On some examples from first-order logic as motivation for categorical equivalence of KPMs

Eleanor March[*]

Abstract

I develop and motivate an extension of the categorical equivalence programme to the full space of KPMs of a theory, beginning with a problem case from first-order logic where categorical equivalence seems too weak a criterion of theoretical equivalence. This has wide-ranging implications for discussions of theoretical equivalence, and the categorical equivalence programme in particular.

1 Introduction

In recent years, there has been a surge of interest in the following criterion of theoretical equivalence:

Categorical equivalence: Theories T_1 and T_2 are equivalent just in case there is an equivalence of categories between their associated categories of models $\mathbf{T}_1$ and $\mathbf{T}_2$ (which preserves empirical content).[1]

Categorical equivalence is motivated by the fact that the collection of models of a theory often has, or can be given, the structure of a category. In general, a category $\mathbf{C}$ consists in a collection of objects $\mathrm{ob}(\mathbf{C})$ and a collection of arrows (or morphisms) $\mathrm{mor}(\mathbf{C})$ between the objects of $\mathbf{C}$, satisfying several conditions. Here, it is useful to introduce the concept of *hom-sets*, which are the sets of arrows between two objects in a category, i.e. for $a, b \in \mathrm{ob}(\mathbf{C})$, one defines $\hom_{\mathbf{C}}(a, b) := \{f | f : a \to b, f \in \mathrm{mor}(\mathbf{C})\}$. Then a category $\mathbf{C}$ is:

- A collection of objects $\mathrm{ob}(\mathbf{C})$;
- For all $a, b \in \mathrm{ob}(\mathbf{C})$, a function $(a, b) \to \hom_{\mathbf{C}}(a, b)$;
- (Composition of arrows) For all $a, b, c \in \mathrm{ob}(\mathbf{C})$, all $g \in \hom_{\mathbf{C}}(b, c)$, and all $f \in \hom_{\mathbf{C}}(a, b)$, a function $(g, f) \to g \circ f$;

[*] Faculty of Philosophy, University of Oxford. eleanor.march@philosophy.ox.ac.uk

[1] Recall that an equivalence of categories $\mathbf{C}$ and $\mathbf{D}$ is a functor $F : \mathbf{C} \to \mathbf{D}$ which is full, faithful, and essentially surjective (see e.g. Mac Lane (1998) for details, and Weatherall (2017) for a philosophically-oriented presentation). What it means for an equivalence of categories to 'preserve empirical content' is difficult to spell out in general terms, and I won't attempt to do so here, but is usually reasonably clear within a given context.

- (Identity arrow) For all $a \in \mathrm{ob}(\mathbf{C})$, an arrow $\mathrm{id}_a \in \hom_{\mathbf{C}}(a, a)$;

satisfying

1. (Associativity) For all $a, b, c, d \in \mathrm{ob}(\mathbf{C})$, all $h \in \hom_{\mathbf{C}}(c, d)$, all $g \in \hom_{\mathbf{C}}(b, c)$, all $f \in \hom_{\mathbf{C}}(a, b)$, $h \circ (g \circ f) = (h \circ g) \circ f$.
2. (Unity) For all $a, b, c \in \mathrm{ob}(\mathbf{C})$, all $g \in \hom_{\mathbf{C}}(b, c)$, and all $f \in \hom_{\mathbf{C}}(a, b)$, $\mathrm{id}_b \circ f = f$ and $g \circ \mathrm{id}_b = g$.
3. (Disjointness) If $(a, b) \neq (a', b')$ then $\hom_{\mathbf{C}}(a, b) \cap \hom_{\mathbf{C}}(a', b') = \varnothing$.

Given a theory T, one can take the objects of its associated category of models $\mathbf{T}$ to be the models of T, and its arrows to be maps between models which preserve physical content (which maps these are will depend on one's interpretation of T, though for all the first-order theories considered in this article, I will take the arrows of their associated categories of models to be elementary embeddings unless stated otherwise).[2]

Beginning with suggestions made by Halvorson (2012), Weatherall (2016), and Tsementzis and Halvorson (2017), categorical equivalence has been used to provide insight into the relationships between several different pairs of physical theories. For example, Weatherall (2016) discusses the relationship between Galilean gravitation and Newton-Cartan theory, and distinguishes two different categories of models for Galilean gravitation, only one of which is categorically equivalent to Newton-Cartan theory. Rosenstock, Barrett, and Weatherall (2015) use categorical equivalence to substantiate claims by Rynasiewicz (1992) about the equivalence of general relativity and the theory of Einstein algebras. And Barrett (2019) makes use of categorical equivalence to adjudicate a dispute between North (2009) and Curiel (2014) about the (in)equivalence of Lagrangian and Hamiltonian mechanics, showing how different categories of models for these theories can be used to articulate North's view, Curiel's view, and the 'standard' view that Lagrangian and Hamiltonian mechanics are equivalent.

All these, I take it, have been success stories for the categorical equivalence programme. And yet, categorical equivalence is also known to face a serious problem: it seems far too weak a criterion of theoretical equivalence. This point has been made at length by Hans Halvorson and co-authors in a series of papers (Halvorson 2012; Barrett and Halvorson 2016; Tsementzis and Halvorson 2017; Barrett and Halvorson 2022),[3] with reference to the following example of

[2] This approach is in the spirit of e.g. Weatherall (2016), Rosenstock, Barrett, and Weatherall (2015), Barrett (2019), and Nguyen, Teh, and Wells (2020), though note that when Barrett and Halvorson (2016, 2022) and Tsementzis and Halvorson (2017) talk about theories as categories, they have in mind first-order theories, and categories whose objects are the theory's models and whose arrows are elementary embeddings. Whilst this approach is sensible (and actively useful, when it comes to comparing categorical equivalence with other criteria of theoretical equivalence for first-order theories), I also think it is too restrictive. The reason for this is that there might be cases in which one wishes to interpret some of the predicate/relation/sort/function symbols in some signature as not representing anything physical, in which case, there will be good reason to consider some pairs of elementarily inequivalent models of one's theory physically equivalent. We will return to this issue later.

[3] See also Weatherall (2020).

a pair of categorically equivalent, but not definitionally equivalent nor Morita equivalent[4] theories from first-order logic:

Example 1. Let T_1 be a theory in signature Σ_1 containing a single sort symbol σ_1 and a countable infinity of predicate symbols p_0, p_1, p_2, etc. And let T_2 be a theory in signature Σ_2 containing a single sort symbol σ_2 and a countable infinity of predicate symbols q_0, q_1, q_2 etc. T_1 has as its axioms the sentence $\exists!_{\sigma_1} x x = x$. T_2 has as its axioms the sentences $\exists!_{\sigma_2} x x = x$ and a countable infinity of sentences $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, \ldots$.

The purpose of this paper is to point out that there is a relatively straightforward and general way of handling this counterexample, which has not so far received any attention in the literature, but which seems to me to diagnose exactly what goes wrong in the case of the theories T_1 and T_2 which Halvorson and co-authors discuss. In particular, the way of handling example 1 which I will propose does not involve appealing to the fact that the categories of models of T_1 and T_2 are not object-finite—a point which, though Barrett and Halvorson (2022) dwell on it in some detail, seems to me to be largely irrelevant to the intuition that T_1 and T_2 are inequivalent theories. Instead, I will begin by getting clear on what is behind the intuition that T_1 and T_2 are inequivalent theories, use this to motivate an extension of the categorical equivalence programme which can handle the example of T_1 and T_2, and then show how my approach allows us to make sense of a variety of other cases, including the kind of interpretations under which it would make sense to think of T_1 and T_2 as equivalent theories.

