Spacetime and Physical Equivalence

Sebastian De Haro

Trinity College, Cambridge, CB2 1TQ, United Kingdom
Department of History and Philosophy of Science, University of Cambridge
Free School Lane, Cambridge CB2 3RH, United Kingdom

sd696@cam.ac.uk

Abstract

In this paper I begin to lay out a conceptual scheme for: (i) analysing dualities as cases of theoretical equivalence; (ii) assessing when cases of theoretical equivalence are also cases of physical equivalence. The scheme is applied to gauge/gravity dualities. I expound what I argue to be the contribution of gauge/gravity dualities to questions about: (iii) the nature of spacetime in quantum gravity; (iv) broader philosophical and phyiscal discussions of gauge/gravity dualities.

(i)-(ii) proceed by analysing duality through four contrasts. A *duality* will be a suitable isomorphism between models: and the four relevant contrasts are as follows:

(a) *Bare theory*: a triple of states, quantities, and dynamics endowed with appropriate structures and symmetries; *vs. interpreted theory*: which is endowed with, in addition, a suitable pair of interpretative surjections from the triples to a domain in the world.

(b) *Extendable vs. unextendable theories*: which can, respectively cannot, be extended as regards their domains of applicability.

(c) *External vs. internal intepretations*: which are constructed by coupling the theory to another interpreted theory, respectively from within the theory itself.

(d) *Theoretical vs. physical equivalence*: which distinguish formal equivalence from the equivalence of fully interpreted theories.

I also discuss three meshing conditions: between symmetries and duality, between symmetries and interpretation, and between duality and interpretation. The meshing conditions lead to a characterisation of symmetries as *redundant* or as *physical*, and to a characterisation of physical equivalence as a commutativity condition between two maps.

I will apply the above scheme to answering questions (iii)-(iv) for gauge/gravity dualities. I will argue that the things that are physically relevant are those that stand in a bijective correspondence under gauge/gravity duality: the *common core* of the two models. I therefore conclude that most of the mathematical and physical structures that we are familiar with in these models (the dimension of spacetime, tensor fields, Lie groups, and the classical-quantum distinction) are largely, though crucially never entirely, *not* part of that common core. Thus, the interpretation of dualities for theories of quantum gravity compels us to rethink the roles that spacetime, and many other tools in theoretical physics, play in theories of spacetime.

Contents

Introduction

Dualities raise interesting philosophical questions regarding the physical equivalence of theories, the relation between duality and symmetry, and the nature and properties of the 'common core' shared by two dual theories. Philosophers of physics have started to address the philosophical significance of dualities in recent years.[1]

In this paper I begin to develop a scheme for answering the question of physical equivalence and the relation between duality and symmetry. I will apply the scheme to the case of gauge/gravity dualities, thereby exhibiting part of the common core which is shared by the two theories, with implications for the nature of spacetime as described by these theories, and for the content which two dual theories take to be 'physical'.

The paper has two parts. In the first part of the paper (Section 1), I will lay out the conceptual scheme which I begin to develop here, and which will be used to answer the questions of: (i) theoretical equivalence, (ii) physical equivalence. In the second part of the paper (Sections 2-4), I will discuss the questions: (iii) the nature of spacetime in theories of quantum gravity, and (iv) the broader philosophical and physical implications, for the case of gauge/gravity dualities.

My argument in the first part of the paper will proceed by analysing dualities in terms of four contrasts, as follows.

A duality is an isomorphism between theories (more specifically: between *bare* theories, see (a) immediately below). Then the four contrasts are:

(a) *Bare theory vs. interpreted theory:* A *bare theory* is a triple of states, quantities, and dynamics, each of which are construed as structured sets, invariant under appropriate symmetries. An *interpreted theory* has, in addition, a pair of surjections to physical quantities (at the very least: but an interpretation will typically include more): which are what we call the *interpretation*.

(b) *Extendable vs. unextendable theories*: theories which do, respectively do not, admit suitable extensions in their domains of applicability. I will also allow for a weaker conception of 'unextendable theory', according to which unextendable theories may admit an extension via e.g. *couplings* to other theories in their domain, but such that their interpretations are *robust*, i.e. unchanged under such extensions.

(c) *External vs. internal intepretations*:[2] interpretations which are obtained from outside (i.e. by coupling the theory to a second theory which has already been interpreted), respectively from inside, the theory. Here, "inside the theory" means that the interpretation stems from the structure and symmetries of the bare theories, i.e. the triples, themselves: from the role which states, quantities, and dynamics have within the theoretical structure.

(d) *Theoretical vs. physical equivalence*: formal equivalence (i.e. agreement of the bare theories, but with possible disagreement of the interpretations) vs. full equivalence of the interpreted theories: i.e. agreement of both the bare theory and the interpretive maps.

These contrasts build on each other: so that (a) is used in the analysis of (b); (a) and

[1] See e.g. Rickles (2012), Matsubara (2013), Dieks et al. (2015), De Haro (2016, 2016a), De Haro et al. (2016, 2016a, 2016b), Huggett (2016), Fraser (2016), McKenzie (2016), Rickles (2016), Castellani (2016), Teh et al. (2016). For a comparison with these works, see Section 4.

[2] See Dieks et al. (2016: §3.3.2).

(b) are jointly used in the interpretative analysis of (c); and (a)-(c) are all needed in order to reach a verdict distinguishing theoretical vs. physical equivalence, as (d) intends.

The scheme (a)-(d) contains three related maps, viz. symmetry transformations, interpretative map, and duality. Since these maps have the theory concerned as a relevant domain or codomain, there are three meshing conditions that we need to consider for the corresponding notions, viz. between: (1) duality and symmetry, (2) symmetry and interpretation, (3) duality and interpretation. I express these meshing conditions as appropriate commutativity conditions for the respective maps:

(1: SymT) A symmetry transformation of the theory must commute with the duality map between the models (§1.1.1 and §1.2).

There is a second kind of meshing condition which holds for symmetries of the models which are not symmetries of the theory, and which is more restrictive:

(1': PSymM) A proper symmetry of the model (one that is not a symmetry of the theory) is represented trivially on the duality map, i.e. $d \circ \sigma = d$, where d is the duality map, and σ is the symmetry transformation of the model (§1.1.1).

(2: Physical) A symmetry transformation of the theory commutes with the internal interpretation of the theory. This makes the corresponding symmetry transformation *physical*, on the internal interpretation (§1.1.2.b).

There is a similar kind of strengthening of this condition to the one in (1'), when the symmetries considered are proper symmetries of the models:

(2': Redundant) A proper symmetry σ of the model is represented trivially on the internal interpretation map, i.e. $I_T \circ \sigma = I_T$, where I_T is the internal interpretation map. This makes the corresponding symmetry transformation *redundant*, on the internal interpretation (§1.1.2.b).

(3: PhysEquiv) The internal interpretation of a theory commutes with the duality map between its models. Given appropriate conditions for the theory, this makes the two models *physically equivalent* (§1.3.1).

My account will provide sufficient details so that the scheme introduced in this paper can be readily applied to other cases, and I will give several examples that will work toward applying the scheme to gauge/gravity dualities. However, a *full* account of theoretical and physical equivalence, doing full justice to the intricacies of the matter, will have to be left for the future.

Of course, not all of the above notions are completely new. But my construal of them is largely novel (the only exception being the contrast (c), for which I am in full agreement with, and just develop further, the position of Dieks et al. (2015) and De Haro (2016)). In particular, the way I here articulate the notions of theoretical and physical equivalence in terms of the contrasts (a)-(d), so that I can successfully analyse dualities, and the way I here discuss the meshing conditions between duality, symmetry, and interpretation in (1)-(3), are novel and are intended to add to the literature on both dualities and equivalence of theories.

In the second part of the paper (Sections 2-4) I will apply the scheme (a) to (d), with the meshing conditions (1)-(3), to gauge/gravity dualities in more detail, so as to clarify the contribution of gauge/gravity dualities to answering the questions of: (iii) the nature of spacetime in quantum gravity, (iv) the broader philosophical and physical implications, including the question of the content which the theories regard as physical. Gauge/gravity

duality is one particular approach to quantum gravity.[3] It was developed in the context of string and M theory, but it has broader ramifications: e.g. applications to condensed matter physics and heavy-ion collisions (Ammon and Erdmenger (2015: III)). Briefly, gauge/gravity duality is an equivalence between:

(i: Gravity) on the one hand: a theory of quantum gravity in a *volume* bounded inside a certain surface.

(ii: QFT) On the other: a quantum field theory defined on that *surface*, which is usually, in most models, at 'spatial infinity', relative to the volume.

Such a relation between a $(d+1)$-dimensional theory and its d-dimensional image is also called 'holographic'. I will argue that, despite the apparent innocence of the references to spacetime appearing in the previous sentence summarising the duality: the physical interpretation of this duality calls for a revision of the role that most of our physical and mathematical concepts are supposed to play in a fundamental theory—most notably, the role of spacetime. And the interpretation of the duality itself requires us to carefully reconsider the philosophical concepts of theoretical equivalence and physical equivalence, as in the first part of the paper.

The overall plan of the paper is thus as follows. Section 1 introduces the notions of theory (§1.1), duality (§1.2), theoretical equivalence, and physical equivalence (§1.3); and I compare with Glymour's notion of equivalence (§1.4). In Section 2, I analyse gauge/gravity duality as a theory with two models, whose 'common core' is what the theory says (§2.1): and I discuss whether the content of this common core includes spacetime (§2.2). In Section 3, I discuss the broader implications of duality for: (i) theory construction (§3.1); (ii) metaphysics (§3.2). In Section 4, I discuss the relation of my scheme to recent work on dualities. The final Section concludes.

1 Theories, duality, and physical equivalence

In §1.1, I will specify more exactly what I mean by 'theory' and related notions, especially a 'quantity' and an 'interpretation'. In §1.2, I will give my conception of a duality between such theories. §1.3 describes how, for theories 'of the whole world', duality is tantamount to physical equivalence, i.e. the theories at issue being really the same theory. In §1.4, I compare my account with Glymour's notion of equivalence.

1.1 The conception of a theory

Before we are ready to engage with the interpretation of dualities (which we will do in §1.2-§1.3), there is some work to do: we need to have conceptions of theory which are sufficiently articulated that they make it possible to build an analysis of physical equivalence on them. I first introduce, in §1.1.1, the notion of a *bare theory*. Then, in §1.1.2, I discuss what an interpretation is, and thus introduce the notion of *interpreted theory*.

[3] For an expository overview, see e.g. Ammon and Erdmenger (2015). For a conceptual review, see De Haro et al. (2016b).

1.1.1 Bare theory

I take a *bare theory* to be a triple $T = \langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$ consisting of: (i) a set $\mathcal{H}$ of states, endowed with appropriate structure; (ii) a set of physical quantities $\mathcal{Q}$, endowed with appropriate structure; (iii) a dynamics $\mathcal{D}$. Such a triple will generally also be endowed with *symmetries*, which are automorphisms $s : \mathcal{H} \to \mathcal{H}$ preserving (a subset of) the valuations of the physical quantities on the states (for details, see De Haro et al. (2016: §3.3)), and which commute with (are suitably equivariant for) the dynamics $\mathcal{D}$.

For a *quantum* theory, which will be our main focus, we will take $\mathcal{H}$ to be a Hilbert space; $\mathcal{Q}$ will be a specific subset of operators on the Hilbert space[4]; and $\mathcal{D}$ can be taken to be a choice of a unique (perhaps up to addition by a constant) Hamiltonian operator from the set $\mathcal{Q}$ of physical quantities.[5] In a quantum theory, the appropriate structures are matrix elements of operators evaluated on states; and the symmetries are represented by unitary operators.

A theory may contain many more quantities but it is only after we have singled out the ones which have a physical significance that we have a *physical*, rather than a *mathematical*, theory or model. The quantities $\mathcal{Q}$, the states $\mathcal{H}$, and the dynamics $\mathcal{D}$ have a physical significance at a possible world W, and within it a domain D, though it has not yet been specified what this significance may be, nor what the possible world looks like[6]. To determine the physical significance of the triple, a physical interpretation needs to be provided: which I do in the next subsection.

So I will dub as the *bare theory*: just the formal triple $T = \langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$, together with its structure, symmetries, and rules for forming propositions, such as: "the value of the operator $Q \in \mathcal{Q}$ in the state $s \in \mathcal{H}$ is such and such". But there is no talk of empirical adequacy yet.

We normally study a theory through its models. A *model* is construed as a representation of the theory, the triple $\langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$,[7] and will be denoted by M. A model M of a theory T may include, in addition to the triple, some variables which are part of the descriptive apparatus but have no physical significance (in the sense of the paragraph before the previous paragraph) from the point of view of the theory: I will call this the *surplus representational structure* of the model M. A theory $\langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$ will then be taken to be an equivalence class of such models. Since the class can be represented by any of its models, my account does not seek to eliminate the surplus representational structure, which varies from one model to another, but rather to identify the core $\langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$

[4]Namely, self-adjoint, renormalizable operators that are invariant under the relevant symmetries.

[5]But more generally: the dynamics does not always have to give rise to, or be limited to, a *time evolution*. A non-trivial constraint imposed on the set of states will also count as dynamics.

