski spacetime, simultaneity in special relativity is objective rather than conventional. I simply intend to extend this criterion to physical structure more broadly.

The criterion clearly departs from Quine's view that one is committed only to whatever entities one quantifies over. There are many properties and relations which are definable in terms of a theory's fundamental posits, but which are not themselves amongst these posits. Some of those properties and relations are famously problematic, for instance the property of being grue. The criterion of commitment which I propose implies that such properties are real, albeit in a derivative sense; it is consistent with this claim that the property of being green is more fundamental than the property of being grue (i.e. the property of being observed before time t and green, or not being observed before time t and blue). But the departure from Quine here is no surprise, since the external approach does not directly quantify over physical structure. Instead, it defines this structure extrinsically in terms of invariance under certain symmetries. Therefore, a non-Quinean criterion is apposite in order to draw out the external approach's ontological commitments.

In terms of the criterion of conditional commitment, the question is this: which relations are definable in terms of a structure such as $\langle M, [x] \rangle$? In response to this question, I will prove that any symmetry-invariant piece of structure is (implicitly) definable from $\langle M, [x] \rangle$, so the ontological commitments of internal and external sophistication are the same. I will rely on the following theorem due to Barrett (2017, Theorem 2).[14] Here, $\mathfrak{m}$ is a model of a theory T and R is some particular relation. We define $\Sigma = \mathbf{R} \setminus R$, where $\mathbf{R}$ is a set of relations defined over the domain D of $\mathfrak{m}$. In other words, $\mathfrak{m}|_\Sigma$

[13] Cf. Mundy (1986); Debs and Redhead (1996); Barrett (2017).

[14] For a precursor of this theorem, see Earman (1989, 58-60).

is the *reduct* of $\mathfrak{m}$ obtained by removing R.[15] Barrett's theorem then states:

> **If**, for any models $\mathfrak{m}, \mathfrak{n}$ of T, if $h : \mathfrak{m}|_\Sigma \to \mathfrak{n}|_\Sigma$ is an isomorphism then $h[R^\mathfrak{m}] = R^\mathfrak{n}$ **then**, for any models $\mathfrak{m}, \mathfrak{n}$ of T, if $\mathfrak{m}|_\Sigma = \mathfrak{n}|_\Sigma$ then $R^\mathfrak{m} = R^\mathfrak{n}$.

The antecedent captures the idea that some piece of structure R is *invariant* under the symmetries of a class of models. Suppose that we start with a model $\mathfrak{m}|_\Sigma$ and append some piece of structure R—for example, a new relation on the model's domain. We then consider the isomorphisms of the *original* model. If, under these isomorphisms, the extension of $R^\mathfrak{m}$ carries over to the extension of $R^\mathfrak{n}$, then R is invariant under the symmetries of $\mathfrak{m}|_\Sigma$. The consequent captures the idea that R is *implicitly definable* from $\mathfrak{m}|_\Sigma$, in the sense that any models which agree on the structure of $\mathfrak{m}|_\Sigma$ must also agree on the extension of R. Barrett's theorem thus states that a piece of structure is implicitly definable from the theory's solutions if it is invariant under the isomorphisms of these models. The theorem does *not* depend on the assumption that T has first-order formulation.[16] We can use the theorem even if we adopt the semantic view on which theories are specified directly by their class of models, as we have done here. That being said, the theorem is limited insofar as it presumes that the theory's 'language' consists of ordinary predicates and relations. Further work is needed to apply the theorem to theories formulated in the language of differential geometry; but at the very least the theorem is highly suggestive.[17]

