European Journal for Philosophy of Science (2023) 13:40
https://doi.org/10.1007/s13194-023-00539-4

PAPER IN PHILOSOPHY OF THE NATURAL SCIENCES

Motivationalism vs. interpretationalism about symmetries: some options overlooked in the debate about the relationship between symmetries and physical equivalence

Joanna Luc[1]

Received: 26 July 2022 / Accepted: 5 July 2023 / Published online: 22 August 2023

Abstract

In the recent philosophical debate about the relationship between symmetries and physical equivalence, two approaches have been distinguished: motivationalism and interpretationalism. In this paper, I point out that there are variants of interpretationalism that have not been taken into account by the proponents of motivationalism. I also argue that some of these overlooked variants of interpretationalism are not prone to the motivationalists' criticism and overall are the most attractive positions available.

Keywords Symmetries · Physical equivalence · Motivationalism · Interpretationalism

1 Introduction

Symmetries of a physical theory are often regarded as providing important clues for the interpretation of that theory.[1] Symmetry-related models are often said to be physically equivalent, and symmetry-variant quantities are often said to be unphysical and in this sense redundant (see, e.g., Castellani, 2003; Saunders, 2003; Baker, 2010). This view has been challenged in the recent philosophical literature in at least two ways. First, some authors claim that there are cases in which symmetries are not

[1] It is controversial how exactly symmetries should be defined. The most liberal definition is that any bijection on the set of models of a theory counts as a symmetry (cf. Belot, 2013:3), but it is surely too broad, as then any two models would be related by some symmetry. Here, by "symmetry" I mean "dynamical symmetry", understood as any transformation of the variables of the theory that does not change the form of the dynamical laws of this theory (see, e.g., Brading & Castellani, 2007:1342–1343). In the physics literature, this notion of symmetries seems to be the most popular one (although there are also other important notions of symmetries used in physics, most notably variational symmetries, which form a subset of dynamical symmetries). This list does not exhaust all possible options; for example, some authors consider definitions of symmetries that have built in the empirical equivalence of symmetry-related models (cf. Section 4.1).

✉ Joanna Luc
joanna.luc.poczta@gmail.com

[1] Institute of Philosophy, Jagiellonian University, Grodzka 52, 31-044 Kraków, Poland

associated with the physical equivalence of models/redundancy of quantities; that is, symmetry-related models can be physically inequivalent and/or symmetry-variant quantities can be physically significant (e.g., Belot, 2013, 2018; Fletcher, 2020). I discuss such examples elsewhere (Luc, 2022), so they will not be analysed here. Second, it has been argued that we should think about the link between symmetries and the physical equivalence of models/redundancy of quantities in a motivational rather than in an interpretational way (Møller-Nielsen, 2017; Read & Møller-Nielsen, 2020a, b; Martens & Read, 2020). In short, this means that when faced with a case of a symmetry-variant quantity, we should not automatically regard it as unphysical, but we should be motivated to find a reformulation of the theory that does not involve this quantity[2]; only if we succeed are we allowed to endorse its redundancy. Similarly, when faced with a case of symmetry-related models of a theory, we should not automatically regard these models as physically equivalent, but we should be motivated to find a *perspicuous* account of the shared ontology of these models; only if we succeed are we allowed to endorse the physical equivalence of these models.

What is the relation between these two argumentative strategies? They are both aimed at undermining the thesis that symmetry-related models are physically equivalent, but their positive conclusions are different: if there really are symmetry-related models that are not physically equivalent, then no account according to which they have the same ontology can be adequate, so we should *not* be motivated to find such an account, which means that the motivationalist attitude towards these models is not warranted. Therefore, the success of the first argumentative strategy would undermine motivationalism as a universal stance towards symmetries; however, it also might be used to argue for a more limited form of motivationalism, concerning theories that do not fall in the scope of the first strategy.

The aim of this paper is to show that the division between the motivational and interpretational approaches to symmetries, as it appears in the literature, is not exhaustive or at least not fine-grained enough. There are certain more nuanced forms of interpretationalism that might be attractive for someone who is convinced to some extent by the arguments for motivationalism but does not want to loosen the connection between symmetries and physical equivalence/redundancy as radically as motivationalism does. Moreover, I will argue that taking into account the arguments of both sides of the debate, these nuanced forms of interpretationalism seem to be the most attractive positions available.

This paper is organised as follows. In Section 2, the distinction between the motivational and interpretational approaches to symmetries is reviewed. Next, some arguments for motivationalism and interpretationalism from the literature are presented in Sections 3 and 4, respectively. In Section 5, the observation is made that the subject matters of motivationalism and interpretationalism are not exactly the

[2] This might be any change of the formalism of the theory after which the quantity in question is no longer represented in that formalism. Not any change would count; for example, replacing models by their equivalence classes under symmetries is not enough, because if the quantity in question was needed to define models, it is still needed to define their equivalence classes (which requires defining models first).

same. This is made precise by distinguishing three questions, which concern the preferred initial reaction towards symmetry-related models (should we regard them as physically equivalent or inequivalent?), our further research on the theory (should we search for a perspicuous account of the ontology shared by these models?) and the preferred way of updating of our beliefs about (in)equivalence depending on the outcomes of this research. All these questions are answered by motivationalism, and only the first one is answered by interpretationalism (as such). Therefore, one can distinguish different variants of interpretationalism, depending on how it answers the remaining two questions. In Section 6, I argue that some variants of interpretationalism might be preferable to motivationalism, even for those who endorse to some extent the arguments for motivationalism. Section 7 provides a summary.

2 Motivationalism and interpretationalism about symmetries

Møller-Nielsen characterises the interpretational approach to symmetries as follows (2017:1256):

> Symmetries allow us to interpret theories as being committed solely to the existence of invariant quantities, even in the absence of a metaphysically perspicuous characterization of the reality that is alleged to underlie symmetry-related models.

In contrast, the motivational approach to symmetries asserts that:

> Symmetries only motivate us to find a metaphysically perspicuous characterization of the reality that is alleged to underlie symmetry-related models, but they do not allow us to interpret that theory as being solely committed to the existence of invariant quantities in the absence of any such characterization.

Different definitions, not in terms of the existence of symmetry-variant quantities but in terms of the physical equivalence of symmetry-related models, are given by Read and Møller-Nielsen (2020b:88):

> (...) we endorse a so-called *motivational* approach to symmetry transformations (...), according to which models of a given theory related by a symmetry transformation may not invariably be regarded *ab initio* as being physically equivalent; rather, such transformations at most *motivate us to seek* a clear ('metaphysically perspicuous') conception of physical reality according to which such models may indeed be so regarded. This is situated against the so-called *interpretational* approach, according to which models related by a symmetry transformation may invariably be regarded as physically equivalent.

It has been pointed out by Jacobs (2021a) that the issues of the existence of symmetry-variant quantities and of physical equivalence of symmetry-related models should not be conflated. According to him, one can defend the thesis that symmetry-related models are physically equivalent without being committed to anti-realism with respect to symmetry-variant quantities (although the realism with respect to them must be appropriately qualified to be consistent with the physical equivalence of symmetry-related models).[3] If we take this into account, then the two above formulations of the distinction between motivationalism and interpretationalism should be regarded as substantially different. In what follows, I will use the formulation in terms of the physical equivalence of models (but my classification of interpretationalist positions developed in Section 5 might be applied analogously to the case of symmetry-variant quantities—see footnote 20).

The above quotes give the impression that motivationalism is a more modest position than interpretationalism. Interpretationalism is presented as a bold claim (that we should interpret symmetry-related models as physically equivalent), whereas motivationalism is presented as merely a suggestion about how to investigate the theory further once we realise that it has symmetry-related models. However, there is no asymmetry in boldness between these two claims, as the motivational view also prescribes what we should do "in the meantime", that is, while we are looking for a perspicuous account of the shared ontology of symmetry-related models. Namely, it requires us to interpret symmetry-related models as representing different possible worlds, which is said explicitly by Read and Møller-Nielsen (2020a:266), who write that according to the motivational approach:

> (...) the existence of symmetry-related models first (a) motivates us to provide an explication of the shared ontology of these models; but only once such an explication is forthcoming should we (b) interpret those models as representing the same possible world; and (potentially) (c) identify those models to construct a reduced space of KPMs [i.e., kinematically possible models].

This is also stressed by Møller-Nielsen (2017:1261, emphasis mine):

> According to the motivational view, then (to repeat slightly), absent a metaphysically perspicuous characterization of the reality underlying these symmetry-related models, *we have no choice but to regard them as representing physically distinct states of affairs*.

Therefore, motivationalism is not only a claim about what we should seek once we are faced with symmetry-related models and symmetry-variant quantities (namely, "a metaphysically perspicuous characterization of the reality" underlying these models or "an explication of the shared ontology of these models") but also

[3] Jacobs (2021a) construes his position in an analogy to sophisticated substantivalism, which is the view that spacetime points are real, but symmetry-related models are physically equivalent, which is made consistent by assuming anti-haecceitism regarding spacetime points (i.e., the thesis that "[r]ather than possessing primitive identities, spacetime points are individuated via their qualitative relations to each other and to the universe's matter content"; 2021a:9). Jacobs considers a similar move with respect to physical quantities by denying what he calls the "Value-Magnitude Link", which states that "[t]he values of a quantity invariably represent the same magnitude across models" (2021a:6). This is made consistent by an appeal to anti-quidditism (an analogue of anti-haecceitism for quantities), which states that "physical properties are individuated via their position in a structure of qualitative relations" (2021a:10). As a consequence, if distinct values of a given quantity "occupy the same structural role across symmetry-related models, they also represent the same magnitude" (2021a:10); the quantity itself is regarded as real.

a claim about what we should think about these models and quantities in the meantime. Namely, according to motivationalism, we should regard symmetry-related models as physically *inequivalent*, and we should be *realists* with respect to symmetry-variant quantities.

In this exposition, one final clarification is needed, namely, what is meant by "a metaphysically perspicuous characterization of the reality" underlying symmetry-related models or "an explication of the shared ontology of these models".[4]

Møller-Nielsen recognises two ways of giving a perspicuous account of the ontology shared by symmetry-related models. According to him, if symmetry-related models are non-isomorphic,[5] then what is needed is a mathematical reformulation of the theory that does not involve symmetry-variant quantities and such that the whole family of symmetry-related models of the original formulation corresponds in the reformulated theory to a single model. An example is electrodynamics formulated in terms of the electromagnetic potential (which is not invariant under the internal symmetry $A_\mu \mapsto A'_\mu = A_\mu + \partial\phi$, where ϕ is a smooth scalar function). The assumption here is that A_μ is part of the structure of these models, so that changing its value amounts to changing the structure of these models. The perspicuous picture of the reality underlying these models is given here by the reformulation of electrodynamics in terms of the electromagnetic field $F_{\mu\nu}$, which gets rid of the symmetry-variant potential. Models of the original formulation related by an internal symmetry (with different values of the potential) correspond to one model in the reformulated theory (with one value of the electromagnetic field). Another example is Newtonian gravity theory (NGT), whose models are related by various symmetries of the so-called Maxwell's group. By moving to Newton–Cartan theory (NCT), we can remove the redundancy due to some of these symmetries in the sense that a single model of NCT corresponds to the whole equivalence class of symmetry-related models of

[4] Various equivalent expressions are used by the supporters of motivationalism. For example, they also talk about a coherent explication (or picture or understanding) of the common ontology of symmetry-related models, a coherent account of their shared ontology, a coherent explication of the ontology underpinning their physical equivalence, etc. Clearly, by "coherent" something more than logical consistency is meant (cf. footnote 23); I will prefer the term "perspicuous".