[6]It is important to recognize that this framework allows for the construction of theories and models that do not describe actual physical reality. While one can give various metaphysical construals to concepts such as 'possible worlds', I am here just considering the ways in which the possible worlds can be, in the context of a 'putatively fundamental' spacetime physics. Thus I do not mean to subscribe to a specific metaphysics of possible worlds. See also footnotes 17 and 21.

[7]The representations should be equivalent. In the context of duality, the potential problem of having inequivalent representations is avoided by the construal of a duality as an isomorphism between models, so that the inequivalent representations are automatically ruled out by the duality. Thus a theory will be an equivalence class of models. See §1.2.

of the theory: as that structure which is preserved across equivalent models (this kind of isomorphism will be discussed in §1.2). In De Haro et al. (2016: cf. §3.1(1)) such a construal of theory was called a 'via media'. This account of model is more general than a prevalent understanding of 'model of a theory' as a specific history of a system (but also more specific in another respect, as explained in the same reference). In particular, in our duality of interest, gauge/gravity duality, I will regard the two sides to be two different models of the theory.

The models of the theory can, in their turn, have a set of models, in the more restricted sense of 'solutions to the equations' describing histories of the system. On a suitable physical interpretation, many of these solutions will be inequivalent. So we quotient a theory by its equivalence class of models on the main sense of models which I use in this paper, but not on the more restricted sense. In what follows, I will use 'model' exclusively in the sense defined above, viz. as a representation of the triple T.

Having introduced theories and their models, we should now distinguish the *symmetries of the theory* from the *proper symmetries of the models*:

(SymT) [version 1] The *symmetries of the theory* are automorphisms $s : \mathcal{H} \to \mathcal{H}$ preserving (a subset of) the valuations of the physical quantities on the states (for details, see De Haro et al. (2016: §3.3)), and which commute with (are suitably equivariant for) the dynamics $\mathcal{D}$. In the conception of a duality which I will introduce in the next subsection, I will require the duality to preserve the symmetries of the theory. Because the symmetries of the theory T are constitutive to the its definition as a triple, and the models M instantiate the theory, it follows that the symmetries of the theory are also symmetries of the models (though the way in which these symmetries are represented does, of course, vary from model to model: and the more interesting a duality is, the more these representations differ!). More precisely, a symmetry s of the theory T induces a symmetry $\sigma(s)$ of the model M. For an example of a symmetry of the theory, see the discussion of conformal symmetry, in §2.2, as being part of the 'common core' of the two models.

(PSymM) [version 1] The *proper symmetries of the models* are those symmetries of the models (viz. of a *single* model) that are not symmetries of the theory, in the sense of being trivially represented (as identity maps) by the theory (and by the other models of the theory). In other words, they are symmetries of the *surplus representational structure* of a particular model. Since a proper symmetry of a model is not a symmetry of the theory (nor is it a symmetry of the other models of the theory), these symmetries do not map (they map trivially) across duality.

In the seemingly intermediate case in which a symmetry of a model maps only partially (through a forgetful map), but nontrivially, to another model under duality, the part of that symmetry which is common to all the models still counts as a *symmetry of the theory*. 'Mapping partially' here should be understood in terms of the generators of the symmetry: a (PSymM), e.g. a diffeomorphism, usually has several generators: some of which map, and some of which do not map, under duality. In such a case, in which the generators form a symmetry which maps under duality and hence generate a symmetry of the triple, the part of the symmetry which maps counts as a symmetry of the theory.

In other words, from this intermediate symmetry we can derive a true symmetry of the theory. Thus we count as 'symmetries of the theory' any symmetries of the models which furnish nontrivial representations of some symmetry of the theory. An example of this are the conformal transformations in gauge/gravity duality, as represented on the gravity side.

The above view on symmetries was active, i.e. it was defined as a map changing the states $s : \mathcal{H} \to \mathcal{H}$. This immediately prompts the idea of a passive symmetry, through s's *dual map* on quantities, $s^* : \mathcal{Q} \to \mathcal{Q}$ (De Haro et al. (2016: §3.3)).

In what follows, it will be convenient to regard a symmetry, whether active or passive, as a map on a *triple*. For an active symmetry, the map $s : T \to T$ acts on non-trivially the states, $s : \mathcal{H} \to \mathcal{H}$, and it commutes with (it acts trivailly on) the elements of $\mathcal{Q}$ and with the dynamics. A passive symmetry $s^* : T \to T$ acts as the unit map on the states, it maps $s^* : \mathcal{Q} \to \mathcal{Q}$, and it commutes with the dynamics. As I argued above, the symmetries of the theory are also symmetries of the models, hence a symmetry s of a theory T induces a map $\sigma(s) : M \to M$ on each of its models, i.e. an automorphism of the model.

Notice that the question whether symmetries should map across duality is a different one from the question which symmetries should be regarded as physical; though the two questions are, of course, related. The reason for the difference is that the question of what is physical can only be answered once an interpretation has been given (which I will do in the next subsection). The distinction between symmetries which are redundant and symmetries which are non-redundant (hence physical) was discussed, in the case of dualities, in De Haro et al. (2016: §2).

1.1.2 Interpreted theory

There is a certain minimalism to the above definition of theory: since in scientific practice one must be able to tell, in a given experiment or physical situation to which the theory is supposed to apply, what the relevant quantities are which correspond to the empirical data. The above specification of a theory as a triple makes no reference to this as yet: only the existence of some such relation, for some possible world W, is assumed. So, when interpreting a theory one wishes to do the following (the numbering below follows the numbering of the interpretation maps, in §1.1.2.a):

(0) establish the meaning of certain theoretical entities (if one is a realist), whether directly measurable or not.

(1) establish some kind of bridge principles between the physically significant parts of the theory and the world.

These two desiderata will be fulfilled by the two interpretative maps (which will be denoted $I^0_{T,D}$ and $I^1_{T,D}$, respectively, in 1.1.2.a), where T is the theory and D is the relevant domain which it describes, at the world W.[8] Furthermore, one may also wish to establish *theoretical principles* (1.1.2.b) which, for example:

[8] D is called the 'domain' because it is the physical domain of the world which is described by the theory. But D is, mathematically, the *co*domain, rather than the domain, of the interpretative map. There should be no confusion between the physical and mathematical uses of 'domain'.

(2) interconnect various experimental results (causality, locality, and symmetry being just three examples of such theoretical principles often considered in physics);

(3) plug interpretations into some of the things that go into the choices made by the experimenter.

Doing this is the role of the 'physical interpretation', which consists of two parts:

1.1.2.a The two interpretation maps

I take a *physical interpretation* to be a pair of surjective maps, preserving appropriate structure, from the theory to some suitable set of physical quantities. I will denote the maps as $I_{T,W} := (I^0_{T,W}, I^1_{T,W}) : T \to D$, where T is the triple (or Cartesian products thereof)[9] and D is the domain at the possible world W. (Since, once the theory is specified, it is also clear which domain it purports to describe, I will often drop the second subscript and write the map as I_T.)

The first surjective map, I^0_T, is from the triple, $\mathcal{H}, \mathcal{Q}$ and $\mathcal{D}$, to the quantities in the world (potential energy, magnetic flux, etc.), realized in a particular laboratory experiment, in the domain D to which the theory applies.

An example of such a map is: $I^0_T(Q) = q$, where q is the quantity[10] (endowed with appropriate units) in the laboratory experiment corresponding to $Q \in \mathcal{Q}$. For classical theories, q is itself a map recording the time evolution of the possessed value, $q : \mathbb{R} \to \mathbb{R}$, given by $t \mapsto q(t)$, where t represents time, and $q(t)$ is the possessed value of the quantity at time t. For instance, in a classical theory that is invariant under time translations, Q could represent the Hamiltonian and q the energy, which is a constant map sending each time instant $t \in \mathbb{R}$ to the (time-constant) value $q(t)$ of the energy in the system.

The second surjective map, I^1_T, preserving appropriate structure, is from the triple, $\mathcal{H}, \mathcal{Q}$ and $\mathcal{D}$ (or Cartesian products thereof), to the set of values (the set of numbers formed by all possible experimental outcomes) in the domain D to which the theory applies, at some world W. That set of values is typically (minimally) structured, and such minimal structure is to be preserved by the map. Thus typically, the second map maps: $I^1_T : \mathcal{H} \times \mathcal{Q} \to \mathbb{R}$, where $\mathbb{R}$ is endowed with addition and multiplication (for a concrete example, see §1.3.1), and there is a similar map for the dynamics. As mentioned, Cartesian products of elements of the triple with themselves are needed in some cases.

Alternatively, the second map can also be described as mapping the *values* of the physical quantities $\mathcal{Q}$, evaluated on the appropriate states, to the possible experimental outcomes.[11]

[9] In quantum mechanics, the interpretation map maps e.g. expectation values to real numbers in the world. The expectation values themselves are maps from Cartesian products of states and quantities to real numbers. This is why Cartesian products of theories are needed.

[10] In what follows, 'physical quantity' will always refer to members of $\mathcal{Q}$, so there will be no confusion.

[11] In a classical theory, there is usually no need for the distinction of the two maps (I^0_T, I^1_T), because the physical quantities have definite values. In a quantum theory, the instrumentalist, who is not necessarily commited to a particular interpretation of states and operators as corresponding to definite physical quantities, may wish to do away with the first map, so that only the second map, I^1_T, is left. The latter is the map the values of which are the probabilities (when evaluated on states) and Born-rule expectation values (when evaluated on operators). We will take these differences in our stride, so that the interpretive scheme developed here should apply to all these cases, once the maps are judiciously chosen (or, as the

The codomains of $I^0_{T,D}$ and $I^1_{T,D}$ thus defined seem, at first sight, to be again just formal entities, themselves in need of interpretation—a space of functions in the first example, a set of real numbers endowed with addition and multiplication, in the second. But this is not quite what we are doing. We should really think of the codomain as representing the real world in a straightforward way: so, the reals measuring the energy are to be thought of as, for example, the position of a voltimeter's pointer on a scale: a location in space labelled by a number with appropriate units—or, simply, a real number, with its units, on the voltimeter's electronic display: a specific digital signal (for some explicit examples, see §1.3.1). The point is to map from theories to structured sets of functions and numbers, which do *not* describe more theory, but rather are identified with a set of possible physical situations, physical configurations, or experimental outcomes.

The domain of the interpretation maps, which is the theory T, is large enough that many possible worlds are described by the same theory T and different interpretation maps. For instance, the state space $\mathcal{H}$ will generally contain single-particle as well as many-particle states, whereas a specific domain D, or even a particular possible world W, may only contain one-particle states. So, for instance, take five different domains $D_1, \ldots, D_5$; then, for each of them, we have a map $I_{T,i} : T \to D_i$ $(i = 1, \ldots, 5)$. This means that, for given D_i, most elements of T map to the empty set $\emptyset \subset D_i$, where the empty set straightforwardly corresponds to 'nothing in the domain'. For example, the interpretation map $I_{T,i}$ interprets a state $s \in \mathcal{H}$ mapping to the empty set as 'there is no physical situation, in the domain of the world D_i, corresponding to the state s'. And a *quantity*, mapped by $I_{T,i}$ to the empty set, is interpreted as 'there is no magnitude, in the domain of the world D_i, corresponding to that quantity'. And so on. Because the domains D_i are typically much smaller than what the theory T possibly describes, a large number of elements in the domain map to the empty set for a given map I_{T,D_i} (but *all* the elements in T should map nontrivially to *some* domain D at some possible world W). Alternatively, one can take the interpretation maps to be partial maps, i.e. they do not map all the elements in the domain T to elements in the codomain D. But given such a partial map, one can always construct a total map in the way just discussed: by adding the empty set to the codomain and mapping to it, thus obtaining the interpretation of 'nothing in the domain', as discussed above.

The above still leaves open the question of how these quantities in the world represent the theory, which is what the contrast of external vs. internal interpretation (below, and in §1.3) seeks to clarify.

In the above discussion I have so far assumed that the interpretative maps are from the elements of *the triples themselves*, rather than from their models, to quantities in the world. That is, we have assumed a map $I_{T,D} : T \to D$ from elements of the theory T to the domain D at the possible world W. We will call such an interpretation *internal*: it requires nothing but the theory. But, as I will discuss in detail in §1.3.1, there are often good reasons to pursue a different kind of interpretation: one which couples T to some other theory T_{meas} which is already interpreted, so that T inherits its interpretation from

case may be, they be left empty, i.e. they map to the empty set on most of the domain, see two paragraphs below). Thus the aim of this interpretive scheme is not to decide on metaphysical issues, but to formalise our views once those decisions are made, so that the interpretations can be contrasted.

T_{meas}. Often, the coupling of T to T_{meas} will differ for the various models M of T. In other words, T_{meas} may be coupled to M through M's surplus representational structure. So we need to distinguish:

(1) *External interpretation of* T: a pair of maps, as above, $I_M : M \times T_{\text{meas}} \to D$ from the *model* M of the theory T, that is coupled to the theory of measurement T_{meas}, to the domain $D \subset W$. Obviously, a theory can have as many external interpretations as models and as measurement theories to which it is to be coupled. Again, one may write the map as $I_{M,D}$, but I will often drop the codomain.