Here is an example to clarify the theorem. Consider once more the 'extrinsic' models of Newtonian Gravitation of the form $\langle M, [x], \varphi, \rho, \xi^a \rangle$, where $[x]$ is a set of coordinate systems closed under kinematic shifts. Denote these models $\mathfrak{m}|_\Sigma$. Now, consider an extra piece of structure $R \coloneqq \nabla$, which is an affine connection on M. The unreduced models $\mathfrak{m}$ are then of the form $\langle M, [x], \varphi, \rho, \xi^a, \nabla \rangle$. For the proof, suppose that if it is the case that $d : \langle M, [x], \varphi, \rho, \xi^a \rangle \to \langle M, [x], d^*\varphi, d^*\rho, d^*\xi^a \rangle$ is an isomorphism, then $d^*\nabla^\mathfrak{m} = \nabla^\mathfrak{n}$ (i.e. the LHS of Barrett's theorem). This is indeed the case, since ∇ is invariant under kinematic shifts. Furthermore, suppose that $\mathfrak{m}|_\Sigma = \mathfrak{n}|_\Sigma$ (i.e. the antecedent of the RHS). In that case, the identity map $I : \mathfrak{m}|_\Sigma \to \mathfrak{n}|_\Sigma$

[15] The technical term 'reduct' here has no relation to the process of reduction discussed earlier.

[16] Barrett's proof of the converse theorem *does* depend on that assumption, but see Dewar (2022, Prop. 1.16) for a purely model-theoretic proof.

[17] For attempts to 'reduce' differential geometry to more set-theoric constructions, see foundational treatments such as Malament (2012) and Arntzenius and Dorr (2012).

is an isomorphism. So, from the LHS of Barrett's theorem, $\nabla^{\mathfrak{m}} = I^{*}\nabla^{\mathfrak{m}} = \nabla^{\mathfrak{n}}$. Therefore, if $\mathfrak{m}|_{\Sigma} = \mathfrak{n}|_{\Sigma}$, then $\nabla^{\mathfrak{m}} = \nabla^{\mathfrak{n}}$, so ∇ is implicitly definable from $\mathfrak{m}$. The same holds for any other boost-invariant relations, such as the temporal and spatial metrics defined above.

This proves the first point, that external sophistication has an effective decision procedure for ontological commitment. The second point—that the metaphysical commitments of internal and external sophistication are the same— follows. For on both the internal and the external approach models are constructed such that the symmetries of the theory are isomorphisms between them. This means that *exactly the same structure is (implicitly) definable on* both *approaches*. If definability is our criterion for (conditional) ontological commitment, then these approaches have the same commitments in virtue of their invariance under the same group of transformations. The worry could be raised that the external approach is committed to strictly more structure than the internal approach, since the variant structure from which the models are defined is also (trivially) definable.[18] But this misunderstands external sophistication. The fundamental posits in terms of which models are defined on this approach are not the (variant) coordinatisations x, but their equivalence classes $[x]$. This equivalence class is fully invariant under the theory's symmetries. It may still seem as if there is a certain tension between a commitment to the invariant structure $[x]$ and a lack of commitment to the variant functions x which constitute this structure. In the next section I will argue that it is essentially this tension that should lead us to reject external sophistication.

Moreover, given a class of *non-isomorphic* models with the same domain, external sophistication is just committed to exactly that structure which is invariant under the theory's symmetries, since it is this structure which is definable from the theory's models. This explains the sense in which we can act 'as if' SRMs are isomorphic: we simply commit to their symmetry-invariant content. Consider the shift from Newtonian to Galilean spacetime. The trans-temporal identities of spacetime points are variant under the boost symmetries of Newtonian Gravitation. It follows that a commitment to just the part of the structure of Newtonian spacetime that is invariant under Galilean transformations rules out the existence of such trans-temporal identities. Importantly, this is the case even if no intrinsic account of Galilean spacetime is offered. On the external approach, Galilean spacetime is simply defined as the structure that is obtained from Newtonian spacetime if one 'forgets' the trans-temporal identifications of points.

[18] I thank an anonymous reviewer for bring this to my attention.

This approach follows mathematical practice, which for instance defines an affine space as "nothing more than a vector space whose origin we try to forget about" (Berger, 1987, 32).

So, external sophistication is not quite a metaphysical black box. The external approach *does* offer a perspicuous metaphysics, contrary to the claims of some motivationalists. Nevertheless, an important difference between the two approaches remains. The internal approach presents us with a set of invariant relations which directly represent the physical relations the theory is committed to. The external approach, on the other hand, only provides us with a decision procedure to determine if it is committed to some given piece of structure. It then declares a blanket commitment to any such structure, but we do not know in advance which functions and relations *are* invariant under the theory's symmetries. In this sense the symmetry-first approach is a little like a Freedom of Information request: in order to get the answers you are after, you first have to know the right questions to ask.