[5] What does it mean that two models are isomorphic or non-isomorphic? In general, an isomorphism of a given mathematical object is any transformation that does not change the structure of this object. If that structure is fully and explicitly spelled out, then determining what the isomorphisms of a given object are is a matter of checking for each piece of its structure whether it is changed or not. More formally, in the case of theories formulated in the first-order logic, our models have the form $M_1 = \langle X; R_1, \dots, R_n \rangle$ and $M_2 = \langle Y; R'_1, \dots, R'_n \rangle$, where X and Y are sets, R_i are relations on X and R'_i are relations on Y, then a map $\phi : X \to Y$ is an isomorphism between M_1 and M_2 iff (*i*) ϕ and ϕ^{-1} are both bijections, (ii) for any i, R_i *and* R'_i have the same adicity and (iii) for any $i, R_i(x_1, \dots, x_k) \Leftrightarrow R'_i(\phi(x_1), \dots, \phi(x_n))$, where $x_1, \dots, x_k$ is any k-tuple of elements of X. However, the structure of models is often not specified in all the details, which might lead to controversies concerning which transformations are their isomorphisms. Another complication is that in category theory the notion of an isomorphism is used in a different way: isomorphisms are not determined on the basis of a previously given structure, but rather certain maps between models are *postulated* to be isomorphisms, and whatever is invariant with respect to them counts as the structure of these models. It should also be noted that two models might be related by a dynamical symmetry without being isomorphic; this is because dynamical symmetries are defined in terms of the (form of the) dynamical laws, not in terms of the structure of models.

NGT.[6] In this sense, NCT provides an account of the ontology shared by symmetry-related models of NGT.

If symmetry-related models are isomorphic, then, Møller-Nielsen claims, the mathematical reformulation of the theory is not needed and a merely conceptual re-interpretation of the existing formalism suffices. He gives two examples: spatial translations in NGT and diffeomorphisms in general relativity (GR). The perspicuous picture here is given by an anti-haecceistic interpretation of spacetime points, which is called "sophisticated substantivalism". Translation-related models of NGT and diffeomorphism-related models of GR do not differ qualitatively; they differ only in which particular points in spacetime instantiate which qualities (i.e., they differ merely haecceistically). However, according to sophisticated substantivalism, the only real differences are qualitative differences, so there is no such thing as merely haecceistic differences. This position allows one to reconcile the existence of spacetime points with the physical equivalence of (isomorphic) symmetry-related models by saying that spacetime points are individuated qualitatively. Møller-Nielsen maintains that this suffices to explain why translation-related models of NGT and diffeomorphism-related models of GR are physically equivalent.

To sum up, according to Møller-Nielsen, there are two conceivable ways of providing "a metaphysically perspicuous characterization of the reality" underlying symmetry-related models or "an explication of the shared ontology of these models": if these models are non-isomorphic, one should appropriately reformulate the theory, whereas if these models are isomorphic, one should appeal to an analysis in the spirit of sophisticated substantivalism. It needs to be stressed that the particular form of motivationalism defended by Møller-Nielsen (as well as Read and Martens) should not be confused with a motivationalist thesis in general. I take Møller-Nielsen's two-element list to be not defining for a motivationalist position, so that one can be a motivationalist while acknowledging some other ways of providing a perspicuous account of the shared ontology of symmetry-related models not mentioned by him or while not acknowledging some of his methods (in particular, a motivationalist might endorse the mathematical reformulation of a theory as the only way to specify the common ontology of its symmetry-related models and disregard sophisticated substantivalism).

This list and the exact understanding of its items are indeed controversial. For example, in the case of GR, the two main ways of explaining the physical equivalence of diffeomorphism-related models considered in the literature are sophisticated substantivalism (favoured by Møller-Nielsen) and the reformulation of the theory in terms of Einstein algebras, both of which have their supporters and critics. The former is endorsed by, among others, Brighouse (1994), Hoefer (1996) and Pooley (2006, 2013), but Dasgupta (2011) and Gomes (2022a, b) argue that it needs to meet additional constraints to be acceptable. According to Dasgupta (2011:131), sophisticated substantivalists should not come to a stop after saying that "individualistic facts about the manifold are grounded qualitatively", but in addition they "must

[6] NCT still has some symmetry-related models, but they are isomorphic. Therefore, they fall under the second case (i.e., the case of isomorphic models) discussed in the next paragraph.

(1) clearly articulate what the underlying qualitative facts are like, and (2) show that they are sufficient to explain (in the metaphysical sense) individualistic facts about the manifold", which, according to him, so far has not been done. Gomes (2022b:6) postulates three desiderata for sophistication (the symmetries should be induced by the automorphisms of some "natural" geometric structure, this geometric structure should be axiomatisable in terms of the basic physical predicates assumed for theory and the automorphisms of the structure should correspond to changes between different choices of physical coordinate systems or physical reference frames) and claims that they are satisfied in the case of GR, so in the end he endorses a more restrictive variant of sophistication. The option based on Einstein algebras is developed by Earman (1989, ch. 9), according to whom they provide "a direct characterization of physical reality" underlying the diffeomorphism–equivalence class of models (Earman, 1989:192) and allow us to get rid of spacetime points. However, his proposal is challenged by Rynasiewicz (1992) and by Rosenstock et al. (2015), who claim that this reformulation leads to a theory that is equivalent to the original one, which undermines the idea that it has less ontological commitments.

Even the well-established formulation of electromagnetism in terms of $F_{\mu\nu}$ might be regarded as not sufficiently perspicuous. Dewar (2019:497–498) suggests that it involves an explanatory loss with respect to the formulation in terms of A_μ. Jacobs (2021b:174) expresses this idea in terms of "cosmic coincidences":

> In this case, the coincidence is the Gauss-Faraday law $\partial_{[\mu}F_{\nu\rho]}$. When expressed in terms of A_μ, this law reduces to a mathematical theorem. But if A_μ is merely a mathematical abstraction, then Gauss' law is an additional postulate of the theory. In that case, it is mysterious that $F_{\mu\nu}$ behaves *as if* it is the exterior derivative of a four-potential, even when it is in fact a fundamental quantity.

As a final note for this section, let me make one important restriction. In this paper, I understand both interpretationalism and motivationalism as claims about (the preferred attitudes towards) models taken to be representing *entire possible worlds*. This is because models taken to be representing subsystems sometimes are not physically equivalent, even in very familiar cases. For example, two translation-related models of *n*-particle classical mechanics might represent two different physically possible situations if the translation here is interpreted actively and a reference frame is associated with a physical object (e.g., a laboratory in which experiments are performed). Then, one model represents *n* particles as located differently with respect to this reference object than the other model.[7] Therefore, in this case, *both* interpretationalism and motivationalism give the wrong verdict: we should neither treat these models (understood as subsystem models) as physically equivalent, and nor should we be motivated to look for their shared ontology underpinning their physical equivalence, as we know that in this case there is no physical equivalence to be underpinned.

[7] In such a case, these *n* particles might be said to be explicitly represented in the model and the reference object to be represented implicitly; see Caulton (2015:158, footnote 17), Pooley (2017:137) and Luc (2022:5–6).

3 Some arguments for motivationalism

What are the arguments supporting motivationalism? In this section, I will try to extract them from the writings of the proponents of this view.[8] Here I will analyse four arguments that I was able to identify: the argument based on the notion of the natural understanding of models (M1), the argument from explanatory loss (M2), the argument from the interpretationalist not being able to take a realistic attitude towards the theory (M3), and the argument appealing to there being no guarantee for the existence of a perspicuous account of the shared ontology of symmetry-related models before it is actually found (M4).

3.1 Argument (M1): natural understanding of models

The first argument, (M1), is framed in terms of the natural understanding of a given family of symmetry-related models. Interpretationalists are said to treat as physically equivalent certain models that are naturally understood as physically inequivalent. The argument is illustrated by means of the following example. If we regard symmetry-related models of NGT as physically equivalent, then we must agree that there are no physical differences between the following three kinds of situations: a system that is force-free and stationary with respect to absolute space, a system that is force-free and moving at a constant absolute velocity and a system that is absolutely accelerating under a gravitational force field. According to Møller-Nielsen (2017:1261), interpretationalism gets this case wrong, as these "are naturally understood as representing radically distinct physical situations".

If naturalness here is understood as complying with our everyday physical intuitions, then this argument is dubious because our everyday physical intuitions have proven to be wrong when faced with the advancements of theoretical physics (and they are unlikely to give any verdict in such cases anyway). However, this is not what the author seems to have in mind, as in the end these situations are judged to be naturally understood as physically inequivalent not because we have such intuitions but because they are not isomorphic (Møller-Nielsen, 2017:1261):

[8] In Sections 3 and 4, I often do not distinguish between the proper arguments for a given position on the one hand, and the responses to the criticisms of that position or criticisms of defenses of an opposite position on the other hand. The reason is that I believe that this division can be made in different ways, partially because the initial arguments for a given position often anticipate potential responses. For example, it has been suggested to me that in Section 3, the proper arguments are (M1)–(M3), to which an interpretationalist might reply by (i) saying that we already have a perspicuous picture (cf. (I2)), (ii) declaring instrumentalism or (iii) arguing that even though we currently do not have a perspicuous picture, we are likely to find it (cf. (I3) and footnote 33); then, (M4) can be understood as a response to (iii). However, I prefer to regard (M3) and (M4) as two horns of a dilemma that is posed to an interpretationalist (and as such is a proper argument for motivationalism, which does not need to wait for any interpretationalist's reaction to be properly posed). Similarly, it has been suggested to me that in Section 4, the proper arguments are (I1) and (I4), whereas (I2) and (I3) are responses to the motivationalist's challenge (corresponding to (i) and (iii) above). I agree that the dialectics here might be seen in this way. However, I think that an appropriate response to (M3)–(M4) should be more complex, in order to address various aspects of this dilemma (e.g., whether an interpretationalist is indeed less cautious). The same concerns (M2), where the issue of the proper order of answering the questions about a theory seems to me crucial (and it is not a part of any of (i)-(iii)). See Sections 6.1 and 6.2, respectively.

> For our purposes, the crucial thing to note about all of these models is that none of them are isomorphic—naturally understood, they do not represent at most haecceitistically distinct possible worlds. According to the criterion laid down in the previous section, then, in order to be able to transparently understand how it could be that such models may be said to represent physically equivalent scenarios, a mathematical reformulation of the theory is required.