In a bit more detail, then, the structure of this paper will be as follows. In §2, I begin by outlining some intuitions about why it is that T_1 and T_2 are inequivalent theories, and suggest a way of making precise these intuitions. This takes us to §3, in which I discuss (in §3.1) the kinematics-dynamics distinction, and use this to offer a generalisation of the categorical equivalence programme which can handle the case of T_1 and T_2 (in §3.2). After discussing some implications of this criterion for debates about theoretical equivalence for first-order theories, I then, in §4, consider some further implications, especially for criticisms of the categorical equivalence programme made by Weatherall (2020) and Coffey (2014). §5 concludes.

2 Some intuitions, and a motivation

T_1 and T_2 are (intuitively speaking) inequivalent theories: why? The discussions in Halvorson (2012) and Barrett and Halvorson (2016) of this matter seem to me exemplary, so I will begin by quoting them at length:

> [There] is a sense in which the two theories [T_1 and T_2] do not "say the same thing." According to the theory T_2, there is a special predicate q_0. If the predicate q_0 holds, that completely determines what

[4] For details on definitional and Morita equivalence, see Barrett and Halvorson (2016).

> else is true according to T_2. The theory T_1, however, singles out no such predicate. (Barrett and Halvorson 2016, 23)

> [Our] gut tells us that these two theories are inequivalent. We might reason as follows: [T_1] tells us nothing about the relations between the predicates; but [T_2] stipulates a non-trivial relation between one of the predicates and the rest of them. In this case, our gut feeling is correct: the theories [T_1 and T_2] are not definitionally equivalent. Indeed, similar to the case of propositional theories, the predicate [$q_0(x)$] cannot be defined in terms of the theory [T_1]. (Halvorson 2012, 11)

I think that these intuitions are exactly right, and I want to hold onto them. T_1 and T_2 are inequivalent theories because T_2 'says something' non-trivial about the relationship between the predicate q_0 and the rest.

What I want to point out here is that there is a very straightforward way of articulating this idea in category-theoretic terms. Consider all those Σ_2-structures satisfying $\exists!_{\sigma_2} xx = x$. The category of these Σ_2-structures (whose arrows are elementary embeddings) will be categorically equivalent to the category of models of T_1. But only a proper subset of these Σ_2-structures will be models of T_2. In other words, T_2 says something more than T_1 (and the two theories are inequivalent) because any equivalence of categories between single-object Σ_2-structures and single-object Σ_1-structures will not respect the property of solutionhood of T_1 and T_2.

But note that in order to say this, we had to broaden our focus beyond the space of models of T_1 and T_2. This suggests that taking into account a larger space of structures than just the models of a theory might be just what is needed to handle the case of T_1 and T_2, and to develop the categorical equivalence programme into something workable. What could this larger space of structures be? In what follows, I will suggest that it has to do with the notion of *kinematical possibility.*

3 Kinematical categorical equivalence

3.1 Kinematics and dynamics

Dynamically possibly models (DPMs) represent the worlds which are physically possible according to the laws of a theory. Kinematically possible models (KPMs) represent the worlds which have the right ontological ingredients, but which are not physically possible according to the laws. The laws pick out the space of DPMs, but the space of KPMs is often less clearly articulated. This much is well known.

Still, there is a guiding principle one often hears in discussions of kinematical possibility, and that is the Lewisian principle of free recombination. Very roughly, the principle of free recombination says that arbitrary combinatorial arrangements of objects in the theory's models count as kinematical possibilities of the theory. Free recombination is supposed to be a consequence of what is known as Hume's dictum:

Hume's dictum: There are no necessary connections between wholly distinct entities.

So Hume's dictum says that if a theory's models contain a collection of objects X_i, and all the X_i are wholly distinct, then any combination of values for the X_i defines a kinematical possibility of the theory. To say otherwise would be to say that there are relationships between the objects X_i which cannot, according to the theory, be violated even in principle. *Prima facie*, then, we might say that given a theory whose DPMs are structures of the form $\langle X_1, X_2, ...\rangle$, any combination of values for the X_i is a KPM of the theory.

Care is needed, however, for we can already make sense of this being the wrong move. We can illustrate this by means of the following example. Full Newtonian spacetime, as presented by Earman, is a structure $\langle M, t_a, h^{ab}, \nabla, \xi^a\rangle$, where M is a differentiable four-manifold, t_a and h^{ab} are compatible temporal and spatial metrics, ∇ is a compatible flat derivative operator, and ξ^a is a unit timelike vector field such that $\nabla_a \xi^b = 0$. However, this structure carries some redundancy: given a structure $\langle M, t_a, h^{ab}, \xi^a\rangle$ with $\pounds_\xi h^{ab} = 0$, we can always define ∇ as the unique compatible torsion-free special connection for ξ^a. As a result, several authors choose to present full Newtonian spacetime as a structure $\langle M, t_a, h^{ab}, \xi^a\rangle$.

Now, I take it that the decision to present Newtonian spacetime as $\langle M, t_a, h^{ab}, \nabla, \xi^a\rangle$ or $\langle M, t_a, h^{ab}, \xi^a\rangle$ is essentially a matter of notational preference, and not a substantive one. However, if we take the principle of free recombination at face value, these two presentations of Newtonian spacetime suggest very different spaces of KPMs for the theory (here I am setting aside related questions of whether e.g. metric compatibility conditions should also be included in the principle of free recombination). In particular, the space of structures of the form $\langle M, t_a, h^{ab}, \xi^a\rangle$ will be a proper subspace of structures of the form $\langle M, t_a, h^{ab}, \nabla, \xi^a\rangle$.

What has gone wrong here, and what goes wrong with the related question of compatibility conditions alluded to above, is that in the first presentation of Newtonian spacetime $\langle M, t_a, h^{ab}, \nabla, \xi^a\rangle$, some of the structures in the theory's models are being (perhaps partially) defined with reference to other structures in these models. In other words, the antecedent of Hume's dictum—there are no necessary connections between wholly distinct entities—fails to apply, since we shouldn't think of the various constituent structures which enter into the description of Newtonian spacetime as conceptually independent theoretical posits.

This suggests that care is needed in identifying which of the objects in a theory's models can be regarded as conceptually independent, before we apply the principle of free recombination to define the space of KPMs. In particular, there may be certain equation-like statements which appear in the dynamics of the theory which should be regarded as kinematical constraints. There might also be disagreements about which of the objects in a theory's models should be regarded as independent theoretical posits. For example, consider a slightly modified version of Newtonian gravitation (restricted to the island universe sector) in which we choose to define the 'standard of rest' as the centre of mass velocity of the universe. Plausibly, one might take this to mean that the matter fields represented by the mass-momentum tensor T^{ab} and the unit-timelike vector field ξ^a are no longer conceptually independent, or 'wholly distinct entities' as it were, so that we should take the demand that ξ^a is the centre of mass velocity field as a kinematical constraint (this is my own view). But one might also insist that the two *can* conceptually come apart, in such a theory, so that there will be KPMs in which ξ^a does not coincide with the centre of mass velocity of the universe.