4 Perspicuous Formalism

In this section I will explain in more detail the importance of having a perspicuous *formalism*. I will present three criticisms: (a) that external sophistication does not provide us with causal explanations of physical phenomena (§4.1); (b) that it assumes the representational equivalence of SRMs as a brute fact (§4.2); and (c) that its account of reality involves physically inert structure (§4.3). I claim that these issues arise because external sophistication does not characterise a theory's models *intrinsically*, in the sense of Field (1980). Therefore, the proper demand of motivationalism should be a demand for intrinsic theories.

4.1 Causal Explanations

We are not only interested in a theory's ontology for its own sake. We also use a theory's ontological posits to explain physical events. But as Møller-Nielsen argues, "it is simply not clear what causal-explanatory, realistic picture of the world is being propounded by the defender of the interpretational view" (Møller-Nielsen, 2017, 1264). The first worry is that without transparent ontological commitments, attempts at scientific explanation are obstructed.[19] The problem is not that external sophistication is not meta-

[19] The demand for an explanatory powerful interpretation is in line with what Martens and Read's (2020) 'strong' motivationalism, which I am sympathetic with.

physically committed to a causal story in the first place—the arguments from the previous section have shown the contrary. Rather, the issue is that one cannot read off this story from the theory's formalism, which severely limits our attempts at scientific explanation.

It is unclear, for example, how the externalist can explain the Aharonov-Bohm effect, in which a charged particle that moves around an impenetrable solenoid picks up a phase which is proportional to the flux through the solenoid. The proponent of external sophistication cannot without further argument appeal to the fact that the *holonomies* of the vector potential field are invariant under gauge transformations, since this first requires an expression of holonomies in terms of a representation function from spacetime points to complex numbers. Of course, once one has such an expression it follows from Barrett's theorem that the external approach is committed to the reality of holonomies. But this is a non-trivial conclusion which goes beyond the slogan that we are committed to *whatever it is* that is invariant under the theory's symmetries.

Another example concerns the structure of Maxwellian spacetime, which includes so-called dynamic shifts as spacetime symmetries.[20] Earman (1989) originally defined the structure of Maxwellian spacetime in terms of an equivalence class $[\nabla]$ of covariant derivative operators. But none of the elements of this class are supposed to represent geometrical structure (since each individual ∇ varies under dynamic shifts), and it is therefore unclear what the physical explanation of a particle's trajectory in Maxwellian spacetime consists of. In essence, Earman defines the standard of rotation as 'whatever it is that is common to all $\nabla \in [\nabla]$', but this is not a closed definition of any geometrical object. In contrast, Weatherall (2017) offers an intrinsic characterisation of the standard of rotation. This allows us to define the structure of Maxwellian spacetime in a structure-first way. Although the laws of Newtonian Gravitation have yet to be expressed in terms of this object, Weatherall shows that one can use his standard of rotation to define a notion of relative acceleration which could enter in explanations of physical phenomena such as Newton's famous bucket experiment.

4.2 Unexplained Equivalences

The second criticism of the symmetry-first approach is discussed by Martens and Read (2020): "it is often unclear what grounds or explains or justifies the physical equivalence of models that, on a natural interpretation, repre-

[20] I thank James Read for suggesting this example.

sent distinct possible worlds" (26). Again, the problem here is not that on the external approach to sophistication there is no metaphysical account of this equivalence. Indeed, the account is straightforward: external sophistication is ontologically committed to whatever is invariant under the theory's symmetries, so SRMs are physically equivalent by stipulation.

The issue is that this reverses the natural order of explanation. The external approach first defines SRMs as physically equivalent, and then extracts their physical content from this stipulation. This puts the cart before the horse. We are not committed to some physical structure because certain models are called physically equivalent; we call these models physically equivalent because of our metaphysical commitments! The physical equivalence of SRMs is a thesis about our representations *of* the world, whereas the theory's physical posits concern the world itself. But of course the world does not contain particular physical structures *because* certain representations are considered equivalent.