(2) *Internal interpretation of* T: a pair of maps, as above, $I_T : T \to D$ from the *theory* T to the domain $D \subset W$. There is no coupling to T_{meas} needed.

External and internal interpretations will be taken up in detail in Section 1.3. As remarked before, the domain of the map may involve Cartesian products of elements of M or of T.

Internal interpretations must always map from *theories* not from *models* (I discuss this further at the end of §1.3.1). On the other hand, the domain of an *external* interpretation may be either the model M of the theory T, or the theory itself, depending on the interpretation. So, a map $I_T : T \times T_{\text{meas}} \to D$ always gives rise to an external interpretation, despite its mapping from $T \times T_{\text{meas}}$ rather than $M \times T_{\text{meas}}$. Indeed, the key distinction between the external and the internal interpretations is whether the domain of the interpretative map includes anything *more than* the theory—viz. whether it includes a theory of measurement, an extension of T, or some of the surplus representational structure in a model M of the theory T. In Section 1.3 it will become clear, once we discuss the notion of extendability, when one should apply an external or an internal interpretation.

1.1.2.b Theoretical principles

The physical interpretation provides the formalism with *theoretical principles* (to be realized in an experiment) which constrain the phenomena, play a role in their explanation, and describe various of their properties.[12] Typically, these principles (of which the aforementioned causality, locality, and symmetry are examples: cf. (2)-(3) in the preamble to §1.1.2) are expressed as properties of, or appropriate restrictions on, the interpretive maps: e.g. indicating symmetries, correlations between the possible experimental outcomes, etc. For the purposes of this paper, the only theoretical principle which I will consider in detail is that of symmetries. Working out examples of *other* theoretical principles, which are reflected in the properties of the interpretative maps (e.g. locality as a property of the interpretation map, seting to zero the correlations between quantities evaluated at spatially separated points), will be left for the future. Nevertheless, I discuss them here briefly because they are important to the discussions of the interpretations of theories which one finds in the literature, as well as for the completeness of the interpretive scheme just laid out.

We should now consider appropriate meshing conditions between the maps (1.1.2.a)

[12] A physical interpretation (which, for brevity, I will refer to as an 'interpretation') should not be thought of as something imposed 'from the outside' on the theory (see §1.3). For this reason, I will use the phrase 'physical quantities' rather than 'observables'.

and the principles (1.1.2.b) just introduced. For simplicity, I will concentrate on symmetries and on the internal interpretation. There are two meshing conditions between symmetries and the internal interpretation to be considered. My two classes (Physical) and (Redundant) below correspond to the non-redundant and redundant symmetries, respectively, in the taxonomy of De Haro et al. (2016: §2), in the specific case of the internal interpretation. They also correspond, broadly speaking, to Caulton's synthetic, respectively analytic symmetries (Caulton (2015)). They are as follows:

(Physical) An internal interpretation I_T must commute with every element of (SymT). This is because the symmetries (SymT) are symmetries of the theory, which must be respected by an internal interpretation, since the latter is based on the theory, and the theory only.
By 'commutes' here, I mean that, if $s : T \to T$ is a symmetry of the theory,[13] then there is a corresponding map $s' : D \to D$ on the domain of the physical world, such that there is a commuting diagram: $I_T \circ s = s' \circ I_T$. I will call the map s' a *symmetry of the world*. Furthermore, s' should be non-trivial whenever s is (see two paragraphs below).

(Redundant) A proper symmetry σ of the model M is trivially represented on an internal interpretation: $I_T \circ \sigma = I_T$. Alternatively, in the notation of (Physical), $s' = \mathrm{id}$.[14] This is because the set (PSymM) was defined (in §1.1.1) as symmetries of the models which are *not* symmetries of the theory. Since an internal interpretation has only *triples* in its domain (and nothing else), it *cannot* map a symmetry of the surplus representational structure (which is the structure on which the symmetries (PSymM) act) to a non-trivial symmetry in the world: on pain of giving physical salience to things that are not in the triple—and this is not the job of the internal interpretation. In other words, a proper symmetry transformation should not make any difference on an internal interpretation, since it changes elements in the model without changing anything in the theory.

One might think that there are cases of symmetries s of the *theory* T which are not symmetries of the world (in the sense of $s' : D \to D$ under (Physical)), and thus commute according to the more restrictive (Redundant), viz. as $I_T \circ s = I_T$. These putative symmetries are found in the theory but not in the world. And so one might be inclined to relax the condition (Physical), allowing s' to be trivial. But since these are not symmetries of the world, the putative theory T (more precisely: the formulation of it with which one is working) has a redundancy. So one is actually dealing with a model M of T rather than T itself. On quotenting M by s, one obtains a more perspicuous representation of T, on which s is trivially represented. So the putative (Physical) s is actually a (Redundant) symmetry of M, and one should not relax (Physical).

The commutativity requirement for (SymT) implies that, under the internal interpre-

[13] For how s acts on each of the components of the triple T, see the remark in the penultimate paragraph of §1.1.1.

[14] More precisely, but perhaps a bit pedantically, one might write: $I_T \circ F \circ s = I_T \circ F$, where $F : M \to T$ is a forgetful map, from the model to its corresponding triple. This forgetful map effectively removes the surplus representational structure from M and embeds the remainder into T, by isomorphism. I will continue to use the shorthand above.

tation, these are *physical symmetries*, in the sense that they are found in the world—hence the label (Physical). Typically, such symmetries relate physically different situations, and so they have noticeable effects. The internal interpretation maps the symmetries of the theory T to symmetries of the physical domain D at W. On the other hand, (Redundant) are *not* physical symmetries but just redundancies of the models.

I will call a bare theory, once it is equipped with an interpretation, the *interpreted theory*.[15] It is the physical interpretation that enables the theory to be empirically successful and physically significant. The physical interpretation should also make the ontological commitments explicit (more on this in §1.4).[16]

Both (1.1.2.a) and (1.1.2.b) involve, of course, philosophically laden issues. And the aim here is not to settle these issues, but rather to have a scheme in which the formal, the empirical, and the conceptual are clearly identified and—as much as possible—distinguished, within the structure of the interpreted theory. Indeed, I believe that the conflation of these three aspects can easily lead to confusion.

The above formulation of a bare theory as a triple is minimalistic. But, with (1.1.2.a) added, it is strong enough—because of the complete specification of the set of physical quantities—that, with an additional requirement below, it will be able to determine when two theories are about the same subject matter. Questions concerning the identity of two such triples will be questions concerning the sameness of theories (§1.3.1), rather than standard cases of underdetermination of theory by empirical data: since the triple $T = \langle \mathcal{H}, \mathcal{Q}, D \rangle$, together with the valuations constructed from the syntax, is assumed to be well-defined and consistent, and to encompass all the empirical data, in a certain domain D at W.[17] I will call such a theory *complete*. Completeness is then a necessary condition for there to be a duality.

1.2 The conception of duality

In this subsection, I introduce the conception of a duality, based on the notion of a theory developed in §1.1. In §1.3 I will relate this conception of duality to the discussion of theoretical vs. physical equivalence.

With the conception of a theory considered in §1.1, a *duality* is now construed as an *equivalence of bare theories*. More precisely, it is an isomorphism $d : M_1 \to M_2$ between

[15] Of course, there is a rich literature that conceives 'quantum philosophy' and the controversies on 'the interpretation of quantum mechanics' as a matter of what to add to a 'minimal quantum mechanics'. This is of course not the business of my (a) and (b): cf. also next paragraph.

[16] Ismael and van Fraassen (2003: §2) start with a theoretical ontology and only later add a set of laws; in my approach, the formal and ontological steps are reversed. The reason for this is not my preference for one aspect over the other, but simply because the ontology of 'unextendable' theories (cf. §1.3.1) is not a priori clear, and so we get a certain economy of thought by starting with the formal structure and considering only those interpretations that are consistent with it (see §1.3). As it turns out, the ontologies used to formulate theories independently of any duality are not helpful in interpreting dualities for unextendable theories (cf. §1.3.1). But I believe that several aspects of what I have to say are broadly in line with Ismael and van Fraassen's discussion of symmetries.

[17] The qualifications of a *domain* and a *possible world* are important because the theory need not be complete in *our world*, and the relevant domain may change from one world to another. A theory may, for example, be complete within a given range of parameters which does not contain the values they take in our world. Hence, completeness of a theory is to be construed as relative to D and W.

two models M_1 and M_2 of a theory T: there exist bijections between the models' respective sets of states and of quantities, such that the values of the quantities on the states, e.g. in the case of quantum theories the set of numbers $\langle s_1|Q|s_2\rangle$, where $s_1, s_2 \in \mathcal{H}$, $Q \in \mathcal{Q}$, are preserved under the bijections; the duality also commutes with (is equivariant for) the two models' dynamics and preserves the symmetries of the theory, (SymT), as defined in §1.1.1.[18]

In defining duality this way, one should keep in mind the conception (§1.1.1) of what I am calling a *model*: since what is here called a 'model' is often called a 'theory'. And one should clearly distinguish the notions of theory and model, understood as given mathematical structures, from the way one *gets to* recognise these notions in an example. So one may start with two theories T_1 and T_2 (as triples) and find a duality map between them: in which case, one identifies the triples as being one and the same bare theory. In such a case, the original theories (with their surplus representational structure) are now revealed to be models of one theory. But this is mere historical record: what matters, for a duality, is that one has two models which are isomorphic in terms of their triples. Also, I do not mean to limit duality to just *two* theories: in principle, there can be a duality between any number of theories (cf. for example: S-duality in quantum field theory, where the duality group is $\mathrm{SL}(2,\mathbb{Z})$ (Vafa and Witten (1994: §1))).

On the above conception of duality as an equivalence between models, we can now understand the two kinds of symmetries, (SymT) and (PSymM), introduced in §1.1.1, as meshing conditions between duality and symmetry, and reformulate them using the duality map:

(SymT) [**version 2**] is the condition that the symmetry in question *commutes* with the duality map, so this is a natural meshing condition between symmetry and duality. By 'commutes', I here mean the following. Given two models M_1 and M_2 of a theory T, and given two symmetries $\sigma_1 : M_1 \to M_1$, $\sigma_2 : M_2 \to M_2$ (defined as in the penultimate paragraph of §1.1.1), the following two conditions must be met:

(Visible) σ_1 and σ_2 are both *non-trivial representations*, $\sigma_1(s)$ and $\sigma_2(s)$, of a symmetry s of the theory, where 'non-trivial' here means that they are not the identity map. Thus, for a symmetry to be 'visible' in the theory, is for it to be *non-trivially represented* by the theory.

(Commutative) Duality and symmetry form a commutative diagram: $d \circ \sigma_1(s) = \sigma_2(s) \circ d$.

This generalises to the intermediate case, mentioned in §1.1.1 just after the introduction of (SymT) and (PSymM), of symmetries which act not only on the triples but also on some of the surplus representational structure.

[18] In mathematics, duality is a diverse phenomenon, of which there is no single definition. At the most basic level, duality boils down to the notion of a *natural* isomorphism (in category-theoretic language: an adjoint functor). I am refraining from this and other refinements of the notions of duality and of theory: for to make my main points, I will not need to articulate such mathematical details. For an account of theoretical equivalence in physics using category-theoretic terms, see Weatherall (2015). Note also that I am using $\langle s_1|Q|s_2\rangle$ (i.e. matrix elements using two different states), not the more obvious $\langle s|Q|s\rangle$ (i.e. an expectation value in a single state), for reasons of quantum theory.

(PSymM) [**version 2**] is the condition that the symmetry $\sigma_1 : M_1 \to M_1$ is represented trivially on the codomain of the duality, M_2. (This is analogous to what happens in the case of the internal interpretation (Redundant): cf. §1.1.2.b.) Here, σ_1 maps non-trivially the surplus representational structure of M_1, rather than the triple. (PSymM) can helpfully be expressed in terms of the following invisibility condition on the symmetry σ_1 of the model M_1 and the duality map d:

$M-$**(Invisible)** $d \circ \sigma_1 = d$, where $d : M_1 \to M_2$. In such a case, we say that σ_1 is M_2-(Invisible). On the other hand, given a (PSymM) $\sigma_2 : M_2 \to M_2$, then σ_2 is M_1-(Invisible) just in case $\sigma_2 \circ d = d$. Alternatively, we require that either σ_1 or σ_2 in (Commutative) are identity maps, i.e. one of the two representations is trivial. If σ_1 is trivial, then σ_2 is M_1-(Invisible); if σ_1 is trivial, then σ_2 is M_2-(Invisible). Notice that σ_1 and σ_2 are not (Visible). Hence, they do not represent an underlying symmetry s of the theory T.

A symmetry σ of the model M is a (PSymM) if it is $N-$(Invisible) for *some* model N.

For more on the visibility and invisibility conditions in the case of diffeomorphism symmetry, see De Haro (2016a: especially §3).

The notion of duality in this section is motivated by both physics and mathematics. Duality in mathematics is a formal phenomenon: it does not deal with physically interpreted structures (even though, of course, several of the mathematical dualities do turn out to have a physical significance). But this is also how the term is used by physicists: it is attached to the equivalence of the formal structures of the theories, regardless of their interpretations, i.e. without it necessarily implying the *physical* equivalence of the theories which describe two concrete systems.