Therefore, the concept of the "natural understanding" of models is partially characterised here by the condition that only isomorphic models are naturally understood as representing at most haecceitistically distinct possible worlds (i.e., possible worlds that are not qualitatively different). From this it follows that the three types of NGT models listed above are naturally understood as qualitatively different, as they are non-isomorphic (the first differs from the second by the value of absolute velocity, whereas the third differs from the other two by its involving gravitational force). This approach to the notion of "natural understanding" is more promising than the one appealing to our pre-theoretical intuitions, but as it is used here, it seems to be only a re-statement of motivationalism (in its particular version advocated by Møller-Nielsen) rather than an argument in favour of it. In the presented quote, the author explicitly appeals to his criterion "laid down in the previous section", but this criterion is a part of the position defended by him. Perhaps Møller-Nielsen did not even mean this to be a proper argument for motivationalism but only an illustration of how a motivationalist might proceed.

It has been suggested to me that one could interpret the notion of "natural understanding" in terms of Quine's (1951) notion of ontological commitment: whatever the theory quantifies over does belong to this theory's ontological commitments. Then, one could argue that because NGT quantifies over, for example, the frame of absolute rest, by endorsing NGT we should thereby endorse the existence of the frame of absolute rest. However, this Quinean approach relies on the assumption that one can simply read off the theory's ontological commitments from its formalism, which is not in line with Møller-Nielsen's (2017) way of thinking, as he explicitly claims that one can change the ontological interpretation of the theory without changing its formalism (and in some cases he regards this option as preferable—namely, if the models are isomorphic). Another problem with applying Quinean strategy here is that it is not clear what NGT quantifies over—the mere fact that something belongs to its formalism does not mean that it is the scope of its quantifiers. Settling this issue would require the precise formulation of NGT in first-order logic, which has not been provided. Finally, showing that some entities belong to the theory's ontology does not straightforwardly lead to the conclusion that models differing on these entities are physically inequivalent; for example, Jacobs (2021a) endorses the view that symmetry-variant quantities exist even though models differing merely by their values are physically equivalent.

Summing up, I do not find much potential in this argument and in the notion of "natural understanding" of models in general, unless it is explicated in different terms (in which case this would become a different argument). The other three arguments seem to be more important (also because they appear in other papers as well, in contrast to the first one).

3.2 Argument (M2): explanatory loss

The second argument, (M2), appeals to the loss of explanatory transparency by an interpretationalist who regards as physically equivalent symmetry-related models despite lacking a perspicuous account of their shared ontology. For such a person (Møller-Nielsen, 2017:1263),

> (...) the reality in terms of which this physical equivalence is to be understood will (absent a reformulation of the theory) remain opaque to her; she is offered no immediate explanation as to how such physical equivalence is to be construed or how it could even be said to arise.

Similarly, Read and Møller-Nielsen (2020a:276) claim that.

> (...) without further work, the advocate of the interpretational approach offers no explanation as to how such physical equivalence is to be construed, or how it could even be said to arise.

The main weakness of this argument, which Møller-Nielsen himself recognises, is that it relies on an unclear notion of explanatory transparency. However, a partial characterisation of the loss of explanatory transparency is given; namely, it is claimed to hold whenever one treats as unreal a piece of the formalism of the theory that has an explanatory role within this formalism. For example, in NGT, facts about absolute velocities explain facts about other quantities in the theory (e.g., relative velocities), so a commitment to their non-existence leads to an explanatory loss (cf. Møller-Nielsen, 2017:1263). Another example (perhaps the most important one, as the above quotations suggest) of a fact that cannot be explained by an interpretationalist is the physical equivalence of symmetry-related models itself.

3.3 Argument (M3): realistic attitude towards a theory

The third argument, (M3), is that an interpretationalist who does not have at his disposal an account of the shared ontology of symmetry-related models of the theory is not able to take a realistic attitude towards this theory. According to Møller-Nielsen (2017:1263–1264), an interpretationalist is then forced to think about this theory in an instrumentalist way. This is because to think about a theory realistically, one needs to be able to specify what, according to this theory, the world is really like (Read & Møller-Nielsen, 2020a:276).[9]

3.4 Argument (M4): no guarantee for the existence of a perspicuous account of the shared ontology of symmetry-related models

The fourth argument, (M4), is based on the observation that from the fact that certain models of a theory T are symmetry-related it does not follow that we can find a

[9] In the paper by Read and Møller-Nielsen (2020a), what I call (M2) and (M3) are regarded as one argument. However, it is worth distinguishing them because they use different concepts: (M2) is about explanatory transparency, whereas (M3) is about being (un)able to take a realistic attitude towards a theory.

perspicuous account of their common ontology. There is simply no guarantee that such an account exists in the sense of being logically possible, "waiting in logical space to be discovered" (Møller-Nielsen, 2017:1262). Relatedly, "there appears to exist no set of a priori principles by which one may deductively infer" that there is such an account (Read & Møller-Nielsen, 2020a:284).[10] An interpretationalist here is accused of not being cautious enough in his assertion that a perspicuous account of the shared ontology of symmetry-related models of T can be constructed before being able to explicitly provide an example of such an account. It seems more reasonable, a motivationalist says, to first find an account of this kind and only then make a claim that it exists. Indeed, motivationalism is said by its supporters to be "the most epistemically cautious interpretative attitude possible towards models of physical theories related by symmetries" (Read & Møller-Nielsen, 2020a:286).[11] They also claim that the burden of justification should always lie on riskier positions—in this case on interpretationalism.

I will come back to these arguments in Section 6, where I will discuss their strength and, assuming their viability, the extent to which they threaten various variants of interpretationalism. Here, I will make only one remark about the relation between (M3) and (M4).

Arguments (M3) and (M4) seem to apply to two different construals of interpretationalism. Under the first construal (in (M3)), an interpretationalist is forced to be an instrumentalist (even if he originally intended otherwise) because the only differences that he acknowledges as real are observational differences (which is what enables him to regard two observationally equivalent models as physically equivalent). The objection to interpretationalism understood in this way presumes that scientific realism is in general a more appropriate stance towards scientific theories than instrumentalism, so by showing that interpretationalism is an instrumentalist view we perform a kind of *reductio* of it. Under the second construal (in (M4)), an interpretationalist manages to be a realist with respect to a physical theory and *does* take into account aspects of the theory other than its observational consequences. However, an interpretationalist is understood here as overconfident in his tendency of making assertions about possible reformulations of the theory before such reformulations are worked out. Here, the objection is that one can be more reasonable by being more cautious and treating as physically equivalent only those symmetry-related models whose shared ontology is known.

Are these two arguments incompatible, given that they make different assumptions about what an interpretationalist position amounts to? I do not think so; rather, they might be read as working together by providing a dilemma for an interpretationalist, who needs to choose between two unfavourable options: either to be an instrumentalist or to be an incautious realist.

[10] Another problem is that even if such an account exists, we might not be capable of discovering it (i.e., it might be too complicated for any human being to understand).

[11] In the paper by Read and Møller-Nielsen (2020a), two versions of motivationalism are distinguished, namely, confident (assuming that the shared ontology of symmetry-related models is guaranteed to exist) and cautious (assuming that the shared ontology of symmetry-related models is not guaranteed to exist). In my paper, I identify motivationalism with its cautious version, which is the one defended by the authors.

4 Some arguments for interpretationalism

In this section, I will review some arguments for interpretationalism, without aiming for completeness in this respect. I will present four arguments: the argument from the Occam's razor/from undetectability (I1), the argument from the universal availability of a method for changing non-isomorphic models into isomorphic ones (I2), the inductive argument (I3), and (for local symmetries only) the Hole Argument together with its generalisations (I4).[12]

4.1 Argument (I1): the Occam's razor and undetectability

The first and perhaps the most popular argument for interpretationalism, (I1), appeals to a certain version of the Occam's razor, namely, "other things being equal, our preferred scientific theories should not allow for solutions that represent physically distinct but nevertheless empirically indistinguishable possible worlds" (Møller-Nielsen, 2017:1261). In this version of the Occam's razor, it is the physical differences between models that should be not multiplied beyond necessity.[13] It might also be called the argument from undetectability.

The non-obvious move in this argument is from the models being symmetry-related to their being empirically equivalent. This implication can be established in at least three ways (and a lot depends here on how exactly symmetries are defined).[14] First, this might be a case-by-case analysis: we take transformations that are regarded in the physics literature as symmetries of given theories and realise that for each of them, models related by them are empirically equivalent. This is the weakest strategy (because we establish the claim only for a given set of theories and their symmetries, not in general) but perhaps also the least controversial. Second, one might simply define symmetries in terms of empirical equivalence or some related notion (see, e.g., Ismael & van Fraassen, 2003; Dasgupta, 2016:866–871;

[12] Some other arguments that are not mentioned in the main text are the argument from redundancy (see, e.g., Dasgupta, 2011:137-141) and the argument from objectivity (see, e.g., Debs & Redhead, 2007:52–75).

[13] This is a different criterion of ontological parsimony than the comparison of "the amount" of entities and/or structures postulated by different theories. These two criteria might not be straightforwardly related. For example, we can eliminate certain differences between models by constructing a reduced theory (i.e., a theory expressed solely in terms of symmetry-invariant quantities), but it is not obvious that its ontology must be a subset of the ontology of the original theory. However, it is impossible that these two criteria come apart in the sense that T_1 postulates less differences than T_2 even though the ontology of T_2 is a proper subset of the ontology of T_1. This is because the addition of any entity or structure to the ontology leads to the addition of new possible physical differences (namely, with regard to this entity or structure) without removing any of the previously recognised differences.

[14] Further complications might arise if we accept Maudlin's (1993) claim that qualitative indistinguishability does not need to lead to empirical indistinguishability because "we have more than purely qualitative vocabulary to describe the actual world" (Maudlin, 1993:191). This non-qualitative vocabulary includes indexicals and demonstratives. With these, we can formulate sentences that are true in the actual world but false in all possible worlds that are related to it by a static shift, and their truth values are knowable to us. This means that if our concept of empirical distinguishability allows the use of such non-qualitative sentences, symmetry-related worlds sometimes are empirically distinguishable, even though they are qualitatively indistinguishable.

Read & Møller-Nielsen, 2020a:267): a transformation of a theory T is a symmetry iff it always transforms models of T into empirically indistinguishable models of T (and perhaps satisfies some further conditions). Third, one can define symmetries without explicit reference to empirical equivalence (or any related notion), for example in terms of leaving invariant the dynamics of the theory (as in my footnote 1), and argue that symmetries defined in this way relate observationally equivalent models (see, e.g., Roberts, 2008; Wallace, 2022; Dewar, 2022, section 6.2). The second strategy seems to resolve the issue of the relation between symmetries and empirical equivalence too easily (do we really want their link to be an analytic truth?). The third strategy does not have this disadvantage and seems to be closer to how symmetries are defined by physicists, but it is also the most difficult to work out completely. As declared at the beginning, in this paper I understand symmetries as dynamical symmetries, so my preferences are on the side of the third strategy.

Møller-Nielsen dismisses the argument from the Occam's razor on the grounds that it is perfectly reasonable that certain aspects of reality are empirically inaccessible. According to him, "prohibitively strong versions of verificationism aside, there is nothing obviously absurd about admitting in-principle undetectable facts into one's ontology" (2017:1262). I will come back to this response in Section 6.1.