One might respond that the claim that SRMs are physically equivalent may follow from a prior commitment to the reality only of structure which is preserved under maps that preserve the theory's *observable* content.[21] However, this only pushes back the question. For if we ask *which* structure is preserved under these maps, the answer is that it is just the structure which is common to equivalence classes of SRMs. Again, the theory's ontological content is encoded in a claim about the representational capacities of the theory's models.

Therefore, external sophistication fails to account for an important asymmetry between a theory's interpretation and formalism. The slogan that we are committed to whatever is invariant across SRMs makes it seem as if the former depends on the latter, where in fact the reverse is true.

4.3 Intrinsic Structure

Finally, external sophistication fails to give an *intrinsic* characterisation of physical structure. This Fieldian objection has not been as prominent in the literature so far, but captures an important source of dissatisfaction with interpretationalist approaches.[22] It seems to me that Field was concerned with exactly the kind of questions that are at issue in debates around symmetries, but this connection has largely gone unnoticed. The core objection is that the external approach characterises the physical content of models in terms of the mathematical relations *between* these models; for instance,

[21] I thank an anonymous reviewer for this point.

[22] Sider (2020) and North (2021) are two recent exceptions.

their isomorphisms. Yet these relations themselves are unphysical, as they merely relate abstract mathematical structures. The worry is that a definition of the physical content of a theory's models in terms of the non-physical relations between these models is in some sense 'impure'. It is as if someone were to characterise the referent of the name 'Newton' as "the man whom we call 'Newton'". Perhaps this fixes the correct extension for the name 'Newton', but it does not seem to tell us anything about who Newton really is.

Field's intrinsicalist programme instead requires us to characterise a theory's structure in purely physical terms. We are required to lay down a set of relations, functions and operators which explicitly represent the world's physical structure. My aim here is not to resurrect Field's programme in its entirety. Instead, I want to emphasise his insight into the importance of intrinsically-defined theories, which is under-appreciated in the contemporary literature. Why accept Field's requirement of intrinsicality? A number of motivations can be extracted from his seminal book *Science Without Numbers* (Field, 1980). The first is a commitment to nominalism about mathematical entities. If theories are intrinsically formulated, then all of its terms refer to physical quantities. However, as Chen (2018, fn. 7) has pointed out, "the intrinsicalist and nominalistic visions can also come apart. For example, we can, in the case of mass, adopt an intrinsic yet platonistic theory of mass ratios." Since mass ratios, which are identified with positive real numbers, are invariant under mass scalings, they do not depend on arbitrary conventions. They intrinsically characterise the second-order relations between determinate mass values. It is therefore not the appeal to numerical structure *per se* that poses a problem for symmetry-first sophistication.

The second motivation concerns the arbitrariness of the coordinate functions f, in the sense that *any* $x' \in [x]$ represents a quantity's structure equally well. The intrinsicalist's aim, on the other hand, is "to explain, in terms of intrinsic facts [...] which are statable without such arbitrary choices, why the choice of functions to be invoked in the extrinsic theory will be arbitrary to precisely the extent that it is" (Field, 1980, 46). But although each individual x is arbitrarily chosen, it is not the case that the equivalence class $[x]$ is arbitrary as a whole. After all, $[x]$ is the unique equivalence class of representation functions that is invariant under the dynamical symmetries of the theory, as we have seen in §2.3.2. It is therefore not obvious that external sophistication as a whole suffer from arbitrariness, although it is still plausible that a collection of individually arbitrary functions is unsatisfactory from a realist's point of view.

There is a third Fieldian concern which I believe to be more basic, namely

the claim that the functions in $[x]$ are not *physically* relevant. The x are functions from (for example) the manifold M into some system of coordinates or units, such as $\mathbb{R}^4$. But coordinates themselves play no causal role; they are physically inert. Field expresses this worry most clearly when writing about the gravitational constant G, the numerical value of which arbitrarily depends on a choice of units:

> The role [G] plays is as an entity extrinsic to the process to be explained, an entity related to the process to be explained only by a function (a rather arbitrarily chosen function at that). Surely then it would be illuminating if we could show that a purely intrinsic explanation of the process was possible, an explanation that did not invoke functions to extrinsic and causally irrelevant entities. (Field, 1980, 44)

I believe that this is the crucial defect of the external approach: it's appeal to physically irrelevant functions in order to characterise the structure of physical quantities such as mass or spacetime. The external approach is metaphysically opaque because we cannot interpret the x's themselves as physically meaningful, since they only encode facts about the representational equivalence of the theory's models.