One way in which a naïve application of the principle of free recombination can fail is if not all the objects in a theory's models are ontologically or conceptually independent. There are two other ways in which it can fail. The first is if some of the axioms of a theory serve as (perhaps partial) definitions of some of the objects in the theory's models. For example, the condition ${R^a}_{bcd} = 0$ in Galilean spacetime $\langle M, t_a, h^{ab}, \nabla \rangle$ is plausibly like this: the Galilean connection is defined to be flat. Or, consider the fact that thermodynamic temperature parameterises a partition on equilibrium states of a system. Plausibly, this is a definition of what it is to be thermodynamic temperature. For a more mathematics-oriented example, consider the fact that the relation $\leq$ (on, e.g. $\mathbb{R}$) is reflexive, antisymmetric, and transitive. In other words, not any value for some given object X_i in the models of a theory will define a kinematical possibility of the theory, because it might be part of the definition of X_i that some of its degrees of freedom are related to one another in non-trivial ways.

The second is if some of the axioms of a theory express domain restrictions. Theories are generally not formulated in the manner of a theory of everything: domain restrictions tell us what kind of systems the theory is able to treat. For instance, the theory of general relativity (GR) is usually formulated on a smooth, connected, paracompact, Hausdorff manifold. These conditions are plausibly thought of as domain restrictions. Of course, one can drop some of these—cf. especially discussions of non-Hausdorff GR (e.g. Luc and Placek (2020))—but this is generally done at the level of dynamics, rather than kinematics.

This is enough to make clear the flavour of my proposal for what we should take the KPMs of a theory to be. First, one identifies which objects (or which degrees of freedom of these objects) in the DPMs of the theory are to be regarded as conceptually and ontologically independent. Second, one identifies which degrees of freedom of the objects in a theory's models are tied up by definitions or domain restrictions. Finally, one applies the principle of free recombination to the remaining (degrees of freedom of these) objects. This can be done by identifying the kinematical constraints of the theory—conditions which describe definitions, domain restrictions, or how those degrees of freedom of objects in the theory's models which are not conceptually independent are related to one another—and defining the space of KPMs as structures which satisfy those constraints. Note that this also goes the other way. If we have stipulated that some equation-like statements are to count as kinematical constraints, then either they express definitions, or domain restrictions, or we cannot interpret the objects (or degrees of freedom of the objects) related by these constraints as 'wholly distinct' i.e. as conceptually independent theoretical posits.

It is instructive to compare my proposal for how one is to construct the kinematical possibilities of a theory with what is (as far as I know) the only other well-developed proposal in the literature, due to Curiel (2016). Curiel's kinematics-dynamics distinction, though he doesn't phrase it in quite these terms, draws on a distinction between what one might call the basic dynamical objects of the theory (positions, velocities, electromagnetic fields etc.) which always take the same concrete form, and what one might call placeholder variables (forces, matter currents, etc.) whose concrete form varies depending on the particular interactions into which the system enters. Then kinematical constraints are equation-like statements which feature only the basic dynamical objects, whereas dynamical laws are equation-like statements which feature placeholder variables.

Whilst the proposal is interesting, I don't think it works. The reason for this is that it is too coarse-grained. Simply put, Curiel's distinction misclassifies as kinematical some constraints which one might (though need not) want to treat as dynamical, and misclassifies as dynamical laws some constraints which one might (though again, need not) want to treat as kinematical. For the first, consider the Gauss-Faraday law in electromagnetism $\mathrm{d}_a F_{bc} = 0$. According to Curiel, this is a kinematical constraint, since it involves only the basic dynamical variables of the theory F_{ab}. However, Jacobs (2021) has argued that in the Faraday tensor formulation of electromagnetism, the fact that F_{ab} is closed is a 'cosmic conspiracy'—if the Faraday tensor is fundamental, then surely it could have had any value. So if Jacobs is right that fundamentality entails modal freedom, and one thinks that the Faraday tensor is fundamental, there seems good reason to take the Gauss-Faraday law as a dynamical law. For the second, Curiel's distinction classifies Newton's second law as a dynamical constraint, since the concrete form of the forces involved will depend on the particular interactions which the system enters into. However, I have argued (March 2024) that Newton's second law should be thought of as an implicit definition of the connection from the matter fields and standard of rotation, and so is a kinematical constraint. Irrespective of whether or not one agrees with my proposal, for it even to make sense requires a finer-grained kinematics-dynamics distinction than Curiel's.

With that said, my kinematics-dynamics distinction does have some interesting features in common with Curiel's—in particular, when it comes to the classification of constraints as kinematical or dynamical in mathematical theories. On Curiel's kinematics-dynamics distinction, the axioms of mathematical theories always express kinematical constraints, since theories in mathematics do not contain placeholder variables which depend on the concrete interactions of physical systems. My kinematics-dynamics distinction also has this consequence, albeit for a slightly different reason. This is because the various axioms relating objects which enter into the description of some kind of mathematical structure (e.g. a vector space, group, ring, smooth manifold, poset, Hausdorff space etc.) *define* what it is to be that kind of mathematical structure,[5] and so will end up expressing definitions, or domain restrictions, or else as relationships between objects which are not conceptually independent from one another. For example, consider the fact that ring multiplication is distributive over addition. Plausibly, this is part of what it is to *be* ring multiplication. But if part of what it is to be ring multiplication is to be distributive over addition, then we shouldn't think of ring multiplication and addition as conceptually independent from one another—rather, they are two operations which jointly describe ring structure. Or, consider the fact that vector addition is associative and commutative. This is just part of the definition of vector addition. Finally, consider the Hausdorff condition in the theory of Hausdorff spaces. This is not obviously thought of as a condition which expresses a relationship between structures in a topological space $\langle X, \tau \rangle$ which are not conceptually independent of each other, nor is it obviously thought of as a definition of either of these objects. But it is very obviously thought of as a domain restriction: the theory of Hausdorff spaces is concerned with just those topological spaces which are (surprise, surprise!) Hausdorff.

As I will discuss in §4, I think that the fact that the axioms of theories in mathematics are plausibly thought of as kinematical constraints goes some way towards explaining why, though categorical equivalence often seems to be too weak a criterion of theoretical equivalence for physical theories, it also often (if not always) seems to give the right verdicts for mathematical theories. If the axioms of mathematical theories often (if not always) express kinematical constraints, this means that the dynamical and kinematical possibilities of mathematical theories will coincide. And when they do coincide, we should expect categorical equivalence to be a good criterion of theoretical equivalence (this point will be made precise later on).

3.2 A new criterion

With the kinematics-dynamics distinction on the table, we can now articulate a straightforward generalisation of the categorical equivalence programme. Let T be a theory. We can define a subtheory T^k of T whose axioms are just the kinematical constraints of T. Call T^k its associated kinematical theory. Given a theory T with associated kinematical theory T_k, we can associate a category of models $\mathbf{T}^k$ to T^k in much the same way as before: objects of $\mathbf{T}^k$ are models of T^k, and arrows are maps between models which preserve physical content, subject to the following conditions:

- If T has associated category of models $\mathbf{T}$, then $\mathbf{T}$ is a full subcategory of $\mathbf{T}^k$ (this captures the idea that which maps between models are physical

[5] Consider e.g. the fact that structures which fail to satisfy the axioms of some mathematical theory are said to fall outside the domain of that theory (rather than the theory being false of those structures), or that mathematically useful structures which fail to satisfy all the axioms of some mathematical theory are often dignified with alternative names (presheafs, rngs, psuedometrics etc.).

equivalences shouldn't depend on whether one is looking at the models of T or the broader collection of models of T^k).

- If $\mathfrak{M} \in \text{ob}(\mathbf{T})$ and $\mathfrak{M}' \in \text{ob}(\mathbf{T}^k)\backslash\text{ob}(\mathbf{T})$ then $\hom_{\mathbf{T}^k}(\mathfrak{M}, \mathfrak{M}') = \hom_{\mathbf{T}^k}(\mathfrak{M}', \mathfrak{M}) = \varnothing$ (this captures the idea that the solutions of a theory's equations of motion should not be taken as physically equivalent to non-solutions of the equations of motion).