Duality, as a formal equivalence between two triples without the requirement of identical interpretations, is thus a special case of theoretical equivalence. For an account of theoretical equivalence of unformalised theories, cf. e.g. Coffey (2014).

Like the conception of a theory, my conception of a duality is minimalistic: on this definition, for instance, the verdict over position-momentum duality in quantum mechanics is that it is indeed a duality. The duality has two models, namely the formulations based on, respectively, the x- and the p-representations of the Hilbert space: Fourier transformation being the duality map (for more on this case, cf. §3.2). This duality is, of course, somewhat trivial, because the two models contain *the same amounts of surplus representational structure*, in the sense above: namely, a single variable. The two models are already formulated in terms of their triples of states, quantities, and dynamics. And I indeed regard it as a virtue that my conception of duality is general enough that both familiar, and relatively simple, dualities, as well as the more sophisticated ones in string theory and quantum field theory, all qualify as dualities, under the same general conception. Indeed, I take it that one of the lessons of duality is that 'widely differing theories' are (surprisingly) equivalent to each other, in the same sense that two notational variants differ from each other. And my interest in this paper will not be—and I will not need—to distinguish between 'simple' and 'sophisticated' dualities (or cases in between), though in principle my notions do allow for such a characterisation.

Indeed, I submit that my notion of surplus representational structure, and in partic-

ular (PSymM), *can* distinguish 'simple' from 'sophisticated' dualities, and thus can be used to indicate how 'surprising' a duality is supposed to be, by the amount of surplus representational structure by which the models differ. The larger the amount of (PSymM) which the models possess, the more they differ. But as remarked before, for the purposes of this paper (anwering (i)-(iv) in the Abstract) I will not need to further refine the notion of duality along these lines.

Another way in which, in the physics literature, a distinction is sometimes made between 'simple' and 'sophisticated' dualities, is as follows. Physicists sometimes reserve the word 'duality' for cases of *equivalent quantum theories with different classical descriptions* (Aganagic (2016: Abstract), Polchinski (2016: §1)). While I sympathise with the aim of this notion, it seems to me too restrictive, for: (i) it only allows for dualities between quantum theories; (ii) the notion of a classical description of a theory is not always necessary, or even clear: sometimes, quantum theories have more than one classical regime, and then additional criteria would need to be gived as to which classical regimes one should compare; (iii) the notion of 'different' classical regimes is vague and would need to be replaced with 'inequivalent', on pain of counting as different those classical limits which merely differ by being formulated in different terms. But then an account of *classical equivalence* is also needed. An example of this is the quantum mechanics of a point particle in a velocity-independent potential, formulated in terms of a path integral over all positions, or in terms of a path integral over all positions *and* momenta. These two theories are quantum mechanically equivalent, but their classical limits give rise to Lagrangian, respectively Hamiltonian, models of the mechanics of a point particle under a velocity-independent potential. The latter models are of course equivalent at the classical level: but introducing a notion of classical equivalence undermines the need to require that the equivalence can only be quantum mechanical.

Introducing the notion of models for a *single* theory, and construing duality as an isomorphism between such models, solves this problem: if two models are dual, any two other models isomorphic to them are dual as well.

The two-pronged conception of an interpreted theory (or model) as a triple plus an interpretation, together with the formal definition of duality, allow us to introduce the notion of *physical equivalence* of interpreted models. The discussion of duality indeed prompts us to distinguish *theoretical equivalence* from *physical equivalence*: the latter being the equivalence of two theories as descriptions of *physical systems*, i.e. theories with identical interpretations. More on this in §1.3. The difference may be cashed out as follows: theoretically equivalent theories, once interpreted, 'say the same thing' about possibly *different subject matters* (different parts of the world), whereas physically equivalent theories say the same thing about the *same subject matter* (the same part of the world). We will see also later a linguistic/syntactic conception and infer the quasi-linguistic nature of theoretical equivalence, but throughout we will be mainly concerned with theories in the sense of triples, and thus with duality as an isomorphism between such triples.

An example of the former is the diffusion equation with constant diffusion coefficient: it describes a wide variety of different phenomena—including the Brownian motion of particles on a liquid, and the diffusion of heat in a material. These two models, taken formally, are isomorphic: the isomorphism mapping one value of the diffusion coefficient to another, and the density of Brownian particles to the temperature distribution for the

heat equation: but the two isomorphic models, physically interpreted, clearly describe two different parts of the world, viz. Brownian motion of particles, and heat conduction. This is an example of a 'simple' duality, in the sense discussed earlier in this subsection (after the example of quantum mechanics), i.e. of introducing only minimal surplus representational structure, though that structure can appear if we couple the diffusion equation to a theory of hydrodynamics, in the first model, or to thermodynamics, in the second (but in that case there ceases to be theoretical equivalence).

An example of the latter, viz. of isomorphic models describing the *same* part of the world, is provided by the Lagrangian and Hamiltonian formulations of mechanics, which, under standard assumptions on the class of models considered, are isomorphic and widely taken to describe the same physical systems.

Duality, then, is one of the ways in which two theories can be theoretically equivalent, without its automatically implying their physical equivalence. For instance, a duality can relate a real and an imagined or an auxiliary system. In such a case, duality is a useful and powerful calculational device—and nothing more. Thus it is a good idea, in line with both the physicists' and the mathematicians' practice, to keep the notion of duality formal, as an isomorphism between *bare theories*.

But it is, of course, those cases in which dualities do reveal something about the nature of physical reality, that prompts the philosophical interest in dualities: cases in which the interpretation of the duality promotes it to physical equivalence.

1.3 From theoretical equivalence to physical equivalence

Having introduced the notions of theory, interpretation, and duality in the previous two subsections, we now come to the central question in this Section: when does duality amount to physical equivalence? I will first discuss, in §1.3.1, the external and internal interpretations of a theory, already briefly introduced in §1.1.2.a. In §1.3.2, I will give a Newtonian example of an external interpretation; and in §1.3.3, I will discuss the physical equivalence of dual theories that are unextendable. In §1.4, I will compare my account with extant notions of equivalence, specifically Glymour's (1977) notion.

1.3.1 External and internal interpretations of a theory

In §1.2, I emphasised that, according to both physicists and mathematicians, duality is a formal feature of theories and that, before it is given an interpretation along the lines developed in §1.1.2, it implies next to nothing about the world. In this subsection I will develop the external and internal interpretations in more detail, and in particular two cases: (i) cases of external interpretations, in which physical equivalence fails to obtain despite the presence of a duality; (ii) cases in which an external interpretation is not consistently available (where 'consistently' will be qualified below), so that one can only have an internal interpretation, and hence also physical equivalence.

Let me first illustrate the external interpretation with an example which should make clear the difference between duality as a case of theoretical equivalence, and physical equivalence.

Consider, as an elementary example, classical, one-dimensional harmonic oscillator 'duality': an automorphism $\mathcal{H} \to \mathcal{H}$, defined by $\mathcal{H} \ni (x, p) \mapsto (\frac{p}{m\omega}, -m\omega x)$, from one harmonic oscillator state to another, leaving the dynamics $\mathcal{D}$ invariant—namely, the Hamiltonian $H = \frac{p^2}{2m} + \frac{1}{2}\, kx^2$ and the equations of motion are invariant under it.[19] So it is an automorphism of $T_{\text{HO}} = \langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$. But this automorphism of $\langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$ does not imply physical equivalence of the states:[20] the two states are clearly distinct and describe different physical situations: since the map relates an oscillator in a certain state of position and momentum, to an oscillator in a *different* state.

For example, there is an independent way to define the 'position' of the oscillator, as well as to measure its value at a given time: one couples the oscillator to a standard rod and so carries out a measurement of the oscillator's position, $I^1_{\text{HO+meas}}(x, r) \in \mathbb{R}$, where x is the position of the oscillator and r a physical location on the rod. Here, we have coupled the harmonic oscillator theory T_{HO} to our theory of measurement T_{meas} (in fact, several theories: Euclidean geometry, and standard assumptions about the stability of the measurement instruments based on their material properties). For an ideal measurement, we summarise this with the map $I^1_{\text{HO+meas}} : T_{\text{HO}} \times T_{\text{meas}} \to D$ (cf. §1.1.2.a). The role of T_{meas} is to transfer its interpretation (e.g. of r as 'rod position') to T_{HO} (of x as 'oscillator position'), by spatial juxtaposition. For non-ideal measurements (which include non-trivial interactions between the two theories), we have instead a new, extended theory $T_{\text{HO+meas}}$, and the above map is generalised to: $I^1_{\text{HO+meas}} : T_{\text{HO+meas}} \to D$.

The above coupling to a theory of masurement takes place according to some, usually unformalised, principles about the measurement procedures (such as juxtaposition). Since the product form $T_{\text{HO}} \times T_{\text{meas}}$ is only an idealization, in more realistic cases we can find a more comprehensive theory such as $T_{\text{HO+meas}}$, in which the original theories can be embedded: $T_{\text{HO}} \times T_{\text{meas}} \subset T_{\text{HO+meas}}$ (for more on the notion of 'embedding', see the next subsection).

I will call such an interpretation of T_{HO}, obtained by transfering its interpretation from a theory of measurement T_{meas} (in other words, from an already interpreted theory), or by extension to $T_{\text{HO+meas}}$, an *external interpretation* (cf. §1.1.2.a). And I call a theory, that can be coupled or extended in this way, an *extendable* theory. T_{HO} is the extendable theory, and $T_{\text{HO+meas}}$ is the extended theory (itself again extendable).

But there are cases—such as cosmological models of the universe, and models of unification of the four forces of nature—in which these grounds for resisting the inference from duality to physical equivalence—a resistance based on the possibility of finding an external T_{meas}—are *lost*. For the quantum gravity theories under examination—even if they are not *final* theories of the world (whatever that might mean!)—are presented as candidate descriptions of an *entire* domain of (possible) physical world: let us call such a

[19] The angular frequency, here and in what follows, is given by $\omega := \sqrt{k/m}$. The equations of motion of the harmonic oscillator are $p = m\,\dot{x}$ and $\dot{p} = -k\,x$. When the duality map is considered on solutions of the equations of motion, the map effectively adds $\pi/2$ to the phase of the oscillations, so that it maps solutions to solutions with different initial conditions.

[20] In a world consisting of a single harmonic oscillator *and nothing else*, the two situations could not be distinguished, and one might invoke Leibniz's principle to identify them, which would amount to an internal interpretation, in the sense of §1.1.2.

theory T.[21] So there is no independent theory of measurement T_{meas} to which T should be coupled, because T itself should be a closed theory (an *unextendable* theory: see §1.3.3).

The interpretation of an unextendable theory (unextendability will be explicated further in §1.3.3) is called an *internal interpretation*. Such an interpretation was defined in §1.1.2. I will now discuss the bearing of an internal interpretation on the physical significance of the duality map.

We return to dualities, and the interpretation of two models of the whole world, call them M_1 and M_2. In the rest of this Section, the leading idea is that the interpretation has not been a priori fixed (or, if by some historical accident, an interpretation has been fixed, one should now be prepared to drop large parts of it), but will be developed starting from the duality. The reason for this is that, as pointed out in §1.1.2, the internal interpretation starts from T's invariant content. This invariant content is laid bare by the duality. Thus starting from two dual models M_1 and M_2 of T, the duality map lays bare the invariant content $\langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$, which is the starting point of the internal interpretation for both the theory and the models. Therefore, the interpretation of the two models is now the same. Indeed, the two models do say the same thing (in the sense of §1.2):

(i) their formalisms say the same thing, in particular they contain the same states and physical quantities, and (ii) their physical content is also the same: for the interpretation given to the physical quantities and states starts from the duality, and so the interpretative maps are the same for both models, $I_{M_1} = I_{M_2}$. Therefore, such thorough-going dualities lead to *physical equivalence* between apparently very different models. This is likely to be the case throughout the research area of quantum gravity, since any of its putative theories T_{QG} are e.g. cosmological models of the universe, or models of unification of the four forces, which comprise the entire domain of physics.[22] Such an interpretation, while perhaps starting from extant interpretations of the models M_1 and M_2, will consider an equivalence classes of such models. Thus it will be based, ultimately, only on the common features of the models (§1.1.2a) and on the theoretical principles of the theory (§1.1.2.b).

In view of the above, the conception of an internal interpretation, for an unextendable theory, can now be restated as follows. The internal interpretation of T is *the same* for all models M of T, as I just argued; and, furthermore, the internal interpretation does not depend on the addition of T_{meas}, which would break the duality. Therefore, the interpretation is invariant across the duality: for unextendable theories, therefore, the interpretation of the models of T is the *internal interpretation*. Thus we arrive at the following principle, which is a meshing condition between the duality map and the internal interpretation:

(PhysEquiv) For unextendable theories, the internal interpretative map I_T *commutes*

[21] These are candidate descriptions of *possible* worlds, rather than the actual world, because the models which we will consider in the second part of the paper—viz. Ammon and Erdmenger (2015: §§5-7)—entail a negative cosmological constant, whereas our universe seems best described by a *positive* cosmological constant. But the interest in such models is, of course, that: (1) given their rarity, *any* consistent four-dimensional theory of quantum gravity is interesting; and (2) such idealised models contain helpful lessons for the case of a positive cosmological constant.