4.2 Argument (I2): a universal method for changing non-isomorphic models into isomorphic ones

The second argument for interpretationalism, (I2), attempts to undermine the argument (M4) for motivationalism. Even though it is not easy to find a reformulation of a theory with non-isomorphic symmetry-related models that gets rid of all the structure responsible for this non-isomorphism, one can always find—the proposal goes—a reformulation in which these models become isomorphic and then declare physical equivalence of these models in the spirit of sophisticated substantivalism, which is one of the options allowed by Møller-Nielsen's (2017) motivationalism. In this way, the sought-for perspicuous account of the shared ontology of symmetry-related models might always be obtained.

Dewar (2019) seems to embark on this kind of project (although locating his views within the framework of Section 5 is a subtle matter—see the second paragraph of Section 6.4). He calls this approach "sophistication", to be contrasted with "reduction", which amounts to formulating "a theory that deals only in quantities that are invariant under the relevant symmetry" (Dewar, 2019:486). In more detail, his method of sophistication is external in the following sense (Dewar, 2019:502–503):

> 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). (...) Hence, the proposal is that the pictures on the new semantics are simply what we obtain by taking the old objects, and declaring, by fiat, that the symmetry transformations are now going to 'count' as isomorphisms.

For this proposal to work, we need to understand the structure of models as determined by their isomorphisms, not the other way around (cf. footnote 5).[15] Wallace (2022:336) also considers an approach of this kind and points out that the category-theoretic notion of isomorphism is very flexible, so if we decide to understand isomorphism in this way, changing a family of non-isomorphic models to a family of isomorphic models would be relatively easy.

However, both of these approaches might seem to be rather artificial, as they use a very weak notion of isomorphism. Usually, we seem to have something more substantial in mind when talking about isomorphisms. Martens and Read (2020:13) criticise this approach by claiming that "external sophistication in itself does not afford a perspicuous explication of the ontology of symmetry-related models". The isomorphism of models is not something that can just be declared; it should require substantial work to ensure that it really holds.[16] I essentially agree with this criticism of (I2), so I will not rely on it in the further discussion of interpretationalism (but see Section 6.4 for some more remarks on (I2)).

4.3 The inductive argument (I3)

The third argument, (I3), might be called an inductive argument, although the inductive basis here is rather atypical because it does not consist of empirical observations but rather of cases of our well-established and relatively well-understood physical theories. The premise of the argument is that in all such cases, we (i.e., the community of researchers) succeeded in finding a perspicuous account of the shared ontology of symmetry-related models—either they were isomorphic from the start (as Dewar, 2019:504 says, "modern theories are typically born sophisticated") or we managed to find appropriate reformulations of our theories that get rid of symmetry-variant structure. Therefore, we have inductive grounds to believe that this might be done for new theories with symmetries. Of course, this argument has all the difficulties usually associated with induction and relies on the strong assumption that symmetries in all physical theories have a similar nature; but these problems do not undermine its value.

Wallace (2022:336) is among those who endorse the premise of the inductive argument (although he does not formulate this argument in the way I have suggested above):

[15] This is why Jacobs (2021b:91) calls Dewar's approach "symmetry-first sophistication", which is contrasted with "structure-first sophistication".

[16] Jacobs (2022) has argued that external sophistication leads to perspicuous interpretation (i.e., an account of the models' common ontology) but not to perspicuous formalism—that is, a formalism "whose mathematical structures 'intrinsically' represent the physical world" (2022:1) so that "one can 'read off' the theory's metaphysics from the formalism" (2022:7). According to him, "external sophistication has an effective decision procedure for ontological commitment" (2022:10): it is committed to any structures that are implicitly defined as those structures of the original models that are invariant under symmetries. However, I take it that a perspicuous account of the shared ontology of symmetry-related models, as understood by motivationalists, was supposed to include both perspicuous interpretation and perspicuous formalism in Jacobs's sense: they require not just *some* account of the shared ontology of symmetry-related models but one that is "clear" and "metaphysically perspicuous" (and for this aim implicit definitions seem to be insufficient).

> In any case, the concern [that we are not guaranteed to always find an appropriate reformulation of a theory with non-isomorphic symmetry-related models] is fairly theoretical: I am not aware of any theory in extant physics (even construing 'extant' fairly broadly) which does not have a well-understood reformulation in which dynamical symmetries and automorphisms coincide, even if one eschews Kleinian and categorical tricks and insists on a more purist conception of reformulation.

Remarkably, all examples used by Møller-Nielsen (2017) confirm this claim. He appeals to two theories with non-isomorphic symmetry-related models, namely NGT and electrodynamics, but in both cases the appropriate reformulations are known.[17,18]

4.4 The hole argument and its generalisations (I4)

Finally, let me mention a less general but very important argument, that is, the Hole Argument (I4), according to which unless we regard diffeomorphism-related models as representing the same possible world, our theories would suffer a radical form of indeterminism (Earman & Norton, 1987). As stated, this argument applies only to diffeomorphisms in generally covariant theories (such as GR), but it can be extended to other local symmetries (see, e.g., Healey, 2001). It should be mentioned that the Hole Argument was conceived by Earman and Norton as an argument against spacetime substantivalism and as such was later questioned by the advocates of sophisticated substantivalism (see, e.g., Brighouse, 1994; Hoefer, 1996; Pooley, 2006 and 2013), but most of the participants in the debate would agree that it at least establishes that diffeomorphism-related models represent the same possible world.

Another point to note, which is not always emphasised, is that the radical indeterminism involved here concerns properties that are not detectable, so this argument should be seen as closely associated with (I1). If we were able to empirically distinguish between models related by local symmetries, then the mentioned radical indeterminism would just be an empirical hypothesis of our theories and not their methodological vice.

5 Four variants of interpretationalism

If we look closely at how motivationalism and interpretationalism are defined by Møller-Nielsen (2017), we can see that the subject matters of these theses are not exactly the same. Interpretationalism only says how we should interpret a theory

[17] There is an interesting historical difference between these cases; namely, electrodynamics started with symmetry-invariant formulation in terms of electromagnetic fields and symmetry-variant formulation in terms of potentials was introduced later, whereas NGT started as a theory with non-isomorphic symmetry-related models and the formulation getting rid of absolute velocities was found much later.

[18] Argument (I1) also might be supported in an inductive way. Even if one is not convinced by the general arguments that for any physical theory, any difference between its symmetry-related models is empirically undetectable, it is at least clear that no such difference has ever been empirically detected for any well-established physical theory.

with symmetry-related models, whereas motivationalism in addition says how we should pursue our further research of this theory (namely, that we should, among other things, seek a perspicuous account of the ontology common to its symmetry-related models).[19] In this section, I will make this observation more precise, which will allow me to distinguish four different variants of interpretationalism.

Consider a theory T and assume that we have just discovered that some of its models are symmetry-related. What should be our interpretation of T, given this discovery? It can be described by providing answers to the following three questions[20]:

1. What should our *initial* reaction towards symmetry-related models of T be—should we regard them as physically equivalent or as physically inequivalent?
2. Should we look for a perspicuous account of the ontology shared by symmetry-related models of T?
3. How should we update our interpretation of symmetry-related models of T depending on the outcomes of the research mentioned in question (2)?

Question (1) concerns what the most appropriate *initial* reaction to the discovery that certain models of T are symmetry-related is before we undertake any further analyses of T and irrespective of how much we know about T. Question (2) concerns the recommended ways of undertaking further research on T: should they include searching for a perspicuous account of the ontology shared by symmetry-related models of T or not necessarily?[21] Question (3) presupposes that the answer to question (2) is different than an unqualified "no", so it should be considered only in these cases. It concerns the stage of the investigation of T that comes after putting some significant effort into the research that question (2) was about and asks how our initial interpretation should be changed in response to the results of this research.

[19] It seems that there is another difference between interpretationalism and motivationalism, namely that the former is directly about metaphysics ("symmetry-related models are physically equivalent"), whereas the latter is about our rational epistemic stances towards metaphysical theses ("we should regard symmetry-related models as physically inequivalent until a perspicuous account of their shared ontology is found"). To make them comparable, interpretationalism needs to be given a similar epistemic formulation, which is done in the main text. However, I do not find this change problematic because the transition from the metaphysical to the epistemic formulation (and back) is rather straightforward: one should endorse those metaphysical theses that one regards as the most rational.

[20] One can also formulate a version of our three questions for the debate about symmetry-variant quantities. Consider a quantity V that is postulated by a theory T. Assume that we have just discovered that V is variant under some symmetry S of T. Our interpretative stance towards V can be described by providing answers to the following three questions:

1. What should our *initial* reaction towards V be—should we be realists or anti-realists towards V?
2. Should we look for a reformulation of T that gets rid of V?
3. How should we update our interpretation of V depending on the outcomes of the research mentioned in question (2)?

As I declared in Section 1, I will consider in the main text only the debate about physical (in)equivalence of symmetry-related models.

[21] We assume that our aims with regard to T are purely epistemic, so answers like "no, we should prioritise the search for practical applications of the theory" are irrelevant here, even if reasonable in certain contexts.

As we see, questions (1), (2) and (3) concern different stages of the analysis of the theory, which might be temporally separated. They do not need to be because it might happen that the discovery that certain models of T are symmetry-related is made at the time when we already have at our disposal the account of their shared ontology (which is the historical case of electrodynamics). Therefore, the order of the questions is logical and only sometimes temporal as well.[22]

Motivationalism gives us determinate answers to each of the above three questions:

1. We should initially interpret symmetry-related models of T as physically inequivalent.[23]
2. Yes.
3. If we find such an account, we should change our initial interpretation and begin to regard symetry-related models of T as physically equivalent; if we do not find such an account, we should retain our initial interpretation.

In contrast, interpretationalism can be understood as defined only by the answer to the first of these questions, namely, as a view according to which in a situation of finding that certain models of T are symmetry-related, our first reaction should be to interpret these models as physically equivalent.[24] Therefore, being an interpretationalist is consistent with giving different answers to the remaining two questions, which leads us to different variants of interpretationalism. I will consider four such variants, which I will

[22] What is the point of distinguishing the initial and final attitude towards symmetry-related models if we have at our disposal a perspicuous account of their shared ontology from the very beginning? In such a case, we can still (i) ask counterfactually: "If we did not have such an account, should we then also regard these symmetry-related models as physically equivalent?"; (ii) ask about reasons for our stance: "Do we currently regard these symmetry-related models as physically equivalent *only because* we have a perspicuous account of their shared ontology or *partially independently* of this fact?" (of course the independence here might be at most partial because the existence of such an account is always an important argument for physical equivalence—the question is whether it is the only important argument). The answer "no" to (i) and "only because" to (ii) corresponds to motivationalism (i.e., the answer "physically inequivalent" to question (1)); the answer "yes" to (i) and "partially independently" to (ii) corresponds to interpretationalism (i.e., the answer "physically equivalent" to question (1)).

[23] It has been suggested to me that I mischaracterise motivationalism here because it does not only claim that one *should not* interpret symmetry-related models as physically equivalent without having a perspicuous account of their shared ontology, but it makes a stronger claim that *it is not even possible* to interpret symmetry-related models as physically equivalent in such a case. In other words, the disagreement between interpretationalism and motivationalism might concern not just what option one should choose initially but even what the options are that one can choose from. In response, one might ask: What kind of impossibility would be involved here? There is no contradiction in answering "physically equivalent" to question (1), so a weaker kind of impossibility must be meant. However, it is not clear what the difference is between some interpretation being impossible in some weaker sense than logical contradiction and its being just worse than its competitors. What is more, in considering motivationalism and interpretationalism, I take them to be attitudes towards a particular claim about physical theories (i.e., the claim that symmetry-related models are physically equivalent), not full-blown interpretations of physical theories (which would involve taking a stance towards many more interpretative issues and not just this one). Given this thin understanding of an interpretation, one should not expect that there are very many options to choose from in this context.