I can now state the following criterion of theoretical equivalence:

Kinematical categorical equivalence Let T_1, T_2 be theories, and let T_1^k, T_2^k be their associated kinematical theories. Let $\mathbf{T}_1^k$, $\mathbf{T}_2^k$ denote their associated categories of models. Then T_1, T_2 are kinematically categorically equivalent just in case there is an equivalence of categories $F : \mathbf{T}_1^k \to \mathbf{T}_2^k$ such that for all $M_1^k \in \text{ob}(\mathbf{T}_1^k)$, $F(M_1^k) \in \text{ob}(\mathbf{T}_2)$ iff $M_1^k \in \text{ob}(\mathbf{T}_1)$ (which preserves empirical content).

Note that kinematical categorical equivalence is a strictly stronger criterion of theoretical equivalence than categorical equivalence:

Proposition 1. *Kinematical categorical equivalence entails categorical equivalence.*

Proof. This follows immediately from the definition of kinematical categorical equivalence, using the fact that the category of DPMs of a theory is a full subcategory of its category of KPMs. □

Proposition 2. *Categorical equivalence does not entail kinematical categorical equivalence.*

Proof. Recall the theories T_1 and T_2 from example 1. Let their associated kinematical theories T_1^k and T_2^k be as follows: $T_1^k = \{\exists!_{\sigma_1} x x = x\}$ and $T_2^k = \{\exists!_{\sigma_2} x x = x\}$. T_1 and T_2 are categorically equivalent. But they are not kinematically categorically equivalent: since $T_1 = T_1^k$ but T_2 is logically stronger than T_2^k, any functor $F : \mathbf{T}_2^k \to \mathbf{T}_1^k$ must take objects in $\text{ob}(\mathbf{T}_2^k)\backslash\text{ob}(\mathbf{T}_2)$ to objects in $\text{ob}(\mathbf{T}_1)$. □

I take it that this is the right result. Moreover, it is the right result for the right reason. Proposition 2 captures precisely our earlier intution that T_1 and T_2 are inequivalent theories because unlike T_1, T_2 says something non-trivial about the relationship between the predicate symbol p_0 and all the rest, which is reflected in the fact that there are mere KPMs of T_2 with only a single object in their domain in which this relationship fails to hold, which have no counterpart in T_1.

We also have, somewhat more interestingly, the following result:

Proposition 3. *Morita equivalence does not entail kinematical categorical equivalence.*

Proof. Consider the theories T_3 and T_4 defined as follows (I take this example from Barrett and Halvorson (2016)). T_3 is formulated in the signature Σ_3 which contains a single sort symbol σ_3 and two predicate symbols p and q, and has the following axioms: $\exists_\sigma x p(x)$, $\exists_\sigma x q(x)$, $\forall_\sigma x(p(x) \leftrightarrow \neg q(x))$. T_4 is the empty two-sorted theory in the signature Σ_4. Let $T_3^k = \{\}$ and $T_4^k = T_4$. T_3 and T_4 are Morita equivalent theories, but they are not kinematically categorically equivalent. Indeed, as with proposition 2, any functor $F : \mathbf{T}_3^k \to \mathbf{T}_4^k$ must take objects in $\mathrm{ob}(\mathbf{T}_3^k) \backslash \mathrm{ob}(\mathbf{T}_3)$ to objects in $\mathrm{ob}(\mathbf{T}_4)$. □

Now, here I want to say that kinematical categorical equivalence gives the right result, and Morita equivalence does not. We might reason as follows. According to the theory T_4, it is literally impossible for there not to be a partition of all things into two non-empty sets. If there is some domain on which no such partition exists, it does not even make sense to talk of T_4 applying to that domain, let alone being true or false of it. By contrast, precisely one of the ways in which T_3 can fail is if the predicates p and q fail to define a partition, or if the extension of p or q is empty. To put the point a different way, what T_4 dictates as a precondition for its applicability, T_3 dictates as a condition for its success.

Conversely, kinematical categorical equivalence can also make sense of the conditions under which it would be sensible to think of T_3 and T_4 as equivalent theories:

Proposition 4. *Again let T_3 and T_4 be as in proposition 3, but this time let $T_3^k = T_3$ and $T_4^k = T_4$. Then T_3 and T_4 are kinematically categorically equivalent.*

Proof. This follows immediately from theorem 5.1 of Barrett and Halvorson (2016) (using that T_3 and T_4 are Morita equivalent and that $T_3^k = T_3$ and $T_4^k = T_4$). □

In this case T_3 and T_4 both dictate the same conditions on their domains of applicability (i.e. that they can be partitioned into two non-empty sets), and both go on to say exactly the same things about those domains (i.e. precisely nothing). I take it that T_3 and T_4 are then uncontroversially equivalent theories.

A similar point goes for theories which are definitionally equivalent at the level of dynamics. For example, consider the theories T_5 and T_6 defined as follows: T_5 is formulated in the signature Σ_5 containing a single sort symbol σ and a single predicate symbol p, and has as its axioms the sentence $\exists!_\sigma x x = x$. T_6 is formulated in the signature Σ_6 containing a single sort symbol σ and a countable infinity of predicate symbols q_0, q_1, q_2 etc., and has as its axioms the sentences $\exists!_\sigma x x = x$, $\forall_\sigma x(q_0(x) \leftrightarrow q_i(x))$, $i = 1, 2, 3, ...$ T_5 and T_6 are definitionally equivalent theories (the relevant definitional extensions are $\delta_5 = \{\forall_\sigma x(q_i(x) \leftrightarrow p(x)), i = 0, 1, 2, 3, ...\}$ and $\delta_6 = \{\forall_\sigma x(p(x) \leftrightarrow q_0(x))\}$).

Proposition 5. *Let T_5 and T_6 be as above, and let $T_5^k = T_5$ and $T_6^k = \{\exists!_\sigma x x = x\}$. Then T_5 and T_6 are not kinematically categorically equivalent.*

Proof. This follows by the same argument used in the proofs of propositions 2 and 3. □

Proposition 6. *Again, let T_5 and T_6 be as above, but this time let $T_5^k = T_5$ and $T_6^k = T_6$. Then T_5 and T_6 are kinematically categorically equivalent.*

Proof. This follows immediately from theorems 5.1 of Barrett and Halvorson (2016) (using that definitional equivalence entails Morita equivalence and that $T_5^k = T_5$ and $T_6^k = T_6$). □

Again, there is a compelling intuition behind these results. In proposition 5, we have taken the sentences $\forall_\sigma x(q_0(x) \leftrightarrow q_i(x))$, $i = 1, 2, 3, ...$ to be dynamical laws. And in this case, it seems to me that there is good reason to consider T_5 and T_6 inequivalent: all the predicates in Σ_5 are guaranteed to be coextensive, whereas precisely one of the ways in which T_6 can fail is if some of the predicates in Σ_6 are not coextensive. Just as with proposition 3, what T_5 dictates as a precondition for its applicability, T_6 dictates as a condition for its success.

On the other hand, in proposition 6 we have taken the sentences $\forall_\sigma x(q_0(x) \leftrightarrow q_i(x))$, $i = 1, 2, 3, ...$ as kinematical constraints. Since they are kinematical constraints, we can think of the q_i as coextensive *by definition* in T_6. But then it is clear that T_6 does not really 'say anything more' than T_5: the q_i, $i = 1, 2, 3, ...$ are not conceptually or ontologically independent from q_0, but simply alternative (albeit redundant) bits of notation for the predicate q_0.