[22] A complete theory of quantum gravity will have to address the problem of measurement in quantum mechanics. But for the aims of the current discussion, and the simple models discussed in this paper, we do not need to have this problem solved. The point I am making applies equally well to classical systems (see the previous paragraphs).

with the duality map.

Consequently also, if the interpretative map fails to commute with the duality map, the interpretation is *external*.

On my conception of physical equivalence, discussed in §1.2, two theories are physically equivalent if they are theoretically equivalent and, in addition, have the same interpretations. Thus (PhysEquiv) is the condition of physical equivalence for two dual theories.

What prevents us from concluding, in the consequence after (PhysEquiv), that "The interpretation is external *iff* the interpretative map fails to commute with the duality map"? This would make the notion of an external interpretation the contrary of the internal interpretation, for an unextendable theory. That is, it would imply that whenever there is an internal interpretation, there can be no external interpretation, and viceversa. The reason why this stronger thesis does not follow is because external interpretations can always be constructed, no matter whether the theory is extendable or unextendable. But these two cases do differ: for unextendable theories, we cannot construct an external interpretation that is *physically consistent* over the entire domain of the theory. By unextendability, the theory T cannot be coupled consistently to a theory of measurement T_{meas}, and there is no embedding of T into some other theory either. One may, of course, couple the theory *by hand* to such a theory of measurement: but the result will not give a consistent theory, at least not over the entire domain. So, if one would require the interpretation to be consistent over the entire domain, then one could rule out external interpretations constructed by such means, on physical grounds. And in that case the external and the internal interpretations *would be* each other's contraries. But this is a strong condition to require—it would, at least, require further explication. For we can *always* construct inconsistent external interpretations, or external interpretations which apply to limited domains. So there is no clear reason why we should impose such a strong requirement. Thus I restrict the discussion to the weaker (PhysEquiv).

Another way in which one may wish to adopt an external interpretation, of an unextendable model, is if the surplus representational structure is supposed to have some physical significance, i.e. to map non-trivially to the world. But this contradicts the assumption that only the triple had a physical significance. In other words, *if* it is the case that the surplus representational structure has a physical significance, then we are no longer dealing with a model of the *same* theory, i.e. the theory without surplus representational structure. Rather, the model whose surplus representational structure is physical represents a new theory, which is not theoretically equivalent to the other models of the old theory. So this is not an objection, and we should not change our scheme here: it is simply a change of theory. There is, despite there being some formal equivalence, no theoretical equivalence between the old theory and the new theory.

A central aspect of my conception of physical equivalence, as stated in (PhysEquiv), is that it crucially depends on the domain of the interpretative map. Thus one may never conclude that two theories are physically equivalent without it being the case that they describe the same unextendable domain. I will discuss the significance of the unextendability condition, and of the domain, for (PhysEquiv), in more detail in §1.3.3, after we discuss, in the next subsection, an example of an extendable theory.

1.3.2 A Newtonian example of external interpretation

In this subsection, I will work out the external interpretation, and the concept of an *extendable theory*, further, in an example. By contrast, this will give us the notion of an *unextendable* theory, which we will use in the next subsection. The discussion will also clarify the role of the *domain* of a theory, for its interpretation. Consider, for instance, how an *external* interpretation is worked out, from the following uses of T_{HO}, which will render it an *extendable* theory (see §1.1.2).

Rufus is a pet monkey attached to the end of a spring, described by T_{HO}, which is hanging next to the window in Sophia's room. The displacement of the bob at the end of the spring, in T_{HO}, is interpreted as the height of Rufus' up-and-down movements, which make Sophia laugh. Their description is valid to the accuracy specified by the parameters of T_{HO}. But to give that interpretation—the displacement as the height of *Rufus'* jump—we have to specify more than the idealized T_{HO} and its interpretive maps $I^0_{\text{HO}}(x)$ = displacement and $I^0_{\text{HO}}(t)$ = time (I^0_T and I^1_T are defined in §1.1.2(a) and, as mentioned there, in a classical theory the distinction between I^0_T and I^1_T is not always needed). To interpret the jump as being Rufus' we must at least specify which monkey, for example by his distinct shape and colours: so, if the two monkeys differ in their shape and colours and nothing else, T_{HO} alone cannot distinguish them: we need, in addition, a theory of shape and colour. For mapping $I^0_{\text{HO}}(x, k, m)$ = Rufus, where k, m are the spring constant and the mass, respectively, automatically rules out a second pet monkey, Kadee, with different shape and colours, to be mapped from T_{HO}, because this theory does not contain 'shape' or 'colour' parameters.

And to interpret the gravitational field, which modifies the rest length of the spring, as being the gravitational field, inside Sophia's room, of the earth (the same field that keeps the moon on its path), we need some facts about Newton's theory which explain why the value of this field is different for Rufus and for the moon: and so we embed $T_{\text{HO}} \subset T_{\text{HO+Newton}}$. This embedding clearly allows us to distinguish, for instance, the interpretation of the spring displacement, as: (1) 'Rufus' jumping height in Sophia's room at W_1', from: (2) 'Kadee's jumping height in Oliver's room at W_N', where W_N differs from W_1 in that the radius of earth is smaller by a factor of $1/N$, and its mass is smaller by a factor of $1/N^2$. Thus in the replacement of W_1 by W_N, the gravitational acceleration $g := GM/R^2$ remains constant and nothing in the formulation changes: a fact which we can only derive in Newton's theory, since the quantities R and M do not appear in the harmonic oscillator theory.

The Newtonian theory, on the other hand, *can* distinguish between the two cases, since the radius and mass enter the equation of motion for the pet monkey not through the *constant* acceleration g, but through the *local* acceleration $GM/(R + x)^2$, where x is the displacement from the rest position.[23] Thus the extended theory $T_{\text{HO+Newton}}$ can

[23]Explicitly: the angular frequency of oscillation $\omega = \sqrt{k/m}$ is *smaller*, and equal to $\Omega := \sqrt{\omega^2 - 2g/R}$, in $T_{\text{HO+Newton}}$. This result is derived from Newton's theory, by expanding the force in the small quantity x/R, and keeping only the leading term. Thus, as long as $\Omega < \omega$, this theory can distinguish between W_1 and W_N by a measurement of the frequency. Indeed, $\Omega_1^2 = \omega^2 - 2g/R$ at W_1, whereas $\Omega_N^2 = \omega^2 - 2gN/R$ at W_N. Even if $\Omega_1 \simeq \omega$ at W_1, Ω_N will differ significantly from ω at W_N, for sufficiently large N: so that the effect will not be negligible. This is how $T_{\text{HO+Newton}}$ is able to distinguish the two worlds. T_{HO}

distinguish between W_1 and W_N, while the extendable theory T_{HO} cannot.

The way to see this from the interpretative maps is by noticing that the embedding $T_{\text{HO}} \subset T_{\text{HO+Newton}}$ changes the interpretative maps in such a way as to make it possible to interpret the displacement as being Rufus', as I will now explain. (In what follows, I will assume that $T_{\text{HO+Newton}}$ has been coupled to a suitable theory of measurement T_{meas}, as discussed in the beginning of this subsection. So that, effectively, the T_{Newton} underlying $T_{\text{HO+Newton}}$ is an already interpreted theory, from which T_{HO} inherits its interpretation.)

In T_{HO}, g has the interpretative map: $I^0_{\text{HO}}(g) = a$, i.e. the acceleration at zero displacement, which is a function $a : \mathbb{R} \to \mathbb{R}$ mapping $t \mapsto a(t)$. In the present case, the acceleration at zero displacement is a constant function, and so the second interpretative map simply gives its value, which I denote with the same symbol a: $I^1_{\text{HO}}(g) = a \in \mathbb{R}$. As mentioned in the previous paragraph, it is the coupling to T_{meas} that gives a its operational meaning as a possible measurement of acceleration.

On the other hand, the displacement x at equilibrium is interpreted as: $I^0_{\text{HO}}(x|_{\text{equilibrium}}) = \Delta\ell$, i.e. the change in the spring's rest length, which is due to the pull of gravity. Its value is given by: $I^1_{\text{HO}}(x|_{\text{equilibrium}}) = I^1_{\text{HO}}(g/\omega^2) = a/\omega^2 = \Delta\ell \in \mathbb{R}$, i.e. the theory relates the displacement of the equilibrium position of the spring to the pull of gravity (and, again, $\Delta\ell$ is a constant function and so I will use the same notation for its value). But there is no independent definition of $I^1_{\text{HO}}(g)$ which relates it to other quantities, since there is no independent account of the gravitational acceleration in the theory. So, though the equilibrium displacement is interpreted in terms of the acceleration at zero displacement and the angular frequency: and though these quantities can be measured directly, the maps contain no information about Rufus' distinguishing properties.

In $T_{\text{HO+Newton}}$, the four values of the two interpretative maps calculated above are upheld, with two changes. First, there is an important Newtonian correction to the rest length:

$$I^1_{\text{HO+Newton}}(x|_{\text{equilibrium}}) = a/\omega^2 + 2a^2/\omega^4 R = \Delta\ell \ \in \mathbb{R} \ , \tag{1}$$

again a constant. Second, we have to take into account the fact that Newton's gravitational law now *does* give an account of the acceleration in terms of other quantities, viz. $g := GM/R^2$. So we have the additional interpretive map:

$$\begin{aligned} I^1_{\text{HO+Newton}}(g) &= a \ := I^1_{\text{HO+Newton}}(GM/R^2) \\ &= \frac{I^1_{\text{HO+Newton}}(G) \times I^1_{\text{HO+Newton}}(M)}{I^1_{\text{HO+Newton}}(R^2)} \quad \in \mathbb{R} \ , \end{aligned} \tag{2}$$

i.e. the approximate local acceleration a is interpreted via GM/R^2's interpretation, where G, M, and R each have their own independent interpretations in $T_{\text{HO+Newton}}$, thus adding a new interpretation to g as the local gravitational acceleration on earth's surface, i.e. a planet of given radius and mass. Here, I have made use of the structure-preserving property of the interpretative map (see §1.1.2). (Of course, there are also 0-maps for each of the newly introduced quantities.)

The correction to $\Delta\ell$, Eq. (1), together with the new map map (2), can now be used to reinterpret the displacement at equilibrium $I^1_{\text{HO}}(x|_{\text{equilibrium}})$ in terms of properties of

cannot do this because it predicts the angular frequency ω for all worlds.

Rufus—namely, the radius and mass of the planet that he is hanging above. So Eq. (1) now explicitly depends on R, and not just on the combination $g = GM/R^2$ (which is invariant across worlds), as *was* the case in T_{HO}. This is how $I^1_{\text{HO+Newton}}(x|_{\text{equilibrium}})$ is able to distinguish between W_1 and W_N: hence between Rufus and Kadee, if we define them as 'the pet monkey hanging in a children's room at the planet of radius R and mass M'. Alternatively, the modified angular frequency of oscillation can also be used, as in footnote 19.

Now that we have successfully compared, and distinguished, W_1 and W_N using the embedding $T_{\text{HO}} \subset T_{\text{HO+Newton}}$, let us discuss, before we return to *unextendable* theories of quantum gravity and their internal interpretations, the *domains* to which the above theories apply, viz. either D_1 at W_1 or D_N at W_N: since this is part of the interpretive scheme in §1.1.2 and it will explain T_{HO}'s *extendability*. Clearly, the embedding relation is one of entailment: $T_{\text{HO+Newton}}$ entails T_{HO} but not the other way around. The entailment relation is deduction by taking the second term in Eq. (1) to zero, compared to the first one, i.e. $g/\omega^2 R \to 0$: an approximation which is valid for very large radii of earth, compared to the displacement of the spring's rest length. In addition, entailment also requires that the displacement of the string is much smaller than the radius of the earth, $x/R \to 0$. So the 'domain of validity' of $T_{\text{HO+Newton}}$ is clearly larger than that of T_{HO}, where 'domain of validity' is here meant in the innocuous sense (see the next paragraph) of there being many more situations to which the Newtonian theory applies, and that T_{HO} is obtained from $T_{\text{HO+Newton}}$ by taking the limit, $\lim_{x/R\to 0} \lim_{g/\omega^2 R\to 0} T_{\text{HO+Newton}} = T_{\text{HO}}$. The requirement that the first limit can be taken restricts the kinds of models of $T_{\text{HO+Newton}}$ which can be considered. The second is a limit on the parameters of the model.

But this innocuous sense is not what we mean by a 'domain D at W'. For the difference beween the domains of applicability of $T_{\text{HO+Newton}}$ and T_{HO} is purely configurational, i.e. it is due to the specific properties a particular configuration of matter, in the given model. When we consider the domain D at W in the sense intended in §1.1.2, we find no fundamental difference between $T_{\text{HO+Newton}}$ and T_{HO}: they are both models of classical mechanics, both applying in the domain of that theory at W. The domain *would* change if we were to, for instance, turn on quantum effects, or if we were to take relativistic effects into account.