[24] Can motivationalism also be defined only by its answer to the first question? I do not think so, as its name comes rather from its answer to the second question. One could, in principle, give the same answer as motivationalism to the first question and then answer "no" to the second question, and this will be yet a different position (which does not *motivate* us to do anything specific with the theory); this view is not considered in this paper (but see footnote 36).

call “interpretationalism without motivation”, “steadfast interpretationalism with motivation”, “concessive interpretationalism with motivation”[25] and “graded interpretationalism with motivation”. They are defined by the answers that they give to our three questions, which are listed below (see also Fig. 1).[26]

Interpretationalism without motivation:

1. We should initially interpret symmetry-related models of *T* as physically equivalent.
2. There is no need to do this.
3. Irrespective of whether we find such an account or not, we should retain our initial interpretation.

Steadfast interpretationalism with motivation:

1. We should initially interpret symmetry-related models of *T* as physically equivalent.
2. Yes.
3. Irrespective of whether we find such an account or not, we should retain our initial interpretation.

[25] Another distinction, not taken into account here, is between strong and weak interpretationalism as understood in Martens and Read (2020). In that paper, weak interpretationalism is understood as a claim that we should *typically* regard symmetry-related models as physically equivalent, but in some cases, “in virtue of certain e.g. theoretical/metaphysical/super-empirical considerations” (Martens & Read, 2020:6), symmetry-related models might be regarded as physically inequivalent. In contrast, strong interpretationalism removes the “typically” clause and does not allow any exceptions to the physical equivalence claim. I do not consider this sense of weak interpertationalism in my paper. First, it seems to me to be formulated in a very ad hoc manner (we should regard symmetry-related models as physically equivalent unless there are some reasons for not doing so…), and it is not clear to me what a supporter of this position commits to. Second, the arguments of motivationalists are directed against both variants and especially the strong one, so by restricting my considerations to the latter, my paper does not lose its polemic value.

[26] This list does not exhaust all the conceivable options. For example, one can also consider “caveated interpretationalism”, whose answer to the first question is that we should initially interpret symmetry-related models of *T* as physically equivalent unless conditions *C* hold. This is in fact a family of possible positions, whose members differ by the specification of conditions *C* (which surely should be different than not knowing a perspicuous account of the ontology shared by symmetry-related models of *T*, as otherwise this variant of interpretationalism would collapse to motivationalism; cf. also footnote 25). One can also suspend judgement about question (1), thereby becoming neither a motivationalist nor an interpretationalist but just an agnostic. The agnostic view might seem to be recommendable because of its cautiousness. However, some epistemologists point out that being overly epistemically cautious might be as problematic as being overly confident (see, e.g., Simion, 2023), although it is difficult to decide at which point caution starts to be excessive. According to Simion (2023:3), “a subject *S* has an epistemic duty to form a belief that *p* if there is sufficient and undefeated evidence for *S* supporting *p*”. In the case of such subtle and multi-faceted controversies as the debate between motivationalism and interpretationalism, one surely cannot say that by choosing one or the other answer to question (1) a person is epistemically blameworthy because of being resistant to some easily available evidence. Nevertheless, there is a more general idea that can be applied here: remaining entirely agnostic with respect to some question, although it might seem epistemically safer, is the most rational stance only if the arguments for all answers are really equipollent; in other cases, we should regard one of these answers as more likely than others, even if it is far from being certain. (Whether the arguments in our case are equipollent or not is another matter, to be discussed in Section 6.) At any rate, as my interest here is in the variants of “pure” interpretationalism (which in their answer to question (1) do not compromise the main idea of interpretationalism, i.e., which claim that symmetry-related models should initially be regarded as physically equivalent without qualification), I will ignore these options (i.e., caveated and agnostic) in what follows.

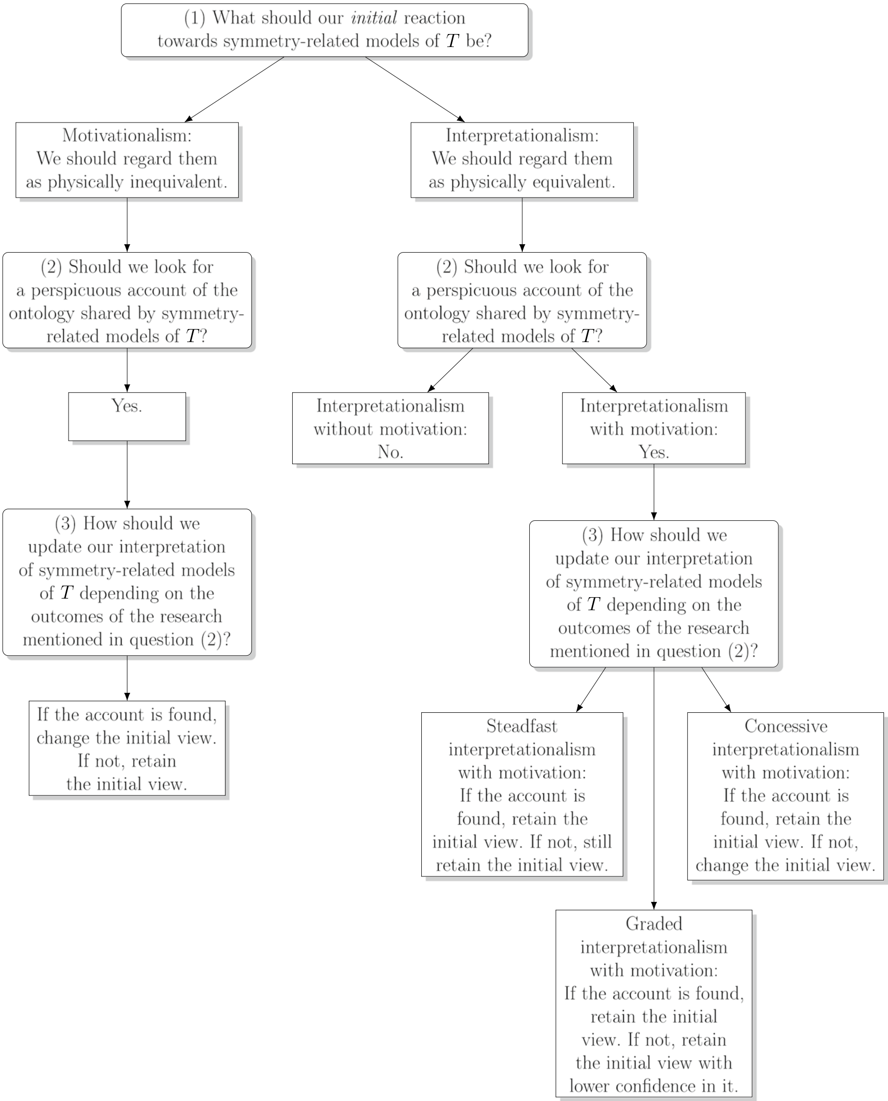
(1) What should our *initial* reaction towards symmetry-related models of T be?
Motivationalism: We should regard them as physically inequivalent.
Interpretationalism: We should regard them as physically equivalent.
(2) Should we look for a perspicuous account of the ontology shared by symmetry-related models of T?
(2) Should we look for a perspicuous account of the ontology shared by symmetry-related models of T?
Yes.
Interpretationalism without motivation: No.
Interpretationalism with motivation: Yes.
(3) How should we update our interpretation of symmetry-related models of T depending on the outcomes of the research mentioned in question (2)?
(3) How should we update our interpretation of symmetry-related models of T depending on the outcomes of the research mentioned in question (2)?
If the account is found, change the initial view. If not, retain the initial view.
Steadfast interpretationalism with motivation: If the account is found, retain the initial view. If not, still retain the initial view.
Concessive interpretationalism with motivation: If the account is found, retain the initial view. If not, change the initial view.
Graded interpretationalism with motivation: If the account is found, retain the initial view. If not, retain the initial view with lower confidence in it.

Fig. 1 Decision tree depicting our three questions and different answers to them given by different positions in the debate. “Changing the initial view” means changing from regarding models as physically equivalent to regarding them as physically inequivalent or the other way around, depending on what the initial view was

Concessive interpretationalism with motivation:

1. We should initially interpret symmetry-related models of T as physically equivalent.
2. Yes.
3. If we find such an account, we should retain our initial interpretation, whereas if despite lots of effort we do not succeed in finding it, we should change our interpretation and begin to regard symmetry-related models of T as physically inequivalent.

Graded interpretationalism with motivation:

1. We should initially interpret symmetry-related models of T as physically equivalent.
2. Yes.
3. If we find such an account, we should retain our initial interpretation, whereas if despite lots of effort we do not succeed in finding it, we should still retain our initial interpretation as more plausible than its opposite, but we should significantly decrease our confidence about this interpretation.

All four variants of interpretationalism give the same answer to the first question, namely, that we should initially interpret symmetry-related models of T as physically equivalent. This is why all of them are counted as interpretationalist positions. The differences come with questions (2) and (3), and in some cases the answers to these questions bring interpretationalism quite close to motivationalism.

Interpretationalism with motivation (in its steadfast, concessive and graded versions) shares with motivationalism the aspiration of finding a perspicuous account of the common ontology of symmetry-related models of T (which is where "with motivation" comes from). These positions differ from motivationalism in what they postulate as the most reasonable attitude for the time being, that is, before significant amounts of time and energy have been put into searching for the shared ontology of symmetry-related models of T; but they do not differ from motivationalism in their view on appropriate directions of reasearch.

The differences between our three variants of interpretationalism with motivation become visible when we take into account the third question. The answer given by concessive interpretationalism with motivation is the same as that given by motivationalism (modulo their difference in the recommended initial reaction). Therefore, concessive interpretationalism with motivation is as close to motivationalism as interpretationalism can be. It differs from motivationalism only in what it claims to be the most reasonable *prima facie* attitude towards models of T that we have recently discovered to be symmetry-related. The answer given by steadfast interpretationalism with motivation is the same as that given by interpretationalism without motivation. Finally, the answer given by graded motivationalism with motivation is more nuanced because it operates on degrees of belief instead of taking beliefs to be an all-or-nothing matter.[27] It enables one to always regard symmetry-related models as physically equivalent, even despite the failures of searching for a perspicuous account of their shared ontology (unlike concessive interpretationalism with motivation), but at the same time is responsive to the outcomes of this research (unlike steadfast interpretationalism with motivation).

[27] This change from beliefs to degrees of belief might suggest that we can use here some tools of Bayesian epistemology. We can start by saying that initially the hypothesis that symmetry-related models of T are physically equivalent should be given some credence between 1 and 0 (excluding these extreme values, as this is neither tautology nor counter-tautology), and then our later credences should be updated on the basis of evidence. However, this approach would require specifying the entire space of possibilities that should be taken into account here and prior probabilities. In the case of such abstract considerations, both tasks are highly non-trivial, if doable at all, and accomplishing them would require a separate paper; also, the concept of evidence in such highly theoretical debates becomes very subtle.