Without saying anything further about how to associate a kinematical theory with some first-order theory, it is difficult to say anything more about the precise relationship between kinematical categorical equivalence and other criteria such as Morita equivalence or definitional equivalence. But this should be seen as a feature, not a bug. Indeed, I think that precisely one of the advantages of allowing, as I have done, for a good deal of flexibility in how the kinematical constraints of a theory are chosen is that it invites us to consider carefully just what it would mean to take some condition or other as a kinematical constraint, which then informs our judgements on whether two theories are equivalent or not. In fact, this gives us the resources to discuss just what it would mean for the two theories T_1 and T_2 from example 1 to be equivalent. In some detail:

Recall our discussion of the kinematics-dynamics distinction in §3.1. There, I introduced three three types of constraints that should count as kinematical: relationships between objects which are not conceptually or ontologically independent of one another, definitions (of a single object), and domain restrictions. One way of getting at what it would mean for T_1 and T_2 to be theoretically equivalent is therefore to ask the following question: what would it mean to assimilate all the $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, ...$ under one of these headings—if, indeed, we can make sense of that at all?

Two of the options here can be ruled out fairly straightforwardly: the sentences $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, ...$ are not obviously thought of as expressing definitions (of a single object), nor domain restrictions. For the first, the sentences $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, ...$ are not to do with properties of a single object, but relationships between different objects. For the second, the sentences $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, ...$ do not tell us anything about the structure of the domains to which T_2 applies, e.g. their cardinality, whether they can be partitioned into two or more non-empty sets, whether they are partially ordered, whether they contain objects with certain properties etc.

This leaves us with the third option, which is to say: the sentences $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, ...$ are kinematical constraints because the predicate q_0 is not conceptually independent of the other q_i, and the sentences $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, ...$ tell us about the ways in which q_0 depends on, or is constructed out of, or defined from, the q_i. What kind of predicate could this be? The sentences $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, ...$ tell us that only when something is q_i for all $i \geq 1$ may we apply the predicate q_0 to that thing. So one way to think of it is that T_2 just introduces q_0 as a convenient shorthand which we may, but need not, use for all and only those things which are q_i for all $i \geq 1$. But then it seems to me as if there is good reason to consider T_1 and T_2 equivalent after all—all that T_2 does over and above T_1 is introduce some (rather redundant) notation—the predicate q_0—but after setting out how that notation is to be used says, like T_1, absolutely nothing. It also seems to me, in this case, that there is good reason to consider elementarily inequivalent pairs of models of T_2 which differ only as to whether some element of the domain is q_0 as equivalent in one's category of models for T_2: the predicate q_0 does not 'say' anything new, but (to reiterate), is just something which we may but need not use to describe just those things which are q_i for all $i \geq 1$. And then there does seem to be a sensible way to 'translate' a model of T_1 into (an equivalence class of) models of T_2, i.e. by taking p_i to q_{i+1} and *vice versa*.

Note that if this is right, then we have identified a sense in which Morita equivalence (and *a fortiori* definitional equivalence) is too strong a criterion of theoretical equivalence for theories. But the way in which we have got here is rather interesting. I began by asking: under what conditions are T_1 and T_2 kinematically categorically equivalent? We identified that as requiring that we take the sentences $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, ...$ as kinematical constraints. Then I asked: what would it mean to take the sentences $\forall_{\sigma_2} x(q_0(x) \rightarrow q_i(x))$, $i = 1, 2, 3, ...$ as kinematical constraints? And we saw that doing so induces an interpretation of the theory T_2 on which it does, after all, make sense to think of it as equivalent to T_1. In other words: getting clear on the kinematics-dynamics distinction induces a (partial) interpretation of a theory, and this interpretation gives us an intuitive handle on the way in which two theories may or may not be equivalent, which, moreover, coincides with the relationship of kinematical categorical equivalence.

The same story plays out for more realistic examples of physical theories as well. For example, I have argued (March 2024) that Maxwell gravitation and Newton-Cartan theory are equivalent at the level KPMs only if (the geometrised version of) Newton's second law (NII) is taken as a kinematical constraint. Roughly, this is because thinking of Maxwell gravitation and Newton-Cartan theory as equivalent requires us to interpret the Newton-Cartan connection as being defined in terms of (or constructed from, or reduced to) the behaviour of matter fields and the rotation standard. But if we begin by taking NII as a kinematical constraint then, I claim, this *already* induces an interpretation of Newton-Cartan theory on which the connection is defined in terms of (or constructed from, or reduced to) the behaviour of matter fields and the rotation standard. Put simply, if the irrotational degrees of freedom of the connection cannot come apart from the behaviour of matter fields, even in principle, then we shouldn't think of them as representing an ontologically and conceptually independent piece of structure in the theory. All of which is to say: paying attention to the kinematics-dynamics distinction helps to illuminate and resolve conceptual puzzles to do with relationships of equivalence and inequivalence between theories, not just for the toy examples from first-order logic which we have been considering, but also for *bona fide* physical theories which are of interest in foundational discussions.

4 Further implications

I will now move on to consider some philosophical payoffs of this focus on the relationship between kinematics and dynamics in discussions of theoretical equivalence. The first has to do with a puzzle which I draw from Weatherall (2020), though he does not quite state it in these terms. The puzzle is this: why is it that some categories seem to suitably capture the structure of the theories they represent, so that relationships of categorical equivalence between these theories (however superficially disparate) are taken to reveal an important sense in which these theories are equivalent, whereas other categories of models do not seem to capture the internal structure of the theories they represent in this way? In particular, the former often seems to be the case for mathematical theories, but is much less obvious for physical theories. This is especially the case given that the categories of models associated to physical theories in foundational discussions are generally groupoids—i.e. categories in which every arrow has an inverse—and there is good reason to think that this kind of structure does not capture all the salient relationships between models of a theory, e.g. the sense in which one model might be embeddable into another but not *vice versa*. Weatherall goes on to consider, and—in my view correctly—dismiss, one *prima facie* plausible way of making sense of this difference, in terms of a property he calls the 'G' property, which says that every autoequivalence of the category is naturally isomorphic to the identity.

Now, I do not think—nor do I want to suggest—that paying proper attention to the full space of KPMs of theories is the full solution to this puzzle. (Indeed, I am inclined to think that the aforementioned worry about the categories of models associated with theories being groupoids suggests that this cannot be the full solution.) But I do think that it is an important part of the solution. In particular, I want to suggest that one of the interesting ways in which the categories of models associated with theories where categorical equivalence of DPMs does seem to give the right results, and theories where it does not, differ, is that in the former case, there always seems to be a relatively natural choice for the space of KPMs of the theory to hand, in a way that the same sort of relationships which play out at the level of DPMs are mirrored at the level of KPMs as well. By contrast, in cases where categorical equivalence of DPMs seems to give the 'wrong' results, it tends to be (if not invariably is) the case that a lot of information about the structure of that theory is encoded in its space of (mere) KPMs. This point is best illustrated by examples; I will consider several here.

First, begin with mathematical theories. As noted above, relationships of categorical equivalence between theories in mathematics are often (if not always) taken to reveal important senses in which they can be thought of as equivalent. I want to suggest that part of the reason for this is that there is good reason to think that the categories of DPMs and KPMs of mathematical theories generally coincide, and so is a special case of the way in which the structure of a theory's category of DPMs can 'mirror' the structure of its category of KPMs—i.e. when the two categories are identical.