There is no claim here of my having made the notion of a domain D at W so sharp that it can be straightforwardly applied to all theories—since for the purposes of this paper the above considerations, which can be generalised on a case-by-case basis, will suffice. A more general specification of the domain will require further formalising the relation $T_{\text{HO}} \subset T_{\text{HO+Newton}}$, something that is beyond the scope of this paper. But the concept of a domain should by now be clear: when I speak of a new domain D at W, I mean a domain that is conceptually, and not merely contingently, distinct from the old domain. Thus I will distinguish the *regime of applicability* R, from the *domain* D, of a theory T at a possible world W. On the former sense, T_{HO} and $T_{\text{HO+Newton}}$ are valid on different regimes of applicability, R_1 and R_N, defined through the above analysis of validity involving limits; on the latter sense, T_{HO} and $T_{\text{HO+Newton}}$ are defined on the same domain (D_1 at W_1 or D_N at W_N), i.e. as both being models of classical mechanics. It is the latter sense of domain that is relevant for the discussion of extendability and which I use throughout this paper.

The above makes precise the sense in which T_{HO} is an extendable theory. T_{HO} ceases to

give an accurate description of D at W when the correction terms in Eq. (1) cease to be negligible. But T_{HO} can be generalised to $T_{\text{HO+Newton}}$, which does describe the larger class of phenomena over the same domain D at W accurately: and T_{HO} is correctly recovered from it in the specified limit. Thus, T_{HO} is *extendable*.

1.3.3 Unextendability implies internal interpretations, and so duality implies physical equivalence

Back to quantum gravity. For the theories T_{QG} under consideration, there is no such extra physics to which the theories can be coupled or extended. Being descriptions of the entire physical universe, or of an entire domain of physics, the interpretation I_{QG} must be *internal* to it. Thus, as a sufficient (though not necessary) condition for the use of an internal interpretation, I will require that T_{QG} be an unextendable theory. I_{QG} only requires the triple $T_{\text{QG}} = \langle \mathcal{H}, \mathcal{Q}, D \rangle$ as input, and it only involves the triple's elements and their relations—it does not involve coupling T_{QG} to other theories. In such a case, duality preserves not only the formalism, but necessarily also the structure of the concepts of two complete and mathematically well-defined models: if one model is entirely self-consistent and describes all the relevant aspects of the world, then so must the other model. And so duality becomes physical equivalence. Thus, in other words, we are really talking about different formulations of a *single theory*.

Let me spell out the (sufficient) condition, suggested by this discussion, for a theory to admit an internal interpretation, since it will be important in §1.4. A theory T in a domain D of a possible world W is *unextendable* iff:

(i) T is a complete theory in the physical domain D at W;

(ii) There is no other theory T'' for the *same* possible world W and domain D, such that: for some T' isomorphic to T, $T' \subset T''$ (proper inclusion).[24]

Remember that the notion of completeness of a theory, in (i), was introduced in §1.1.1: as well-defined, consistent, and encompassing all the empirical data in a certain domain D at W. Condition (ii), in addition, requires that there is no extension of the theory at W: or, in other words, the theory already describes all the physical aspects of the relevant domain at W. Since the relation of isomorphism in (ii) is formal, (ii) is a sort of 'meshing' condition between (i)—or, more generally, between the idea of "not being extendable"—and the formal relation of isomorphism.[25]

I have concluded that, on an internal interpretation, there is no distinction of content between two dual theories. In §1.4, I will further disentangle three different *purposes* for which the distinction between two dual theories is irrelevant: viz. logical, empirical,

[24] T' is a fiducial theory that may well be *identical* to T. But in general, it may be the case that $T \subset T''$ is not true but $T \cong T' \subset T''$ is. In other words, $T \subset T''$ may only be true up to isomorphism.

[25] I have argued that unextendability is a sufficient, though not a necessary, condition for the coherence of an internal interpretation. The condition is not necessary because one can envisage a theory (such as general relativity without matter) receiving an internal interpretation (e.g. points being identified under any active diffeomorphism, as in the hole argument). This interpretation does not change when we couple the theory to matter fields: and I will say that such an internal interpretation is *robust* against extensions. If all possible extensions of a theory preserve an internal interpretation, then such an interpretation is justified. If the extensions suggest diverging interpretations, then we will have to specify the domain of the extension before we are justified in interpreting the theory internally.

and ontological purposes of the kind I have so far discussed. This is not to deny that there are other significant—metaphysical, epistemic, and pragmatic—purposes or uses of physical theories, for which the differences are significant.[26] For instance, one of the main pragmatic virtues of gauge/gravity dualities is that one theory is tractable in a regime of values of the parameters where the other theory is intractable. This fact lies at the heart of their applicability in real systems, such as heavy-ion collisions (cf. e.g. Ammon and Erdmenger (2015: §§14.1.2, 14.2)).

1.4 Comparison with Glymour's notion of equivalence

How does the conclusion, that the two theories related by gauge/gravity duality admit internal interpretations, and that in such cases duality implies physical equivalence, compare with the relevant philosophical literature on equivalence of physical theories? To discuss this, I will recall the usual strategy by which, faced with apparently equivalent theories, physicists try to break the equivalence; and relate this to an influential discussion, of Glymour's. I will agree with Glymour's verdicts for his examples but I will argue that this depends on the theories in the examples being *extendable*.

It is a commonplace of the philosophy of science that, confronted with theoretically inequivalent, but empirically equivalent theories, physicists naturally imagine resorting to some adjacent piece of physics which will enable them to confirm or disconfirm one of the two theories as against the other. The classic case is: confronted with differing identifications of a state of rest in Newtonian mechanics, Maxwell proposes a measurement of the speed of light. There is a parallel for gauge/gravity dualities: when two theories are both theoretically and empirically equivalent, we can still argue, by an extension or by a resort to some adjacent piece of physics, for their physical inequivalence. This is articulated in the contrast, in §1.3.2-§1.3.3, between extendable and unextendable theories.

Glymour's (1977) discussion of equivalence of theories uses the syntactic conception of theory as a set of sentences closed under deducibility. He introduces the notion of 'synonymy': two theories are synonymous when they are, roughly speaking, logically equivalent. That is, there is a well-defined inter-translation between them.[27] Although my use of theories as triples puts me closer to the semantic conception of theory (the syntactic conception's traditional rival), in fact Glymour's criterion of synonymy meshes well with my notion of a duality, construed as an isomorphism of triples. One considers the set of well-formed sentences built from two triples T_1 and T_2, e.g. statements of the type "the value of the quantity $Q_1 \in \mathcal{Q}$ (resp. Q_2), in such and such state, is such and such". Duality then amounts to isomorphism between two such sets of sentences. And this is a case of synonymy in Glymour's sense.

But does this immediately lead to physical equivalence? No. And the reasons pro-

[26] Ismael and van Fraassen (2003: §6.1) point out how, for some metaphysical and epistemic purposes, the Principle of Recombination indeed makes the distinction relevant.

[27] Technically, what is required is a common definitional extension. Barrett and Halvorson (2015: §4, Theorems 1 & 2) show that Glymour's 'synonymy', i.e. there being a common definitional extension, is equivalent to an amendment of Quine's 'translatability'. The amendment is in the notion of 'inter-translatability', which amounts, roughly speaking, to the existence of an isomorphism between two theories, on a syntactic conception of 'theory'. See §2.1.

vided in §1.3.3 are similar to the ones Glymour gives. He envisages theories that are both synonymous—they make the same empirical predictions—and logically equivalent, in the sense just described: yet are not physically equivalent. Recall Glymour's thought experiment (p. 237):

"Hans one day announces that he has an alternative theory which is absolutely as good as Newtonian theory, and there is no reason to prefer Newton's theory to his. According to his theory, there are two distinct quantities, gorce and morce; the sum of gorce and morce acts exactly as Newtonian force does."

Glymour denies that the Newtonian 'force theory' and Hans' 'gorce-and-morce theory' are physically equivalent. He argues that they are empirically equally adequate, but not equally well *tested*. His reasons for this are, partly, ontological ("I am, I admit, in the grip of a philosophical theory", p. 237), and his ontology leads him to prefer the Newtonian theory: the gorce-and-morce theory contains two quantities rather than one, but there is no evidence for the existence of that additional quantity. The argument is from parsimony: he prefers a sparse ontology.

I agree with Glymour about this verdict, in so far as one is concerned with theories that admit external interpretations, and this for two reasons:

(a) His examples deal with classical spacetime theories, i.e. *extendable* theories admitting external interpretations, which can indeed vary widely (e.g. the *same* T_{HO} can be interpreted in terms of either 'Rufus', or 'Kadee', depending on the context: cf. §1.3.2).

(b) The force theory and the gorce-and-morce theory are empirically equivalent on a restricted domain, but their extensions are *not*: "To test these hypotheses, the theory must be expanded still further, and in such a way as to make the universal force term [read instead: 'gorce'] determinable" (ibid, p. 248). (On this point, see the remarks in the second to last paragraph in this subsection, on effective theories.)

But, as I argued before: in cases in which the theory already contains all the physics it can and should contain—in case the theory is unextendable—such extensions are simply not given and the inequivalence does not follow. In such a case, no further relevant theory construction could tell force apart from gorce and morce. The latter phrase then surely does not refer to anything independent and distinct from what is meant by 'force', and the two theories *are* physically equivalent. In other words: on an internal interpretation of a theory, Glymour synonymy leads to physical equivalence.

Of course, these arguments have an ontological component. On an external interpretation, we assumed that we already knew which terms in each sentence referred to *some* things in the world (perhaps without yet knowing *which* things). Hans' theory was interpreted as saying that *two* things exist instead of just one, and this implied the inequivalence of the two theories. To explain how this was possible, given that the theories were Glymour-synonymous, one envisaged extending the theory, thus giving an independent account of what these terms refer to: an account of what the existence of these two things would imply, upon formulation of the theory on a larger range of validity within its domain. Thus, the difference between the two theories was indeed ontological: while Newton postulates one quantity, Hans postulates two.

But on an internal interpretation we cannot assume we possess an account of what 'force' and 'gorce and morce' mean, from outside the theory. The impossibility of an extension, therefore the lack of an independent account of what those terms mean, implies

that we should not make such ontological claims *independently of* the equivalence of the two theories. Because the two theories are Glymour-synonymous, and there is no extension, they are also physically equivalent: and so we are not committed to two quantities but just one. Thus, in this case, the verdicts turn out to be the same on an external and on an internal interpretation[28]. I will now explain how the failure of the unextendability condition (cf. (a)-(b) above) makes the application of an *internal* interpretation problematic for this theory.

Notice that considering extensions of theories is not some purely theoretical, or philosophical, ideal: it is, according to the perspective of modern QFT, a basic desideratum of any serious theory. The breakdown of Newtonian mechanics at short distances should be seen as an indication that it is an *effective theory*. Effective field theories (see Weinberg (1996: I, p. 523ff; II, p. 154)) are theories that are accurate for phenomena in some range of (usually low) energies, but are corrected by higher-order terms in the Hamiltonian which are relevant at high energies.

Beside, there is a more specific relation to the gorce-and-morce proposal. Unless there is an exact symmetry given as part of the theoretical principles in 1.1.2.b, the introduction of new fields will generically introduce higher-order terms which break the seemingly symmetrical way in which those fields appear in the low-energy Hamiltonian. Thus if gorce and morce are indeed distinct fields,[29] most high-energy theories which reduce to the gorce-and-more theory at low energies, will treat gorce and morce differently. They have different interactions (unless an exact symmetry protects them). Thus the framework of effective field theories promises to satisfy Glymour's demand of parsimony, that there should in principle be a way to determine the values of distinct quantities.

Thus, the force and the gorce-and-morce theories are *generically* not physically equivalent, even though they are Glymour-synonymous. For there are very many possibilities for extension to high energies, only some of which lead to theories equivalent to the Newtonian theory's own extension: most of them do *not*. The physical equivalence with the Newtonian force theory can thus only be established if additional requirements are imposed, such as the stipulation of a particular extension of the theory, or a 'protecting' symmetry.

To sum up: Glymour's remarks are concerned with extendable theories—and such theories only admit external interpretations. The case for the physical equivalence of extendable theories is inconclusive unless an extension is stipulated. To secure physical equivalence without such a stipulation, one needs an internal interpretation, and hence two unex-

[28] I am here following Glymour's tacit assumption that one's formulation of the theory is sufficiently perspicuous (e.g. in terms of a triple), so that the physical quantities can be read off from it. If this is not so, it might be equally natural to say that there is no fact of the matter about whether one is committed to one or two quantities—or that such facts are underdetermined by the computation of the relevant physical quantities. In other words, the argument assumes that one *is* dealing with physical quantities, so that the discussion of the theory's ontological commitments makes sense. Cf. the discussion, just below, of effective field theory.

[29] The sum of gorce and morce is the derivative of the sum of two potentials. I envisage these two potentials as pertaining to distinct fields—since Hans declares gorce and morce to be distinct.

tendable theories that are valid for all ranges of the parameters. In physics jargon: such theories are well-defined non-peturbatively and they are not coupled to anything else. For such theories, Barrett and Halvorson's (2015) notion of theoretical equivalence (cf. footnote 27) amounts to physical equivalence.

2 Spacetime eliminated?

Gauge/gravity dualities relate $(d+1)$-dimensional theories (models!) of quantum gravity to d-dimensional quantum field theories (QFT models) with gauge symmetries (hence the name 'gauge/gravity'). Given this one-line description of the duality, one might be tempted to answer the question in the title of this Section in the negative: No, gauge/gravity dualities do not eliminate spacetime as fundamental structure. The suggestion would be that: (1) the duality sets the two models on an equal footing; (2) both models seem to be formulated in terms of space and time; (3) hence, whichever model one likes to choose, duality does *not* eliminate spacetime.