Because external sophistication fails the requirement of intrinsicality, it cannot explain the physical equivalence of SRMs. After all, the symmetry-first approach aims to extract physical content from the *assumption* that SRMs are physically equivalent, and hence an explanation of that equivalence in physical terms becomes viciously circular. The failure of intrinsicality likewise means that an extrinsic account is unable to offer causal explanations: only intrinsic structure is causally efficacious, but external sophistication does not specify this intrinsic structure. The relations between mathematical models, which the external approach uses to extract physical content, have no causal effects. The issue that lies at the foundation of external sophistication, then, is that it tries to characterise the physical world in terms of formal relations between representations of the world. This yields only an indirect picture of the world; and it is one whose content is not readily surveyable.

5 Conclusion

For all of the above reasons, only internal sophistication provides a satisfactory interpretation of symmetry-related models. This emphasises the

importance of a perspicuous formalism, in addition to a perspicuous interpretation. Although I agree with motivationalist authors such as Møller-Nielsen (2017) and Martens and Read (2020) that a perspicuous interpretation is important, the discussion above shifts the focus away from interpretation to formalism as the proper demand of motivationalism. After all, it is the latter which is decisive in the choice between internal and external sophistication. On the view I have defended, we are only warranted to interpret SRMs as physically equivalent when we have *both* a perspicuous account of the theory's ontological commitments, *and* a perspicuous formalism from which we can 'read off' these commitments. The latter amounts to an intrinsic theory in the sense of Field (1980).

These points extend beyond the choice between different forms of sophistication. In §2.2, I mentioned that Dewar contrasts sophistication with reduction. The aim of reduction is to construct a theory whose models uniquely correspond to equivalence classes of SRMs of the old theory. The standard way to achieve this is to reformulate the theory solely in terms of invariant quantities, such as distances and relative velocities. But as with sophistication, there is a way to 'brute-force' this requirement: quotient the space of KPMs of the old theory by the relevant symmetry group, such that the models of the reduced theory are *identified with* equivalence classes of SRMs of the old theory. In the physics literature, this procedure is known as 'quotienting' (Belot, 2003). Yet a quotiented formalism leaves it unclear which quantities are fundamental, that is, which quantities one can coherently define over the quotiented space of models. Just like external sophistication, it defines physical structure in terms of the relations between models, rather than vice versa. Therefore, quotienting falls prey to the same arguments used against external sophistication presented above. The refinement of motivationalism I have presented should therefore be of broad interest to interpreters of physical theories.

Nevertheless, unperspicuous methods such as external sophistication and quotienting present us with powerful mathematical apparatus. I will conclude with the suggestion that these approaches may still be useful as an *interim* interpretation of SRMs. An unperspicuous formalism can be used as a vantage point from which to find an intrinsically-defined structure. The strategy then would be to assume that such structures exist and use educated guesses to find their invariant content. Dewar mentions a similar proposal:

> Assuming that one accepts the external method of definition as mathematically legitimate, then its application gives us a way

> of defining a sophisticated semantics for the theory, by brute force. It then means that we do have a precise target for a sophisticated semantics that is internally defined: we are looking for some internal construction which delivers an equivalent class of structures. (Dewar, 2019, 504)

I agree with this view: as a methodological half-way house, extrinsic methods are unobjectionable. However, the interpretation of symmetry-related models is not finished before we have a perspicuous characterisation of their physical content in purely intrinsic terms. To achieve that aim, an intrinsic formalism is required.

References

Arntzenius, F. 2012. *Space, Time, and Stuff*. Oxford, New York: Oxford University Press.

Arntzenius, F. and C. Dorr. 2012. Calculus as Geometry. In *Space, Time, and Stuff*. Oxford University Press.