To see this in a bit more detail, I want to return to an idea which was raised in §3.1. There, I pointed out that the axioms of mathematical theories serve to define the kind of mathematical structures to which the theory applies, and therefore should be thought of as kinematical constraints. In particular, it is difficult to see what it would mean for a mathematical structure to be aptly described by e.g. the theory of vector spaces (and so fall into its space of KPMs) but nevertheless fail to be a vector space (so that it does not fall into the space of DPMs of the theory). Alternatively, we can see this in the idea that structures which fail to satisfy the axioms of some mathematical theory are said to fall outside the domain of that theory, rather than it being the case that the mathematical theory in question is false of those structures. This stands in contrast to the case of physical theories: we can (perhaps even must) say what it would mean for a physical theory to be applicable to some system or other, but nevertheless be false of that system.[6]

Identity of the KPMs and DPMs of a theory is one way in which the category of DPMs can 'mirror' the structure of its category of KPMs. I have suggested that this might be part of what is going on in examples of categorical equivalence between mathematical theories. But identity of the space of KPMs and DPMs of a theory is not the only way in which this can happen. To illustrate some other ways in which this can work, I now want to consider three examples of relationships between physical theories where I take it that the categorical equivalence of DPMs programme has been illuminating. These are Galilean gravitation and Newton-Cartan theory, Faraday tensor and gauge potential electromagnetism (Weatherall 2016), and Lagrangian and Hamiltonian mechanics (Barrett 2019).[7]

Beginning with Weatherall's (2016) discussion of Galilean gravitation and Newton-Cartan theory, the relationship between models of these theories is due to the Trautman geometrisation and recovery theorems (see e.g. Malament (2012)). These say that given a model $\langle M, t_a, h^{ab}, \nabla, \phi, T^{ab} \rangle$ of Galilean gravitation (formulated using a flat connection and gravitational potential), one

[6] Or, to put the point in slightly more Popperian terms: physical theories dictate not only truth conditions but also falsity conditions on the domains to which they apply.

[7] For reasons of space, I will not consider the discussion of general relativity and Einstein algebras in Rosenstock, Barrett, and Weatherall (2015) here, though I suspect that a similar point goes for these theories as well. It would be worthwhile to examine this in detail.

can define a unique model $\langle M, t_a, h^{ab}, \tilde{\nabla}, T^{ab}\rangle$ of Newton-Cartan theory,[8] and conversely, from a model $\langle M, t_a, h^{ab}, \tilde{\nabla}, T^{ab}\rangle$ of Newton-Cartan theory, one can recover a Trautman gauge orbit of models of Galilean gravitation, related by transformations of the form $\nabla \to (\nabla, t_b t_c \nabla^a \psi)$, $\phi \to \phi + \psi$, $\nabla^a \nabla^b \psi = 0$.

One of the helpful contributions of Teh (2018) is that it establishes just what is necessary for a 'recovery theorem' *à la* Trautman, without appealing to the dynamics of the theory. In particular, sufficient for there to be a correspondence between Newton-Cartan connections and pairs $\langle \nabla, \phi \rangle$ consisting of flat connections and scalar fields, up to transformations of the form $\nabla \to (\nabla, t_b t_c \nabla^a \psi)$, $\phi \to \phi + \psi$, $\nabla^a \nabla^b \psi = 0$ are the two homogeneous Trautman conditions $R^a{}_b{}^c{}_d = R^c{}_d{}^a{}_b$ and $R^{ab}{}_{cd} = 0$. So providing these two conditions are taken as kinematical constraints, the same relationship between DPMs of Newton-Cartan theory and Trautman orbits of Galilean gravitation will play out at the level of KPMs as well, regardless of what the matter fields are doing.

A similar point goes for Faraday tensor and gauge potential electromagnetism. Here, the relationship between DPMs of the two theories is due to Poincaré's lemma, i.e. the fact that the Faraday two-form is closed implies that it is (at least locally) exact, and so can be identified with the exterior (covariant) derivative of the electromagnetic one-form, defined up to exact one-form shifts. So if(f) the Gauss-Faraday law ($\mathrm{d}_a F_{bc} = 0$) is a kinematical constraint, the relationship between the DPMs of these theories will be reflected at the level of KPMs as well.

Our third example—Barrett's (2019) discussion of the equivalence of hyperregular Lagrangian mechanics (on a tangent bundle) and hyperregular Hamiltonian mechanics (on a cotangent bundle)—is even more straightforward, because the way that Barrett sets up the categories of models of the two theories makes it clear that his result is essentially blind to the distinction between kinematically and dynamically possible models of these theories. Barrett defines models of Lagrangian mechanics as pairs $\langle T^* M, L\rangle$ and models of Hamiltonian mechanics as pairs $\langle T_* M, H\rangle$. And from this one can indeed define the DPMs of Lagrangian mechanics and Hamiltonian mechanics: in DPMs of Lagrangian mechanics, the configuration space trajectory of the system is a base integral curve of the Lagrangian vector field $(X_L)^a$, whilst in DPMs of Hamiltonian mechanics, the configuration space trajectory of the system is a base integral curve of the Hamiltonian vector field $(X_H)^a$.

This also makes it immediately clear that neither of the structures $\langle T^* M, L\rangle$ or $\langle T_* M, H\rangle$ carry enough information to distinguish DPMs of the theory from mere KPMs of the theory. In order to tell whether some physical history is a DPM of Lagrangian or Hamiltonian mechanics, we need to introduce some fur-

[8] Where $\tilde{\nabla} = (\nabla, -t_b t_c \nabla^a \phi)$. The notation here follows Malament (2012, proposition 1.7.3): $\nabla' = (\nabla, C^a{}_{bc})$ iff for all smooth tensor fields $\alpha^{a_1 \dots a_r}{}_{b_1 \dots b_s}$ on M,

$$\begin{aligned}(\nabla'_n - \nabla_n)\alpha^{a_1 \dots a_r}{}_{b_1 \dots b_s} = {}& \alpha^{a_1 \dots a_r}{}_{m b_2 \dots b_s} C^m{}_{n b_1} + \dots + \alpha^{a_1 \dots a_r}{}_{b_1 \dots b_{s-1} m} C^m{}_{n b_s} \\ & - \alpha^{m a_2 \dots a_r}{}_{b_1 \dots b_s} C^{a_1}{}_{nm} - \dots - \alpha^{a_1 \dots a_{r-1} m}{}_{b_1 \dots b_s} C^{a_r}{}_{nm}.\end{aligned}$$

ther structure, i.e. a curve γ in M which represents the configuration space trajectory of the system. Then we can say: $\langle T^*M, L, \gamma\rangle$ (respectively $\langle T_*M, H, \gamma\rangle$) is a model of Lagrangian (Hamiltonian) mechanics iff γ is a base integral curve of $(X_L)^a$ (respectively $(X_H)^a$). But curves in the base manifold are preserved when passing (via the Legendre transform) from the tangent bundle to the cotangent bundle and *vice versa*, so Barrett's result applies whether one considers the full space of KPMs or restricts attention to the space of DPMs.

In all three of these cases, the underlying point is that categorical equivalence of DPMs is a good criterion of theoretical equivalence because there is a natural choice for the space of KPMs available such that the mere KPMs of the two theories do not 'tell us' anything new about the structure of the theory's models, or the relationship between them. Contrast this with an example where categorical equivalence of DPMs seems to give the wrong verdict, which I draw from Weatherall (2020). The theory "directions" says 'the cardinal directions form a two-dimensional vector space (over the reals), with 'north' and 'east' physically distinguished.' The theory "baubles" says 'there are two shiny things, one of which is red and one of which is green.' Weatherall associates to these theories two categories **Di** and **Bau** respectively; **Di** has as its objects two-dimensional vector spaces with preferred ordered basis, and as its arrows linear bijections which preserve that ordered basis, and **Bau** has as its objects ordered pairs of distinct elements, and as its arrows bijections which preserve order. These categories are equivalent.