Yet the suggestion is too hasty: it misses the significance of dualities in physics as opposed to, say, mathematics. The aim of this section is to elaborate on a correction of this hasty suggestion.

In order to know whether spacetime is eliminated by a gauge/gravity duality, we have to analyse: (i) what is it that the two models in question say? And: (ii) does this saying include 'spacetime'? In §2.1, I argue that what two dual unextendable theories say, is the content they have in common: the *common core*. In §2.2, I argue that this common core includes some, but not all, spacetime structure.

2.1 What do the theories say?

Even if one is not entirely convinced by the arguments on the internal interpretation of unextendable theories in §1.3.3-§1.4, let us suppose that we *do* have a gauge/gravity duality, that can be interpreted as physical equivalence, between a model of gravity in $(d+1)$ dimensions, and a gravity-free model of a QFT in d dimensions. What might that imply for what the theory says?

Consider the following analogy with natural language. Most languages are not perfectly translatable, sentence by sentence. Understanding particular sentences requires extra-linguistic facts such as gestures or facial expressions, or particular knowledge that not every speaker of another language will possess. So a complete translation of such a sentence is not possible because it may require an *explanation*, which when given in place of the original sentence, distorts key communicative aspects (the sharpness of a joke, the specific emotion expressed). But let us imagine for a moment that there are such things as universal languages: languages which can express every human thought and emotion (i.e. they are unextendable!). Then any two such languages would be inter-translatable—there would be no idiosyncracies which cannot be expressed in any of the languages. Any complete sentence would therefore be translatable between such languages. But the individual words need not be translatable, since their meaning may depend on the context.

Full sentences would be the minimal semantic unit.[30]

An obvious objection to the above may come to mind: when each language is endowed with a *context* (i.e. the facial expressions, ostensive gesturing etc., mentioned above), then any two languages *are* pairwise translatable. Even admitting to the possibile truth of this objection: such extra-linguistic elements are absent in our case: for, by assumption (see the preamble of §2), our theories are *unextendable*, i.e. there is no relevant extra-theoretic context that can be added to them. The reason for the absence of such a context is that in the case of unextendable theories we seek an internal interpretation, rather than an external one based on considerations of context or use, i.e. from outside the theory. In other words, our imaginary agents are to communicate using language and language only.

Back to gauge/gravity duality. In the language analogy, only those things that have a fixed meaning (full sentences) can be translated. This suggests that, in the cases under consideration, the only things that the theories, given the duality between their models, say about the world are the things which are inter-translatable: states, quantities, and their valuations. Those are the things that have a unique physical meaning. They are what I called, at the beginning of this subsection, the 'common core' of the theory. They defined a theory in the first place. I was able to define a theory in this way because these are indeed the things that get an internal interpretation. This is the same reason why internally interpreted dualities gave us physical equivalence in §1.3: the theories were equivalent because *all* they said was in the internal interpretation of the triple T, common to the two models. Interpreting the models did not entail embedding them into some larger theory (see §1.3.1).

2.2 Does what the theories say include 'spacetime'?

Suppose, as we are assuming in this subsection, that we have a duality that amounts to physical equivalence. Thus we have an isomorphism between the two Hilbert spaces and between the physical quantities—so we are in fact considering unitary equivalence. And it doesn't matter which of the two models we choose, even though each representation has its own, very different, intuitive picture of the world (i.e. its own external interpretation, which is based on how the model is normally used, independently of the duality). One model says the dimensionality of spacetime is $d + 1$, the other says it is d; one model has gravity, the other does not. Our question is whether there is a *common core* to the two models and whether this common core is spatio-temporal. That common core, once interpreted on an internal interpretation, is what the theory says about the world (§2.1), i.e. the interpreted triple T, which takes only the equivalent facts under unitary equivalence as physical.

It turns out that there *is* such a *common core*: a d-dimensional spacetime $\mathcal{M}$ whose

[30]For the sake of the analogy with physical theories, I exclude cases in which only *entire texts* can be translated, as irrelevant to the analogy. One may also worry that the linguistic criterion discussed in this subsection may differ from Glymour's definitional extensions, discussed in §1.4. However, Barrett and Halvorson (2015: Theorems 1 and 2) show that the inter-translatability criterion (two-way translatability) and Glymour's criterion are equivalent in first-order logic (cf. footnote 27). Of course, it is a long way from first-order logic to physical theories of the kind here discussed: so in the current context, one should take the linguistic criterion with a pinch of salt.

metric is defined only up to local (spacetime-dependent) conformal transformations (De Haro et al. (2016b: §§6.1.2-6.1.3)), i.e. a conformal manifold. This works as the 'core' theory in the following sense. We examine the structures of the two models under duality: the structure which is mapped by duality will be the content of the triple, which arises as the common core of the two models. Actually, the duality map itself, as sometimes presented in the literature (see especially the formulation in De Haro et al. (2016: §4.2)), already makes explicit what the states and operators of the two models are, and how they map across models. So let us look at the two models and extract their common structure.

On the gravity side, one evaluates the path integral over all metrics and topologies with given *boundary conditions*: this determines the *state* in $\mathcal{H}$.[31] I will now discuss how these two boundary conditions are interpreted in a boundary formulation of the model, which will turn out to be equivalent to the CFT formulation (see the third paragraph below).

There are two boundary conditions to consider. The first is an asymptotic boundary condition on the form of metric, which is itself determined only up to a conformal factor; in other words, one needs to specify a conformal manifold $\mathcal{M}$ together with a conformal class of metrics, denoted $[g]$, induced at the boundary. Thus, regardless of the chosen class $[g]$, the asymptotic symmetry algebra is the conformal d-dimensional algebra, and the representations of this algebra form the states of $\mathcal{H}$. (That this is really a state in the state space $\mathcal{H}$ of the theory will become clear when we see that, on the gauge model side, we have the *same* conformal structure and algebra. See the next paragraph. In the case considered in this subsection, of matter fields set to zero, we only have a subalgebra of the full superconformal algebra, hence only a subset of the states of $\mathcal{H}$. See footnote 31.)

Second, a boundary condition needs to be imposed on the asymptotic value of the canonical momentum Π_g conjugate to the metric induced on the boundary, on the given state.[32] This choice determines a subset of states of the conformal algebra (for instance, $\langle s| \Pi_g |s\rangle = 0$ is a boundary condition that fully preserves conformal symmetry, so that the states $s \in \mathcal{H}$ are representations of the d-dimensional conformal algebra; other choices will break conformal symmetry and thus further constrain the set of states). Thus the latter boundary condition is interpreted as a choice of a *subset of states* in the Hilbert space of the theory, whereas the former corresponds to a choice of a *source* which is turned on for the canonical momentum Π_g, thereby generating new states. We will write the resulting states as $|s\rangle_{\mathcal{M},[g]} \in \mathcal{H}$, where s is the state, modified by the addition of a source $[g]$ on $\mathcal{M}$ coupling

[31] For simplicity, I am now considering the case of quantum gravity without matter. This restricts our considerations to a class of states with zero helicity. Adding matter can be done, and does not affect the philosophical conclusions, but would involve some cumbersome qualifications—so I concentrate on the matter-free case. (Briefly: adding matter on the gravity side corresponds to specific deformations of the gauge theory Lagrangian, see e.g. Ammon and Erdmenger (2015: §5.3)). Also, in the rest of this Section (except for one example with $d = 4$), I restrict the discussion to the case that d is odd, for technical reasons. Also, I am considering Euclidean signature and solutions that are regular throughout the $(d + 1)$-dimensional space, which in most cases determines the solution uniquely: see e.g. De Haro (2009: §2.1), though I am not aware of a general proof. In Lorentzian signature, one needs, in addition, to specify other boundary conditions.

[32] It is technically useful, but not always required, to impose boundary conditions for the metric *at spatial infinity*. For instance, in semi-classical treatments of the gravity side, the second boundary condition is often imposed in the deep interior of the space, rather than at the boundary.

to Π_g. The set of states corresponding to all possible boundary conditions constitutes the subset of the state space $\mathcal{H}$ which we are able to describe without introducing matter fields. The basic physical quantities are the canonical momenta $\Pi_g \in \mathcal{Q}$ conjugate to g (De Haro et al. (2016: §4.2.2.1)).

On the side of the gauge model of the theory, there is a conformal field theory on a d-dimensional manifold whose metric is defined, up to a local conformal factor, by the very form of the asymptotic metric which one gets from the gravity model: in fact, we can identify this, via duality, with the conformal manifold $\mathcal{M}$. The states are representations of the conformal symmetry algebra, the same algebra which we obtain on the gravity model side (and again, we are only considering a subsector, obtained by setting the expectation values of all operators, except the stress-energy tensor, to zero). The canonical momentum Π_g corresponds, through the duality map, to the stress-energy tensor T_{ij} of the CFT (De Haro et al. (2001: §3)): $\Pi_g \equiv T$. The stress-energy tensor is the operator from which the generators of the conformal symmetry algebra can be constructed (see e.g. Ammon et al. (2015: 3.2.3)). Thus, the two models share the d-dimensional conformal manifold $\mathcal{M}$ with its conformal class of metrics $[g]$, the conformal algebra, and the structure of operators, as claimed. The conformal algebra and class determine $\mathcal{H}$ and thereby the valuations of this important subset $\{T_{ij}\}$ of operators of $\mathcal{Q}$, namely the infinite set of correlation functions ${}_{\mathcal{M},[g]}\langle s| T_{i_1 j_1}(x_1) \cdots T_{i_n j_n}(x_n) |s\rangle_{\mathcal{M},[g]}$, for any n. This infinite set of correlation functions contains important dynamical information about the CFT.[33]

The correspondence, through the duality map, between the states and the quantities of the two models, obtained in the previous two paragraphs, justifies the appearance of these states and quantities in the common core of the duality, which is part of $\mathcal{H}$ and $\mathcal{Q}$. Of course, there is no claim here that the common core which we have identified so far gives us the *entire* triple T: only its subsector corresponding to the quantities considered above. Nor is it my intention to do so—I do not aim at a full foundational discussion of gauge/gravity duality here, but at an illustration of the concepts introduced in Section 1, and their practical application in gauge/gravity duality. For instance, we have not discussed the dynamics—though the dynamics is, in fact, not hard to get. But its detailed study would require more technical detail. (Briefly: one picks out a Hamiltonian operator, which on the gravity side is the generator of radial translations, and on the CFT side is the generator of conformal transformations. See Papadimitriou et al. (2005: §4.2)) More importantly, as mentioned in footnote 31, there is no proof that, even for pure gravity, the correlation functions of the stress-energy tensor *exhaust* the set of operators $\mathcal{Q}$: non-local operators such as Wilson loops may also be needed. Finally, the set of states $|s\rangle_{\mathcal{M},[g]} \in \mathcal{H}$ have been obtained under certain very natural assumptions about the regularity of the gravity metrics, asymptotically. Again, there is no proof that there could not be some singular states corresponding to e.g. degenerate boundary metrics in the same Hilbert space. But, as mentioned, we do not need to resolve these technical issues here: the main point being that, by considering the common core of (suitable limits of) the two models, we have obtained an important subset of the states $\mathcal{H}$ and of the quantities $\mathcal{Q}$ of the

[33] There is no claim here that it contains *all* of the information about the CFT, even for the subset of states which we are considering. Nonlocal operators such as Wilson loops (and perhaps additional states) may also be required.

theory.

The common core (as per the conception of a theory in §1.1.1 and §1.2 as an equivalence class of models) only contains those quantities which are physical in the two models. However, the *external* interpretations, which are normally given to the conformal class of metrics and to the operators, are of course completely different in the two formulations of the theory. On the gravity side, the manifold $\mathcal{M}$ is the boundary of a $(d+1)$-dimensional manifold $\hat{\mathcal{M}}$, with a metric which can vary; whereas $\mathcal{M}$ is a manifold in the QFT with a fixed metric. Also, on the gravity side the conformal class of metrics does not stand alone, but is deformed to give a metric on the $(d+1)$-dimensional spacetime (in the semi-classical limit: see De Haro (2016a: §2.1), also Ammon and Erdmenger (2015: §4.1.4, §5.1)). On the gauge model side, the conformal class gives the class of geometries on which the QFT is defined. Quantum mechanically, the d-dimensional conformal class of metrics provides an arbitrary asymptotic boundary condition for the path integral of the quantum gravity model. And the operators are externally interpreted as follows: Π_g (in its semi-classical limit) is the *gravitational*, quasi-local stress-energy tensor at the boundary; whereas the corresponding operator T_{ij} in the CFT is the stress-energy tensor for the *matter* content of the CFT.

The above discussion has an important consequence. We are considering gauge/gravity dualities with *pure gravity*, and no matter fields, in the gravity model of the theory. The QFT, of course, *does* have matter fields. But the *specific set* of matter fields is *not* part of the invariant core. The description of T_{ij} as the 'stress-energy momentum *for a specific set of matter fields of the QFT*' is thus not part of the common core. Only the stress-energy tensor T_{ij} is part of the common core, even though in the gravity model of the theory it figures (under an external interpretation) as a 'gravitational' tensor, while in the boundary model it describes 'matter'. Thus, the qualifications 'gravity' or 'matter' are just parts of our *descriptions* of the theory.