It seems that the authors who use the distinction between motivationalism and interpretationalism usually think of the latter as the same as interpretationalism without motivation. However, Read and Møller-Nielsen (2020a:280) recognise that an interpretationalist might be motivated to search for the shared ontology of symmetry-related models. Therefore, to some extent, they anticipate my distinction between interpretationalism with motivation and without motivation, but they do not appreciate all its consequences and, more importantly, they do not consider anything that corresponds to my question (3). As a result, they do not distinguish different types of interpretationalism with motivation. Instead, they assume that an interpretationalist needs to retain his initial stance, no matter what the outcomes of further research on the theory will be.

What is the value of these subtle distinctions for the debate under consideration? I think that the main benefit of making explicit this larger spectrum of positions is that it shows that we do not need to regard arguments against interpretationalism without motivation as automatically being arguments for motivationalism—they equally strongly speak for various variants of interpretationalism with motivation. As I will argue in more detail in the next section, in fact those previously unrecognised variants of interpretationalism might be perceived as the most promising positions because they are supported by the arguments of both sides of the hitherto debate (reviewed in Sections 3 and 4).

6 Assessing the positions in the debate

In this section, I would like to discuss the merits of different forms of interpretationalism as compared with each other and with motivationalism. I assume that the most important arguments to be taken into account here are (M2), (M3) and (M4) on the side of motivationalism (cf. Section 3), as well as (I1), (I3) and (I4) on the side of interpretationalism (cf. Section 4).[28]

6.1 A response to the dilemma (M3)–(M4)

In this section, I will provide a multi-faceted analysis of (M3) and (M4). I will start with some critical remarks concerning these arguments and later I will show how an interpretationalist might avoid both horns of the dilemma.

Concerning (M3), instrumentalism is not an inherently absurd position, so even showing that an interpretationalist is committed to it does not amount to the *reductio ad absurdum* of interpretationalism. Our assessment of instrumentalism should depend on the strength of arguments for and against scientific realism, which is of course beyond the scope of this paper. Having said this, I will assume from now on, for the sake of argument and in agreement with the defenders of motivationalism, that

[28] I do not regard different arguments as supporting different forms of interpretationalism or motivationalism. Rather, depending on the assessed strength of all arguments for interpretationalism on the one hand and all arguments for motivationalism on the other, one should choose between motivationalism and interpretationalism; and if the chosen option is interpretationalism, further evaluation of the arguments for motivationalism leads to the choice of its specific form (if they are assessed as weak and unimportant, this would be the version without motivation, whereas for the opposite assessment this would be some version of interpretationalism with motivation).

instrumentalism is unwelcome and should be avoided, to see what the chances are for interpretationalism to avoid it.[29]

The argument from incautiousness (i.e., (M4)) seems to presuppose that the assumption of the physical inequivalence of symmetry-related models is somehow more modest than the assumption of their physical equivalence. Otherwise, one could not claim that by endorsing the former one is more cautious than in endorsing the latter. However, this is in tension with (I1). If we agree that the Occam's razor is on the side of an interpretationalist, then the burden of justification should be on anyone who claims that symmetry-related models of *T* are not physically equivalent, including a motivationalist, because it is the supporter of the physical equivalence claim that postulates a sparser ontology. Supposedly, a motivationalist would reply here that it is not clear what ontology an interpretationalist postulates in the first place. However, even if an interpretationalist is not able to provide all the details of his ontology, he surely postulates fewer possible physical differences and acknowledges fewer physical possibilities, so there is a clear sense in which an interpretationalist is more ontologically cautious than a motivationalist.

There is another problem with (M4), namely that it introduces an unwarranted asymmetry in what is required from a motivationalist and from an interpretationalist. It is claimed that an account of the shared ontology of symmetry-related models is not a priori guaranteed to exist. This suggests that an interpretationalist is allowed to assert its existence only if he is fully certain of it. However, it is not clear why we should expect so strong a justification here—why does something less than a certainty not suffice? More importantly, has a motivationalist certainty on his side? I do not think so. Motivationalism assumes that if symmetry-related models are physically equivalent, then there needs to exist a perspicuous account of their shared ontology. But do we have certainty with regard to *this* (conditional) claim? No. We do not have a guarantee that any metaphysically perspicuous picture of physical reality exists at all, which is the presupposition of the motivationalist's imperative to seek one.[30] The reality in itself might be very messy, and we do not have any guarantee that it is nicely ordered. That is, there might be *no true and perspicuous* account of the ontology of the theory *at all* (which is consistent with there being a false and perspicuous one as well as with there being a true and non-perspicuous one). This is a rather radical metaphysical hypothesis, but it does not seem to be self-contradictory and as such cannot be excluded with absolute certainty (so there is no guarantee that it is false).

[29] An anonymous reviewer suggested to me that the point of motivationalists here is not that instrumentalism is wrong, but "that there is no interesting debate left if we're not assuming that we're trying to be realists". However, one can consider a more general debate about "What exists?", which does not presuppose scientific realism, and I took the discussion between motivationalism and interpretationalism to be an instance of this more general debate, which does not exclude instrumentalism with respect to unobservables from the very beginning. In any case, this is for me only an aside point, as for most of this paper I assume that a realist attitude towards theories is a desirable one.

[30] This claim might look surprising, as it was one of the main *motivationalist's* arguments that a perspicuous account of the shared ontology of symmetry-related models might not exist. My point here is that a motivationalist seems to assume that *some* true and perspicuous account of the ontology of the theory must exist, so if there is no perspicuous account according to which symmetry-related models are physically equivalent, then we should regard them as physically inequivalent because our current account of the ontology of the theory *is* perspicuous under the assumption of physical inequivalence. However, this account might not be true, despite its perspicuousness.

We also do not have any guarantee that any physical equivalence of models is underpinned by their sharing the same ontology—there is no contradiction in regarding the facts of physical equivalence as primitive.[31] This would not mean that we can add the relation of physical equivalence to our set of models as we wish, in the spirit of external sophistication. The difference is as follows: external sophistication is an operation on the formalism (which I take to include both syntax and semantics), whereas the physical equivalence being primitive is understood as a metaphysical hypothesis. This hypothesis concerns models, which are abstract entities, but their being physically equivalent or inequivalent is treated here as an objective modal fact that concerns the physical reality and not as something conventional or purely formal. Therefore, we cannot "add" the relation of physical equivalence (or do anything else with it)—it just objectively holds or not; at best, we can try to discover facts about it. Again, I do not claim that this hypothesis is plausible but only that it is logically consistent and as such it cannot be excluded with absolute certainty (so there is no guarantee that it is false).

Perhaps we have good reasons to think that the reality is indeed nicely ordered and that facts of physical equivalence are not primitive (and as such are explainable). The reasons are both inductive, in the broad sense used in this paper (we succeeded several times in describing aspects of this order and in explaining physical equivalences), and methodological (e.g., it is always better to have a theory that is more ordered and gives more explanations). However, we also have good reasons to think that symmetry-related models are physically equivalent (e.g., (I1), (I3) and (I4)). Therefore, the lack of certainty seems to threaten both interpretationalism and motivationalism, although in different ways; but if we lower the epistemic standards below the level of certainty, both approaches can say something in their support.

As I mentioned earlier, I believe that the best way to think about arguments (M3) and (M4) for motivationalism is to regard them as forming a dilemma for interpretationalists: either they are forced to be instrumentalists or they are committed to an incautious speculation. Let us now take a closer look at this dilemma.

Starting with the first horn of the dilemma (i.e., (M3)), none of our four interpretationalist positions is explicitly instrumentalist. Perhaps an interpretationalist without motivation is the closest to this, as an advocate of this view claims that we do not need to look for an account of the ontology underlying the physical equivalence of symmetry-related models. In any case, interpretationalism without motivation seems to be the least recommendable position, given that our aim is to understand our theory T as fully as we can. It is always better to learn something about a theory than not to learn it (cf. footnote 21); in particular, it is always better to know an explicit account of the shared ontology of symmetry-related models, if there is any and if its understanding does not exceed our epistemic capacities, than not to know it. Therefore, we might restrict our considerations to various kinds of interpretationalism with motivation. As they explicitly demand to look for a perspicuous account of the shared ontology of symmetry-related models (cf. their answer to question (2)), they cannot be accused of instrumentalism.

[31] Something along these lines is considered (but rejected) in Dasgupta's (2011) analysis of sophisticated substantivalism: according to him, the difference or identity of two possible worlds might either be regarded as grounded in some facts about them or be a "bare modal claim", where the latter would correspond (in my terminology) to the relation of physical equivalence being primitive.

It should also be noted that scepticism about theoretical distinctions that do not make any difference in what is observable is not the same as excluding unobservable objects from our ontology (cf. Dasgupta, 2011:142). The latter is defining for scientific anti-realism (whose branches are instrumentalism and verificationism),[32] but the former should be worrying for both anti-realists and realists. For example, electrons are unobservable, so they are dismissed from an instrumentalist's ontology; but their postulation makes a difference to our predictions about what is observable, and this is why their existence is endorsed by scientific realists. However, if someone postulated a particle whose existence in principle could not make any difference to what is observable, its existence would not be advocated by instrumentalists and scientific realists alike. Interpretationalism of any kind is sceptical about the (alleged) unobservable physical differences between symmetry-related models—but any empiricist (in the broadest sense of this word) should be sceptical here, a scientific realist no less than an instrumentalist. These remarks, if correct, undermine (M4) as well as the response given by Møller-Nielsen (2017:1262) to the Occamist argument (I1) (see the last paragraph of Section 4.1).

What about the problem raised by motivationalists that to be a realist one needs to know an explicit characterisation of the structure that one is a realist about? A part of the answer is that realism is a general attitude with respect to scientific theories, which should not be sensitive to our level of understanding of the details of a given theory. An interpretationalist surely might say that one should accept as real the ontology posited by our best scientific theories (this is his general realist attitude) while suspending judgement about the details of what is posited by a particular given theory. Such a suspension of judgement does not force an interpretationalist to be an anti-realist. Another part of the answer is that from the fact that we do not have a *full* characterisation of the ontology of a given theory it does not follow that we do not know anything about it. For example, it seems reasonable to say that in the case of NGT, we knew quite a lot about its ontology even before the construction of NCT. Therefore, the mentioned suspension of judgement usually will not amount to complete ignorance.

Let us now turn to the second horn of the dilemma (i.e., (M4)). Again, it can easily be avoided, as an interpretationalist does not need to make a bold claim that an account of the shared ontology of symmetry-related models is always guaranteed to exist, which was the basis for the incautiousness accusation. An interpretationalist's answer to our question (1) does not commit him to this bold claim, and it is this answer that I take to be defining for the interpretationalist position. Instead of making this bold claim, an interpretationalist might just say that the hypothesis that symmetry-related models of T are physically equivalent is more plausible than its opposite because arguments such as (I1), (I3) and (I4) (and perhaps others) tell in favour of it. The fact that certain models of T are symmetry-related alone might not be a sufficient reason for being (almost) certain that they are physically equivalent,

[32] There might be different grounds for restricting acceptable ontologies to observable objects: the statements about unobservables might be regarded as meaningless (which is the case of verificationism), or one can think that they are meaningful but all assertions of the existence of such objects are false (or at least not justified).

but it might be a sufficient reason for regarding the hypothesis that they are physically equivalent as more plausible than the hypothesis that they are physically inequivalent (given the overall knowledge of symmetries and physical theories that humanity has gained so far).[33] This, in turn, might be a sufficient basis for forming a rational belief that they are indeed physically equivalent—that is, for becoming an interpretationalist.