Now, consider the theory baubles. One natural way of defining the space of KPMs of baubles is to take the kinematical constraints of baubles to say 'there are some (perhaps finitely many) shiny things, which are partitioned into two not necessarily non-empty sets—those which are red and those which are green.' Models of this theory correspond to ordered pairs of disjoint sets, at most one of which can be empty. DPMs of baubles then correspond to the special case where both of these sets are singletons. Analogously to before, we can then take arrows in our category $\mathbf{Bau}_k$ to be bijections which preserve order.

On the other hand, consider the theory directions. One natural way of defining the space of KPMs of directions is to relax the assumption that north and east are physically distinguished: any pair of (orthogonal) cardinal directions would do equally well (in physical terms, this amounts to the point that the earth's magnetic dipole moment could (at any given moment in time) have been located at any angle with respect to (e.g.) its axis of rotation). So KPMs of directions are two-dimensional vector spaces with any preferred ordered basis, and arrows in the category $\mathbf{Di}_k$ are linear bijections which preserve this ordered basis.

This already makes it clear that $\mathbf{Di}_k$ and $\mathbf{Bau}_k$ are inequivalent categories. (Why? Because objects in $\mathbf{Bau}_k$ will in general have non-trivial automorphisms induced by arbitrary permutations of the elements of either of the sets in the ordered pairs, whereas the only automorphisms of objects in $\mathbf{Di}_k$ are identity morphisms.) It also makes it clear that neither of these categories will be equivalent to the other examples Weatherall considers—the category **1**, which has a single object and an identity morphism, or the category **Sing**, whose objects are singleton sets and whose morphisms are functions preserving that element, both of whose KPMs and DPMs coincide.[9] But in both these cases, this is because the (mere) KPMs encode important information about the structure of the theory which goes missing when we restrict attention to the category of DPMs. In the case of baubles, the mere KPMs tell us that the theory only distinguishes things by their being red or green, which is encoded in the fact that the mere KPMs of the theory have non-trivial automorphisms. In the case of directions, the mere KPMs tell us that the objects in **Di** have infinitely many elements, which is encoded in the fact that there are infinitely many distinct choices of preferred ordered basis for the vector space.

This brings us on to a second proposal Weatherall discusses, which he calls 'Rosenstock's heuristic,' and which is closely tied up with a number of other discussions of purely formal criteria of theoretical equivalence (in particular, Coffey (2014), Butterfield (2021), and Teitel (2021)). Roughly, this kind of view says that categorical equivalence of DPMs might be necessary, but not sufficient for theoretical equivalence. Theoretical equivalence, on the other hand, would require a more subtle and thoroughgoing analysis of the relationship between the two theories, including perhaps the way in which they are empirically equivalent (to borrow a point made by Weatherall (2020)), or the proposed interpretations of the theories (to borrow an idea from Coffey (2014)).

Now, I do not want to suggest that focussing on the broader space of KPMs of a theory should always be expected to act as a surrogate for this kind of detailed, context-dependent analysis. In particular, I do not think that it can supplant considerations of the way in which two theories are empirically equivalent. On the other hand, I do think that that it goes a large part of the way towards addressing concerns about the role of judgements of interpretation in judgements of theoretical equivalence. A full discussion of this point is beyond the scope of this article, but in brief:

One of the insights of Coffey's (2014) discussion of theoretical equivalence is to raise an important challenge for formal criteria of theoretical equivalence: if we interpret theories realistically, then we should expect questions of theoretical equivalence to supervene (in the end) on what interpretations the theories have, so how could any purely formal criterion of theoretical equivalence be adequate? Part of this challenge was answered by Dewar (2023, 2022), who points out that which arrows one includes in ones category of DPMs of a theory often

[9] This is also the case for plausible ways of associating a category of KPMs to **Sing**. Consider the theory "singleton" which says: 'there is exactly one thing.' We can take **Sing** to be the category of DPMs of the theory singleton. Now, the obvious option for the category of KPMs $\mathbf{Sing}_k$ is to take the kinematical constraints of singleton to say: 'there is at least one thing.' We can take the objects in $\mathbf{Sing}_k$ to be non-empty sets, and its arrows to be bijective total functions. To see that $\mathbf{Sing}_k$ will be kinematically categorically inequivalent to $\mathbf{Di}_k$, we just need to note that, as with $\mathbf{Bau}_k$ there are objects in $\mathbf{Sing}_k$ with non-trivial automorphisms. For $\mathbf{Bau}_k$, consider those objects of $\mathbf{Bau}_k$ which are ordered pairs consisting of a singleton and the empty set. These are not objects of **Bau**. These objects have only a single automorphism (i.e. the identity), so any equivalence functor $F : \mathbf{Bau}_k \to \mathbf{Sing}_k$ must map these objects to objects in $\mathbf{Sing}_k$ with only a single automorphism. But the only objects in $\mathbf{Sing}_k$ with only a single automorphism are objects in **Sing**, so the two theories will not be kinematically categorically equivalent.

encodes important facts about the interpretation of that theory (whether e.g. the standard of rest in full Newtonian gravitation represents something physical or whether it is mere 'descriptive fluff').

Taking into account the full space of KPMs of a theory provides another part of the solution to Coffey's puzzle. Here, I will illustrate how KPMs can be used to articulate interpretative disputes about fundamentality and explanation. For this, consider again the Faraday tensor and gauge potential formulations of electromagnetism. It has often been noted in the literature on these two theories that there is a sense in which the gauge potential formulation of electromagentism 'explains more' than the Faraday tensor formulation—in the latter theory, the Gauss-Faraday law has to be taken as a bare postulate, whereas in the gauge potential formulation, this is a mathematical theorem—and that this is somewhat in tension with the idea that these two theories are equivalent (see, e.g. Dewar (2016) and Jacobs (2021)). But kinematical categorical equivalence gives us the resources to make sense of this. Only if the Gauss-Faraday law is a kinematical constraint will the two theories be kinematically categorically equivalent. But if the Gauss-Faraday law is a kinematical constraint, then there is an obvious sense in which the Faraday tensor formulation is just as committed to the electromagnetic one-form as the gauge potential formulation. For example, we might say that the Gauss-Faraday law is a kinematical constraint because it is part of the definition of F_{ab} that it is closed. Via Poincaré's lemma, this becomes the point that it is part of the definition of F_{ab} that (at least locally) there exists a one-form, defined up to exact one-form shifts, such that $F_{ab} = \mathrm{d}_a A_b$. But if this is how F_{ab} is defined, in what sense is the Faraday tensor formulation not committed to the electromagnetic one-form? And if the Faraday tensor formulation is just as committed to the electromagnetic one-form as the gauge potential formulation, then it doesn't come with any loss of explanatory power. Conversely, if the Gauss-Faraday law is a dynamical constraint, then there is a clear sense in which the Faraday tensor formulation is not committed to the electromagnetic one-form, since there will be mere KPMs of the theory in which F_{ab} is not the exterior derivative of any one-form. In this case, the Faraday tensor formulation does come with a loss of explanatory power, but then it will also be inequivalent to the gauge potential formulation at the level of KPMs.