Baker, D. J. 2010. Symmetry and the Metaphysics of Physics. *Philosophy Compass 5*(12): 1157–1166.

Barrett, T. W. 2017. What Do Symmetries Tell Us About Structure? *Philosophy of Science* (4): 617–639.

Belot, G. 2003. Notes on Symmetries. In *Symmetries in Physics: Philosophical Reflections*, eds. K. A. Brading and E. Castellani, 393–412. Cambridge University Press.

Belot, G. 2013. Symmetry and Equivalence. In *The Oxford Handbook of Philosophy of Physics*, ed. R. Batterman, 318–339. Oxford University Press.

Berger, M. 1987. *Geometry I*. Springer Science & Business Media.

Butterfield, J. 2011. Emergence, Reduction and Supervenience: A Varied Landscape. *Foundations of Physics 41*(6): 920–959.

Caulton, A. 2015. The Role of Symmetry in the Interpretation of Physical Theories. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 52: 153–162.

Chen, E. K. 2018. The Intrinsic Structure of Quantum Mechanics. http://philsci-archive.pitt.edu/15140/.

Dasgupta, S. 2016. Symmetry as an Epistemic Notion (Twice Over). *The British Journal for the Philosophy of Science 67*(3): 837–878.

Debs, T. A. and M. L. G. Redhead. 1996. The twin "paradox" and the conventionality of simultaneity. *American Journal of Physics 64*(4): 384–392.

Dewar, N. 2019. Sophistication about Symmetries. *The British Journal for the Philosophy of Science 70*(2): 485–521.

Dewar, N. 2022. *Structure and Equivalence*. Cambridge University Press.

Earman, J. 1989. *World Enough and Spacetime*. MIT press.

Field, H. H. 1980. *Science without Numbers: A Defence of Nominalism*. Princeton, N.J: Princeton University Press.

Ismael, J. and B. C. van Fraassen. 2003. Symmetry as a Guide to Superfluous Theoretical Structure. In *Symmetries in Physics: Philosophical Reflections*, eds. K. Brading and E. Castellani, 371–92. Cambridge University Press.

Malament, D. 1977. Causal Theories of Time and the Conventionality of Simultaneity. *Noûs 11*(3): 293–300.

Malament, D. B. 2012. *Topics in the Foundations of General Relativity and Newtonian Gravitation Theory*. Chicago: Chicago University Press.

Martens, N. C. M. and J. Read. 2020. Sophistry about symmetries? *Synthese*.

Møller-Nielsen, T. 2017. Invariance, Interpretation, and Motivation. *Philosophy of Science 84*(5): 1253–1264.

Mundy, B. 1986. On the General Theory of Meaningful Representation. *Synthese 67*(3): 391–437.

North, J. 2021. *Physics, Structure, and Reality*. Oxford, New York: Oxford University Press.

Pooley, O. 2013. Substantivalist and Relationalist Approaches to Spacetime. In *The Oxford Handbook of Philosophy of Physics*, ed. R. Batterman. Oxford University Press.

Read, J. 2021. Geometric objects and perspectivalism. http://philsci-archive.pitt.edu/18911/.

Russell, J. S. 2018. Quality and Quantifiers. *Australasian Journal of Philosophy 96*(3): 562–577.

Saunders, S. 2003. Physics and Leibniz's Principles. In *Symmetries in Physics: Philosophical Reflections*, eds. K. Brading and E. Castellani, 289–307. Cambridge University Press.

Sider, T. 2020. *The Tools of Metaphysics and the Metaphysics of Science*. Oxford, New York: Oxford University Press.

Suppes, P. 1967. What Is a Scientific Theory? In *Philosophy of Science Today*, ed. S. Morgenbesser. New York: Basic Books.

Wallace, D. 2019a. Observability, redundancy and modality for dynamical symmetry transformations. http://philsci-archive.pitt.edu/16622/.

Wallace, D. 2019b. Who's Afraid of Coordinate Systems? An Essay on Representation of Spacetime Structure. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 67: 125–136.

Weatherall, J. O. 2017. A Brief Comment on Maxwell(/Newton)[-Huygens] Spacetime. *arXiv:1707.02393 [physics]*.