An interpretationalist (with motivation) might also add that this plausibility assessment is retractable if the outcomes of further research would not support it. Indeed, concessive interpretationalism with motivation and graded interpretationalism with motivation treat the answer to question (1) as tentative and revisable in the light of the outcomes of further research on T. This revision may take the form of changing the interpretation of symmetry-related models of T altogether (in the concessive variant) or only in lowering our confidence in the initial interpretation (in the graded variant). An interpretationalist does not need to engage in any speculation concerning these outcomes and might be fully open to whatever the future investigations will bring (or fail to bring).

A proponent of interpretationalism with motivation endorses the need to have a perspicuous account of the shared ontology of symmetry-related models (together with a motivationalist), but he does not require that we should be able to give this picture at the early stages of theorising. Perhaps this is something that can be achieved only at mature stages of theorising, when the theory will be better understood and formally analysed. The tentative endorsement of the physical equivalence of symmetry-related models does not need to wait for this. In this way, an interpretationalist (with motivation) might fully take into account the fact that we are not guaranteed to find an explication of the shared ontology of symmetry-related models and, therefore, cannot be accused of not being cautious enough.

The situation here might be explained by means of an analogy. If we have a promising scientific theory, which has passed some initial empirical tests and scores well in other relevant respects (such as consistency with well-established theories and non-empirical theoretical virtues), and at some point one of the experiments gives a result inconsistent with it, it is often more reasonable to tentatively assume that there might be some shortcomings in the experiment and to retain our belief in the theory (while looking for what exactly these shortcomings are) instead of declaring the theory false until someone shows that the experiment was unreliable. This does not mean that we are not cautious, as we take into account both options: that our promising theory is indeed false and that the experiment is flawed. Choosing one option as more plausible for the time being does not amount to disregarding the other unconditionally.

[33] An alternative path of argumentation, which I do not take here, is developed by Baker (2023). He argues that in some cases it might be justifiable to believe that there exists a reduced theory even though it is not known. I suppose that in my typology he would be a steadfast interpretationalist with motivation because he does not postulate that our failures in finding a reduced theory should decrease our confidence in its existence. He stresses that it might be the case that such a theory exists but that it is "ineffable" to humans.

Now, in the above, replace a promising scientific theory with a thesis that symmetry-related models of T are physically equivalent,[34] and replace an experiment giving a result inconsistent with the theory with the lack of an account of the shared ontology of symmetry-related models of T. The stance of an interpretationalist with motivation is structurally similar to the stance of the tentative supporter of a promising scientific theory described above: he claims that it is more reasonable to tentatively assume that symmetry-related models of T are physically equivalent (while looking for their shared ontology) rather than declaring them physically inequivalent until such ontology is found. He surely is cautions, as he takes into account both options: that we might find such an ontology, thereby supporting the equivalence claim, and that we might fail to find such an ontology, thereby leaving our equivalence claim without further support. Choosing one option as more plausible for the time being does not amount to disregarding the other unconditionally.

Summing up, the above considerations show that interpretationalism with motivation does not fall into any horn of the motivationalist dilemma: it is committed neither to instrumentalism nor to unwarranted speculation about the possibility of finding a perspicuous account of the shared ontology of symmetry-related models.

6.2 A response to (M2)

My preferred response to (M2) is similar to the above response to (M4). A proponent of interpretationalism with motivation might endorse the value of having as many good explanations as we can and might agree with a motivationalist that providing a perspicuous picture of the shared ontology of symmetry-related models is indispensable to avoid significant explanatory losses that otherwise would arise when those models are interpreted as physically equivalent. However, an interpretationalist might further suppose that the full picture of the explanatory structure of the theory is a matter of its advanced understanding, which is not needed for the initial endorsement of the physical equivalence of symmetry-related models.

In other words, one can ask what is the proper order of settling various interpretational issues about a physical theory. Do we need to have a full grasp of the ontology of the theory and its explanatory structure *before* we endorse the physical equivalence of its models? Why not regard the former (i.e., a full ontological and explanatory picture) as a matter of an advanced understanding of the theory, whereas the latter (i.e., the issue of physical equivalence) as something that could be settled (at

[34] Why not replace it with the thesis that they are physically inequivalent? A counterpart of a promising scientific theory in this analogy should be the one of the two theses that is better supported by our current knowledge about T and symmetries in general. But which of them is it, the thesis that symmetry-related models of T are physically equivalent or the thesis that they are inequivalent? The answer depends on our assessment of the relative strength of the arguments for motivationalism and interpretationalism, such as (M1)–(M4) and (I1)–(I4). This is the key issue in the whole debate, and I did not intend in this paragraph to settle it too easily. Instead, my point here is rather a conditional one: *if* one regards the hypothesis that symmetry-related models of T are physically equivalent as having (at least slightly) stronger arguments on its side, even though a perspicuous account of their shared ontology is not known, then it is reasonable to tentatively accept this hypothesis, and there is no incautiousness in this attitude per se.

least tentatively) at early stages of theorising, provided that there are good reasons for doing so? I do not see any rationale for remaining agnostic about the issue of physical equivalence before we reach an advanced understanding of the theory (even though this is the most cautious option); and regarding them as *in*equivalent in such a circumstance is not more cautious or more modest than the opposite view.

Notice that this response does not amount to instrumentalism (since the need for explanatory transparency is fully appreciated), nor does it engage in a speculation of what our advanced understanding of a theory would look like (before we actually have it). Instead, it is claimed that the question about physical equivalence can be (at least tentatively) answered before many other questions about the theory, without excluding that the answers to those other questions will influence our final answer to the question about physical equivalence (which is allowed by interpretationalism with motivation in its concessive and graded version).[35]

6.3 The overall assessment of interpretationalism

As we have seen, a defender of (some variant of) interpretationalism with motivation can respond to the arguments by a motivationalist (and, moreover, this can be done in a more than one way). Additionally, we have seen that there are various arguments for interpretationalism (e.g., (I1), (I3) and (I4)). These two branches of discussion suggest that some variant of interpretationalism with motivation is the best available position because it does justice to the insights of the both sides of the debate.

Of course, one might try to undermine arguments for interpretationalism. For example, if one's favourite concept of symmetry is a dynamical symmetry, then the link between symmetries in this sense and empirical equivalence (which is assumed in (I1)) might be challenged by claiming that some symmetry-variant quantities are observable and their observability enables us to distinguish empirically between symmetry-related models. However, I am not aware of anyone taking this route; even Middleton and Murgueitio Ramírez (2021), who argue that absolute velocities are measurable, agree that boost-related worlds are empirically indistinguishable. The inductive argument, (I3), is also disputable; a lot depends here on which ways of obtaining a perspicuous account of the shared ontology of symmetry-related models are regarded as valid and what constraints are imposed on them (cf. Section 2). Finally, the status of the Hole Argument, (I4), is subject to an extensive debate and, more importantly for our discussion here, its basic version (i.e., for diffeomorphisms in GR) concerns isomorphic models, while the most controversial cases are those involving symmetry-related but non-isomorphic models.

However, these worries do not threaten the crucial point I want to make here, which is the conditional one: if one finds arguments for interpretationalism convincing but wants to avoid the objections raised by motivationalists, then it is possible to have it all by taking up one of the more nuanced interpretationalist positions.

[35] Concerning other possible responses to (M2), see footnotes 8 and 33.

6.4 Further remarks about (I2)

There is one remaining argument for interpretationalism, namely (I2), which I find implausible and agree with its criticism in the literature, but nevertheless it is worthwhile to see how it relates to my taxonomy of interpretationalist positions. From the point of view of this taxonomy, this argument might seem to be an overshoot: if there is a universal method of finding a perspicuous account of the shared ontology of symmetry-related models, then we do not need to consider what to do in the case of a failure of finding such an account, so all variants of interpretationalism will give the same verdict concerning any particular case. Therefore, if (I2) is correct, the differences between different types of interpretationalism become practically irrelevant in the sense that for any particular case they give the same answer (but see footnote 22 about the senses in which they still remain different). However, also in such a case a difference between interpretationalism and motivationalism practically ceases to matter: one can just regard any symmetry-related models as physically equivalent without wondering what one should think about them if the account of their shared ontology was not known, because it is always known. Of course, a motivationalist might not accept the proposed method, but an interpretationalist might also not accept it. I assume that the question of what it means to provide a perspicuous account of the shared ontology of symmetry-related models is, in general, tangential to the dispute between interpretationalism and motivationalism (cf. the ninth paragraph of Section 2), although some answers to this question might render motivationalism or interpretationalism more difficult to defend. All in all, I think that (I2) is best understood as an argument in favour of the strengthened version of steadfast interpretationalism with motivation, for which in the answer to the third question only the first conditional is relevant (i.e., “If the account is found…”) because its antecedent, according to this view, is always true. The difference between our initial and final attitude towards symmetry-related models seems to lose importance here, as they are always the same. However, they are the same not just by fiat but because we have a universal method of finding a perspicuous account of the ontology underlying the physical equivalence of symmetry-related models. Therefore, such a position is still significantly different from interpretationalism without motivation and it is not the case that for this strengthened version of steadfast interpretationalism everything collapses to the answer to the first question.

What about Dewar’s (2019) proposal of external sophistication (see Section 4.2)? It seems to me that it might be understood within my framework in three different ways. The first question is whether he regards his method of external sophistication as providing a perspicuous account of the shared ontology of symmetry-related models. The answer is not obvious because he does not use this term—instead he talks about “changing one’s theory to incorporate the lessons of a symmetry” or getting “rid of the ‘surplus structure’ the symmetry reveals” (Dewar, 2019:495), which might or might not be the same thing as finding such a perspicuous account. If the answer is “no” (as his opponents seem to assume), then the next question is whether he insists on finding, in addition, a perspicuous account of the ontology underlying the physical equivalence of symmetry-related models. If he does not (which is suggested, e.g., by his claim that “we (…) can *rest content* with a theory whose models are isomorphic under that transformation", where "being isomorphic" is achieved via his method of external sophistication; Dewar, 2019:498, emphases mine), then he should be classified as an interpretationalist without motivation. However, at one point he considers an approach that he calls "more ecumenical", according to which "both kinds of construction [i.e., external and internal] are important for fully understanding the structure—in which case, one would desire an internal construction as well" (Dewar, 2019:503), and this looks like steadfast interpretationalism with motivation. Finally, Dewar can sometimes be understood as claiming that his method of external sophistication *does* provide a perspicuous account that we are interested in (especially on pp. 505–506, where he argues that this method really succeeds "in implementing the idea that we should get rid of 'surplus' (that is, symmetry-variant) structure" and that anti-quidditism about physical properties allows one to get "clear on what ontological commitment has been relinquished in the passage from an unsophisticated to a sophisticated semantics"). Under this reading, his proposal is consistent with any of the options in Fig. 1 and cannot be classified further without answering the question concerning a counterfactual situation, namely: "Should we always look for a perspicuous account of the shared ontology of symmetry-related models *if we did not know* a universal method of finding it?".