Finally, I want to consider one of the morals which Barrett and Halvorson (2022) draw from their discussion of the theories T_1 and T_2 from example 1, which has to do with what they call the Cantor-Bernstein and co-Cantor-Bernstein properties of theories. On the one hand, the Cantor-Bernstein property says that if T is embeddable into T' and T' is embeddable into T then T and T' are equivalent theories. This is supposed to be captured by the existence of mutually conservative translations between T and T'. On the other hand, the co-Cantor-Bernstein property says that if T posits all the structure of T' and T' posits all the structure of T then T and T' are equivalent theories. This is supposed to be captured by the existence of mutually essentially surjective between T and T'. Since there are mutually conservative and mutually essentially surjective translations between the theories T_1 and T_2 from example 1, Barrett and Halvorson conclude that theories lack the Cantor-Bernstein and co-Cantor-Bernstein properties.

Now, I am in agreement with Barrett and Halvorson that theories lack the Cantor-Bernstein and co-Cantor-Bernstein properties when considered at the level of DPMs. However, I want to point out that Barrett and Halvorson have not yet said anything to suggest that theories lack the Cantor-Bernstein or co-Cantor-Bernstein properties at the level of KPMs, particularly given my discussion of the circumstances under which T_1 and T_2 would be kinematically categorically equivalent theories. Of course, it might be the case that there are examples for which the Cantor-Bernstein or co-Cantor-Bernstein properties fail at the level of KPMs as well. But the very least, more work is needed to establish this.

5 Close

In this article, I have aimed to show how the kind of examples from first-order logic which Halvorson and co-authors take to show that categorical equivalence is too weak a criterion of theoretical equivalence can be used to motivate extending the categorical equivalence programme to the full space of KPMs of a theory. I have also discussed how kinematical categorical equivalence can be used to illuminate a number of other puzzles to do with relationships of equivalence and inequivalence between theories.

I do not hope to have convinced the reader that kinematical categorical equivalence is the be all and end all of theoretical equivalence, nor that it is the only proposal on the table for dealing with these kind of cases. But I do hope to have shown that kinematical categorical equivalence offers interesting and compelling insights into a number of toy and realistic examples: both ones which have seemed to be problem cases for the categorical equivalence programme and ones which have seemed to be success stories for the approach. Insofar as one agrees that these insights are valuable, I take myself to have established that kinematical categorical equivalence is a serious contender in the space of criteria of theoretical equivalence—or at least, one that merits exploring further.

Undoubtedly, there is much more to be said on the issues raised here, which stretch far beyond the scope of this article. To name just a few examples:

- I have pointed out that getting clear on the kinematics-dynamics distinction of a theory often seems to induce a (partial) interpretation of that theory. It would be of interest to discuss exactly what kind of interpretative disputes can be cashed out as disputes about the kinematics-dynamics distinction (I would expect these to include, *inter alia*, debates about fundamentality, explanation, and ontology, but there are probably others).[10]

- If I am right that paying attention to the full space of KPMs is the way to go in discussions of theoretical equivalence, one might also wonder about the role that KPMs have to play in other inter-theoretic relationships of

[10] For example, Adam Caulton has pointed out to me that the dynamical-geometrical debate could be cashed out in this way.

interest, such as reduction or instantiation.[11] For example, whether GR reduces Newton-Cartan theory at the level of KPMs as well as DPMs will depend non-trivially on which of the dynamical equations of Newton-Cartan theory are also kinematical constraints.

- Since kinematical categorical equivalence is a finer-grained criterion than categorical equivalence, different choices for the kinematical constraints of a theory will in general induce different results on whether or not they are equivalent. It would be worthwhile to conduct some detailed case studies on the way in which this plays out for theories where the categorical equivalence of DPMs programme has been illuminating.

Acknowedgements

Many thanks to Adam Caulton, Zack Glindon, Helene Lösl, and James Read for helpful comments and discussions. I am grateful to Balliol College, Oxford, and the Faculty of Philosophy, University of Oxford, for financial support.

References

Barrett, Thomas William. 2019. "Equivalent and inequivalent formulations of classical mechanics." *British Journal for the Philosophy of Science* 70 (4): 1167–1199.

Barrett, Thomas William, and Hans Halvorson. 2016. "Morita Equivalence." *The review of symbolic logic* 9 (3): 556–582.

Barrett, Thomas William, and Hans Halvorson. 2022. "Mutual translatability, equivalence, and the structure of theories." *Synthese* 200 (3).

Butterfield, Jeremy. 2021. "On Dualities and Equivalences between Physical Theories." In *Philosophy Beyond Spacetime.* Oxford: Oxford University Press.

Coffey, Kevin. 2014. "Theoretical Equivalence as Interpretative Equivalence." *British Journal for the Philosophy of Science* 65 (4): 821–844. https://doi.org/10.1093/bjps/axt034.

Curiel, Erik. 2014. "Classical mechanics is Lagrangian; it is not Hamiltonian." *British Journal for the Philosophy of Science* 65 (2): 269–321.

Curiel, Erik. 2016. "Kinematics, Dynamics, and the Structure of Physical Theory," arXiv: 1603.02999 [physics.hist-ph].

Dewar, Neil. 2016. "Symmetries in physics, metaphysics, and logic." DPhil thesis, University of Oxford.

[11] I am indebted to Adam Caulton for articulating this idea to me.

Dewar, Neil. 2022. “Structure and Equivalence.” Cambridge: Cambridge University Press.

Dewar, Neil. 2023. “Interpretation and equivalence; or, equivalence and interpretation.” *Synthese* 201 (4).

Halvorson, Hans. 2012. “What scientific theories could not be.” *Philosophy of science* 79 (2): 183–206.

Jacobs, Caspar. 2021. “Symmetries as a Guide to the Structure of Physical Quantities.” DPhil thesis, University of Oxford.

Luc, Joanna, and Tomasz Placek. 2020. “Interpreting Non-Hausdorff (Generalized) Manifolds in General Relativity.” *Philosophy of Science* 87 (1): 21–42. https://doi.org/10.1086/706116.

Mac Lane, Saunders. 1998. *Categories for the working mathematician.* Second edition. Graduate texts in mathematics; volume 5. New York: Springer.

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

March, Eleanor. 2024. “Are Maxwell gravitation and Newton-Cartan theory theoretically equivalent?,” http://philsci-archive.pitt.edu/22922/.

Nguyen, James, Nicholas Teh, and Laura Wells. 2020. “Why surplus structure is not superfluous.” *British Journal for the Philosophy of Science* 71 (2): 665–695.

North, Jill. 2009. “The “Structure” of physics: a case study.” *Journal of philosophy* 106 (2): 57–88.

Rosenstock, Sarita, Thomas William Barrett, and James Owen Weatherall. 2015. “On Einstein algebras and relativistic spacetimes.” *Studies in the History and Philosophy of Modern Physics* 52:309–316.

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

Teh, Nicholas. 2018. “Recovering recovery: on the relationship between gauge symmetry and Trautman recovery.” *Philosophy of Science* 85 (2): 201–224.

Teitel, Trevor. 2021. “What theoretical equivalence could not be.” *Philosophical studies* 178 (12): 4119–4149.

Tsementzis, Dimitris, and Hans Halvorson. 2017. “Categories of Scientific Theories.” In *Categories for the Working Philosopher.* Oxford: Oxford University Press.

Weatherall, James Owen. 2016. “Are Newtonian gravity and geometrised Newtonian gravity theoretically equivalent?” *Erkenntnis* 81 (5): 1073–1091.

Weatherall, James Owen. 2017. “Categories and the foundations of classical space-time theories.” In *Categories for the Working Philosopher,* edited by Elaine Laundry, 329–348. OUP.

Weatherall, James Owen. 2020. “Why not categorical equivalence?,” arXiv: 1812.00943 `[physics.hist-ph]`.