6.5 A historical dimension of the debate

As the last point of this discussion, let me make a historical comment. The question of whether one should regard symmetry-related models of classical mechanics as equivalent is sometimes phrased as a question of whether Newton made a mistake by not regarding them as such (see, e.g., Dasgupta, 2016:854). The motivationalist's answer to this question is "no": Newton did not have an account of their shared ontology, so it was reasonable for him to regard them as physically inequivalent.[36]

What should an interpretationalist's answer be? I believe that it is best to think about interpretationalism as a position that has grown out of the historical cumulation of knowledge about physical theories. Therefore, it does not entail that Newton should have regarded symmetry-related models of his theory as physically equivalent because knowledge about symmetries and physical theories was much less advanced in his times. In particular, he could not use the inductive argument, as in his times the inductive basis that is at our disposal today was not available—it was then only starting to be developed. However, now our inductive basis *is* significant (and we also have other, more theoretical arguments, which are based on decades of formal and conceptual work), so we do have reasons for endorsing the interpretational approach to symmetries that Newton did not have.

[36] Actually, perhaps historical Newton was a representative of a position mentioned in footnote 24 (i.e., the view that symmetry-related models are physically inequivalent and we should *not* be motivated to seek an account of their shared ontology). Not only did he regard space and time as absolute and, as a consequence, possible situations related by spatiotemporal symmetries as inequivalent, but also he did not seek an alternative ontology because the absolutness assumption was for him basic rather than being a side effect of the choice of a formalism. However, developing this point would require a deeper textual study for which there is no place here.

7 Summary

My first aim in this paper was to point out that there are some overlooked options in the debate between motivational and interpretational approaches to symmetries. These options become visible once we carefully distinguish the different theses that constitute motivationalism, formulate the questions to which they are answers and reflect on what possible answers to these questions might be given by an interpretationalist.

My second aim was to argue that some version of interpretationalism with motivation is superior to motivationalism. These positions are neither committed to an instrumentalist way of thinking about physical theories and nor are they less cautious than motivationalism, but they take into account our existing knowledge about symmetries more comprehensively than motivationalism does. Given this knowledge, the initial plausibility of the hypothesis that symmetry-related models of some physical theory are physically equivalent seems to be greater than the initial plausibility of the opposite hypothesis, so it is the former that should be a preferred initial interpretation. However, according to concessive and graded interpretationalism with motivation, this initial interpretation (or at least our level of confidence in it) might change with new results of the research on the theory in question (or with the lack of expected results, despite significant effort).

What are the benefits of drawing these fine-grained distinctions between different types of interpretationalism (besides our awareness of the broader spectrum of options in the debate)? I think that at the general level, the main advantage is the following: once we recognise that besides motivationalism and interpretationalism without motivation there are other possible positions, we can *both* (i) be motivated to find a perspicuous account of the shared ontology of symmetry-related models and (ii) regard these models as physically equivalent in the meantime (or at least think that the hypothesis of their physical equivalence is more plausible than the hypothesis of their physical inequivalence). Motivationalism affirms (i) but rejects (ii), whereas interpretationalism without motivation affirms (ii) and rejects (i); all other variants of interpretationalism discussed in this paper accept both (i) and (ii). In this way, as I have argued in Section 6, we can do justice to the arguments for both interpretationalism and motivationalism.

This general observation transfers rather straightforwardly to the analysis of specific case studies, although their details will be important for the assessment of whether we succeeded in finding a perspicuous account of the ontology underlying the physical equivalence of symmetry-related models, which (depending on our answer to question (3)) might have consequences for our final view on the physical (in)equivalence of the symmetry-related models. For example, in the case of the debate about the Aharonov–Bohm effect, all variants of interpretationalism with motivation would recommend that we initially regard models related by a gauge transformation (but differing by the values of A_μ) as physically equivalent, even if none of the available accounts of their shared ontology is sufficiently perspicuous.[37] If the results of our repeated attempts to find the account

[37] Concerning the known candidates for such an account and their assessment, see, for example, Healey (2007) and Jacobs (2021b, ch. 7).

that is fully perspicuous will be unsatisfactory, we might need to change this initial position (if our chosen variant is concessive interpretationalism with motivation), lower our confidence in it (if our chosen variant is graded interpretationalism with motivation) or adhere to it no matter what these results are but without ceasing to look for such an account (if our chosen variant is steadfast interpretationalism with motivation). An analogous analysis can be applied to other cases, which does not mean that my framework offers an "algorithmic" strategy for dealing with them. This is because a lot depends on answering further questions, such as: under what conditions an account of the shared ontology of symmetry-related models should be regarded as (sufficiently) perspicuous, what time and efforts are sufficient to say (at least tentatively) that we failed in finding it (this is important for concessive and graded interpretationalism with motivation) and so on.

Acknowledgements I would like to thank Tomasz Placek and anonymous reviewers for numerous comments that helped me to formulate better my view and arguments in favour of it. This work was supported by National Science Center in Poland, research grant Preludium 2017/25/N/HS1/00705.

Author contributions This paper has a single author.

Funding National Science Centre in Poland, research grant Preludium 2017/25/N/HS1/00705.

Declarations

Financial or non-financial interests N/A

Ethical approval N/A (my research does not involve human participants)

Informed consent N/A (my research does not involve human participants)

References

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

Baker, D. J. (2023). What are symmetries? *Ergo, 9*(67), 1784–1805.

Belot, G. (2013). Symmetry and equivalence. In R. Batterman (Ed.), *The Oxford Handbook of Philosophy of Physics* (pp. 318–339). Oxford University Press.

Belot, G. (2018). Fifty million Elvis fans can't be wrong. *Noûs, 52*(4), 946–981.

Brading, K., & Castellani, E. (2007). Symmetries and invariances in classical physics. In J. Buttrefield & J. Earman (Eds.), *Philosophy of Physics* (pp. 1331–1367). Elsevier.

Brighouse, C. (1994). Spacetime and holes. *PSA: The Proceedings of the Biennial Meeting of the Philosophy of Science Association* , 117–125. https://doi.org/10.1086/psaprocbienmeetp.1994.1.193017

Castellani, E. (2003). Symmetry and equivalence. In K. Brading & E. Castellani (Eds.), *Symmetries in Physics: Philosophical Reflections* (pp. 425–436). Cambridge University Press.

Caulton, A. (2015). The role of symmetry in the interpretation of physical theories. *Studies in History and Philosophy of Modern Physics, 52*, 153–162.

Dasgupta, S. (2011). The bare necessities. *Philosophical Perspectives, 25*, 115–160.

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., & Redhead, M. L. G. (2007). *Objectivity, Invariance and Convention: Symmetry in Physical Science*. Harvard University Press.

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

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

Earman, J. (1989). *World enough and space-time*. The MIT Press.

Earman, J., & Norton, J. (1987). What Price Spacetime Substantivalism? The Hole Story. *British Journal for the Philosophy of Science, 38*, 515–525.

Fletcher, S. C. (2020). On Representational Capacities, with an Application to General Relativity. *Foundations of Physics, 50*, 228–249.

Gomes, H. (2022a). *Same-diff? Conceptual similarities between gauge transformations and diffeomorphisms*. Part I: Symmetries and isomorphisms. https://arxiv.org/abs/2110.07203v2

Gomes, H. (2022b). *Same-diff? Conceptual similarities between gauge transformations and diffeomorphisms*. Part II: Challenges to sophistication. https://arxiv.org/abs/2110.07204v2

Healey, R. (2001). On the Reality of Gauge Potentials. *Philosophy of Science, 68*(4), 432–455.

Healey, R. (2007). *Gauging What's Real*. Oxford University Press.

Hoefer, C. (1996). The metaphysics of spacetime substantivalism. *Journal of Philosophy, 93*, 5–27.

Ismael, J., & van Fraassen, B. (2003). Symmetry as a guide to superfluous theoretical structure. In K. Brading & E. Castellani (Eds.), *Symmetries in physics: Philosophical reflections* (pp. 371–392). Cambridge University Press.

Jacobs, C. (2021a). Invariance or equivalence: A tale of two principles. *Synthese, 199*, 9337–9357.

Jacobs, C. (2021b). Symmetries as a guide to the structure of physical quantities [PhD thesis]. University of Oxford. https://ora.ox.ac.uk/objects/uuid:2cef8463-70fe-4a27-9cc3-c430eb430c37.

Jacobs, C. (2022). Invariance, intrinsicality and perspicuity. *Synthese, 200*, 135.

Luc, J. (2022). Arguments from scientific practice in the debate about the physical equivalence of symmetry-related models. *Synthese, 200*, 72.

Martens, N. C. M., & Read, J. (2020). Sophistry about symmetries? *Synthese*, *199*, 315–344.

Middleton, B., & Murgueitio Ramírez, S. (2021). Measuring absolute velocity. *Australasian Journal of Philosophy, 99*(4), 806–816.

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

Maudlin, T. (1993). Buckets of water and waves of space why spacetime is probably a substance. *Philosophy of Science, 60*(2), 183–203.

Pooley, O. (2006). Points, Particles, and Structural Realism. In D. Rickles, S. French, & J. Saatsi (Eds.), *The Structural Foundations of Quantum Gravity* (pp. 83–120). Oxford University Press.

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

Pooley, O. (2017). Background Independence, Diffeomorphism Invariance and the Meaning of Coordinates. In D. Lehmkuhl, G. Schiemann, & E. Scholz (Eds.), *Towards a Theory of Spacetime Theories* (pp. 105–143). Birkhäuser.

Quine, W. V. (1951). On what there is. *The Review of Metaphysics, 2*(5), 21–38.

Read, J., & Møller-Nielsen, T. (2020a). Motivating dualities. *Synthese, 197*, 236–291.

Read, J., & Møller-Nielsen, T. (2020b). Redundant epistemic symmetries. *Studies in History and Philosophy of Modern Physics, 70*, 88–97.

Roberts, J. T. (2008). A Puzzle about Laws, Symmetries and Measurability. *British Journal for the Philosophy of Science, 59*(2), 143–168.

Rosenstock, S., Barrett, T., & Weatherall, J. O. (2015). On Einstein algebras and relativistic spacetimes. *Studies in History and Philosophy of Modern Physics, 52B*, 309–316.

Rynasiewicz, R. (1992). Rings, Holes, and Substantivalism: On the Program of Leibniz Algebras. *Philosophy of Science, 59*(4), 572–589.

Saunders, S. (2003). Physics and Leibniz's principles. In K. Brading & E. Castellani (Eds.), *Symmetries in Physics: Philosophical Reflections* (pp. 289–307). Cambridge University Press.

Simion, M. (2023). Resistance to evidence and the duty to believe. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12964

Wallace, D. (2022). Observability, redundancy, and modality for dynamical symmetry transformations. In J. Read & N. Teh (Eds.), *The Philosophy and Physics of Noether's Theorems: A Centenary Volume* (pp. 322–353). Cambridge University Press.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.