Of course, the challenge is now to work out the *internal interpretation* of the bare theory—the common core—independently of the two external interpretations, i.e. work out an interpretation for $|s\rangle_{\mathcal{M},[g]} \in \mathcal{H}$ and $T_{ij} \in \mathcal{Q}$, which are part of the *bare theory* $T = \langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$, together with the conformal symmetry. I will not do this in detail in this paper: since my aim here is to illustrate the content of the bare theory. But the following elements of an internal interpretation can already be worked out. Indeed, the fact that, in both models, the pair $(\mathcal{M}, [g])$ is a conformal manifold with a conformal class of metrics, and T_{ij} describes stress, energy and momentum, suggests that $(\mathcal{M}, [g])$ and T_{ij} are interpreted internally as representing, respectively: (i) a conformal manifold with a conformal class of metrics, and (ii) stress, energy and momentum. There may be other aspects to the internal interpretation of $[g]$ and T_{ij} but (i) and (ii) certainly seem to be important elements of it!

Having discussed the states and quantities in the theory T, let me now make some comments about its *symmetries*. Clearly, the conformal symmetry is a symmetry of the theory, in the sense of §1.1.1, because it is the symmetry that classifies the states and the quantities. Different classes of states are given by different representations of the symmetry algebra. This symmetry is also a symmetry of the conformal manifold with its conformal structure, and it acts non-trivially on the stress-energy tensor (the vanishing of the trace of the expectation value of the stress-energy tensor on any state being a direct

consequence of this symmetry). De Haro (2016a: §3) gives a detailed argument why the conformal symmetry is (Visible), in the sense introduced in §1.2. In this subsection I have argued that the conformal algebra is also (Commutative), since it is represented in both models, in two different ways which I have discussed above. Thus, conformal symmetry is part of (SymT) [version 2] in §1.2. On the other hand, De Haro (2016a: §3.1) argues that an important class of the general diffeomorphism invariance of the gravity model is *trivially represented* on the theory T, so it is QFT-(Invisible) in the sense of §1.2, and hence it is part of (PSymM) [version 2] in §1.2. This analysis implies that, once an internal interpretation is adopted, the conformal symmetry is a physical symmetry, of the type (Physical) discussed in 1.1.2.b; while the remaining diffeomorphism symmetry is (Redundant).

As just mentioned, De Haro (2016a) shows that a subgroup of the diffeomorphism group of the gravity model is (Visible) under the duality map and corresponds to the group of conformal transformations of the theory. Other diffeomorphisms of the gravity model are QFT-(Invisible), and map trivially under the duality. This analysis generalises to other symmetries: in general, a part of the symmetries will be (Visible) to the theory and hence belong to (SymT); while the remaining symmetries will be simply redundancies of one of the models, being trivially represented on the theory, hence (Invisible) relative to the other models, and they belong to (PSymM).

For this reason, 'bag-type' symmetry concepts, such as e.g. 'gauge symmetry', cannot distinguish between symmetries which are (SymT) and symmetries which are (PSymM): let alone symmetries which are (Redundant) and symmetries which are (Physical). An analysis along the lines of De Haro (2016: §3), distinguishing the (Visible) from the (Invisible) for the specific symmetry at hand, must be carried out before one can conclude that a symmetry belongs to the model or belongs to the theory. Typically, 'bag-type' concepts, such as 'gauge symmetry', even specific ones (e.g. local U(1) gauge symmetry in the gravity theory) are partly (SymT) and partly (PSymM).

Let us now discuss the common core and the surplus representative structure further, in another example, of $d = 4$, i.e. the QFT is 4-dimensional and the gravity model is 5-dimensional. Specifically, let us take a look at the symmetries. The four-dimensional gauge theory[34] has an internal SU(N) gauge symmetry (this is the theory's gauge symmetry group) that is Gravity-(Invisible), in the sense of §1.2. The theory is formulated so that this symmetry is explicit—the states $\mathcal{H}$ and the observables $\mathcal{Q}$ are invariant under it. Since the common core only contains states and quantities constructed from the triple $T = \langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$, which are invariant under gauge symmetry, this gauge symmetry does not belong to the common core: it lies in the set (PSymM).

There is thus surprisingly little invariant under the duality between these two theories, yet they describe all that is physical about the theories. The $(d+1)$-dimensional manifold $\hat{\mathcal{M}}$, most of the diffeomorphism group (viz. diffeomorphisms that do not generate transformations of the boundary metric), gauge symmetries: in the present context, these are apparently all part of our description, rather than facts of nature. So are many global aspects of the topology of a spacetime (i.e. of one model's description, but not of the

[34]The theory in question is a 'super Yang-Mills theory', a specific supersymmetric variant of Yang-Mills theory (Ammon and Erdmenger (2015: §3.3.6)), but the details are irrelevant here.

other's), such as the number of dimensions. Similarly also the concepts of vector fields, tensor fields, Lie groups, differential geometry: though we use them to formulate our theories, each such a concept is not part of nature, at least not part of the common core of the models of a gauge/gravity duality. For example, tensor quantities in $d+1$ dimensions do *not* map to tensor quantities in d dimensions under duality, and so do not belong to $\mathcal{Q}$. Rather, such quantities are part of the surplus representational structure of one of the models. Nor does a sharp classical-quantum divide seem to be upheld, since one-loop effects such as anomalies can be mapped to anomalies of classically defined quantities.

Thus, the answer to the question we posed in the title of this Section requires careful articulation, as follows. I state it in terms of what is eliminated on the gravity side:

(i) In the gravity model of the theory, most of the spacetime structure is eliminated: the entire interior region of the manifold $\hat{\mathcal{M}}$, on which a gravity model is defined (including its topology), is eliminated.

(ii) All that remains is the *asymptotic* boundary manifold $(\mathcal{M}, [g])$ with its conformal structure, which plays the role of an asymptotic boundary condition for the equations of motion, and so is arbitrary but fixed; the states $|s\rangle_{\mathcal{M},[g]} \in \mathcal{H}$, and the stress-energy tensor $T_{ij} \in \mathcal{Q}$ at spacelike infinity. So our theory T is made up of, in any case, of the common core just identified. This common core is furthermore endowed with an action of the conformal group, which is a (SymT). So, in particular:

(iii) All local gravitational structure has been eliminated: there are no *local* dynamical gravitational degrees of freedom left. This agrees with the general expectation in the quantum gravity literature.

(iv) It is, of course, already the case within general relativity that some of this spacetime structure is non-invariant: spacetime points in $\hat{\mathcal{M}}$ are not the invariant objects, being rather equivalence classes that are orbits of the relevant active diffeomorphisms. But here we find that under diffeomorphisms that reduce to a conformal transformation at the boundary, only a conformal manifold $(\mathcal{M}, [g])$ with its conformal class of metrics, which provides the boundary state $|s\rangle_{\mathcal{M},[g]} \in \mathcal{H}$ for the path integral, is left, in addition to the stress-energy tensor $T_{ij} \in \mathcal{Q}$ at spacelike infinity.[35]

To summarize: at the end of the day, most, but not all, of the spacetime structure (as specified in (iv)) has been eliminated.

3 What are the broader implications of duality?

In this Section, I consider the implications of gauge/gravity dualities for: (i) theory construction (§3.1) and (ii) metaphysical accounts of the world (§3.2).

[35]The remaining question is: *should* and, indeed, *can* this remaining lower-dimensional structure be eliminated? This question has been taken up by De Haro (2015: §§2.3.2-2.3.4), but I cannot consider it here.

3.1 Implications for theory construction: support for holography

To see what gauge/gravity duality might imply for theory construction, let us first review the invariant content of this duality. For this invariant content counts as the duality's 'common core', the set of statements which are independent of a specific model: they are what the theory says (see the discussion in §2). For pure gravity without matter, this common core was identified as $|s\rangle_{\mathcal{M},[g]} \in \mathcal{H}$ and $T_{ij} \in \mathcal{Q}$: a d-dimensional conformal manifold $(\mathcal{M}, [g])$ with its conformal class of metrics (§2.2, (iv)), the corresponding states as representations of the conformal algebra, and the stress-energy tensor. The theory comes with an action of the conformal group, which is a (SymT). General diffeomorphism symmetry is not part of the common core because it is not part of the gauge model's formulation, though a subgroup of it is—namely, the subgroup that induces conformal transformations at infinity. I emphasise that this conclusion—that most diffeomorphisms are *not* part of the common core—does not change if we analyse other examples of gauge/gravity dualities.

I believe that here may lie the main contribution of dualities to our understanding of the world: a duality can be used as a tool for figuring out, in concrete terms, the common core of a set of models, thus giving us the theory.

What positive lessons can one draw from gauge/gravity dualities for *other theories of quantum gravity*? That is, if one were to quantise another theory similarly to how gauge/gravity duality quantises gravity (on its gravity side), what would one do?

Let us consider how the Hilbert space and the physical quantities are obtained from the *gravity* model of the theory: we find that physical quantities are of codimension one and that Hilbert spaces are associated with boundaries. It has been known for a long time that the physical quantities of gravity could not be associated with points in spacetime, which have no intrinsic meaning because they are moved around by active diffeomorphisms. In classical general relativity, the received view is that physical quantities satisfying all the necessary requirements are very rare and almost exclusively limited to the simplest ones: ADM energy, angular momentum, and charge. But these are hardly enough to describe an entire dynamical universe.

Gauge/gravity duality suggests that physical quantities are indeed associated with global properties of the spacetime—in particular, with boundaries both in time and in space. Thus these quantities are global (better, and usually, called 'quasi-local') in the way expected for a theory of gravity, and they also have a well-defined classical approximation. But in general, in any quantum theory, physical quantities are operator insertions in correlation functions that are calculated between two states. Therefore, to obtain quantities which are independent of the spacetime points, one integrates the insertions over a (maybe) smeared spacetime region.

This discussion vindicates a suggestion that 't Hooft (1993) made when he first proposed the *holographic principle*: he interpreted the holographic principle as entailing that quantum gravity is a topological quantum field theory: "[The holographic principle] means that, given any closed surface, we can represent all that happens inside it by degrees of freedom on this surface itself. This, one may argue, suggests that quantum gravity should be described entirely by a *topological* quantum field theory, in which all physical degrees

of freedom can be projected onto the boundary" (p. 6).

We can now see that gauge/gravity duality provides evidence for this idea. One can now clearly see how duality suggests that quantum gravity is a topological quantum field theory. For the duality indeed assigns a Hilbert space to each connected piece of the d-dimensional boundary of a $(d+1)$-dimensional spacetime, for which one sums over geometries and topologies. Category theory seems the natural language in which to formulate such a statement (for an application of category theory in two-dimensional topological quantum field theories, see e.g. Koch (2003)).

This discussion raises another issue: namely, that for quantum gravity we may need new mathematics. As we have seen, gauge/gravity dualities have significant implications for theory construction. Namely, though various mathematical techniques (as found in differential geometry, topology, and Lie groups) play some role in the common core defined by the duality: the dualities, in fact, map between models in which these structures take very different forms (§2.2). And so, describing the common core of dual models in appropriate language may well require the development of new mathematics.

Remember the case of mirror symmetry (one of the prime examples of a duality in string theory), where complex geometry and symplectic geometry were related in a way that mathematicians found both unexpected and (at first) unbelievable. Atiyah (2007: p. 83) describes it as follows:

"It is a spectacular coup: physicists go up into the sky, they land by parachute in the middle of algebraic geometers and they capture immediately the whole city. The discovery of mirror symmetry is certainly one of the most remarkable developments of the last part of the 20th century. It provided an example of two different classical theories, two different algebraic varieties, giving rise to the same quantum theory, and with spectacular applications."

Whether gauge/gravity dualities will turn out to have similar spectacular applications in mathematics is yet to be seen. But arriving at a mathematically rigorous formulation of them is likely to uncover new mathematical structures.

3.2 Metaphysical implications

I will restrict my comments to two main implications of gauge/gravity dualities for discussions in philosophy of science and metaphysics. Both remarks concern the relation of theories of quantum gravity to the notion of physical equivalence.

The first comment proceeds from physical equivalence to the interpretation of physical theories. As I argued in §1.3.1, in cases where a duality supports an internal interpretation, duality amounts to physical equivalence. Recognising that two models are physically equivalent can thus function as the first step to making precise an internal interpretation. Namely, the duality yields a common core that is identified as what the theory says (§2.2). This common core can itself be presented as a triple $T = \langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$, together with its syntax: it is what the two models have in common. And so, the physical interpretation of a quantum gravity theory, and in particular the articulation of the ontology that may underlie such a theory, starts from that common core.

The second comment is in a different direction: from the formalism to physical equivalence. Duality can be found in both unextendable and extendable theories. Gauge/gravity

dualities between string or M theory and QFTs are presumed to be exact dualities between unextendable theories, which in particular should have non-perturbative formulations.[36] Similarly, position-momentum duality in elementary quantum mechanics (with respect to a simple possible world, containing only the given quantum system) is a duality between *unextendable* models, i.e. equivalent formulations of a single theory: the two formulations are therefore also physically equivalent.

Other dualities, on the other hand, are also exactly defined, but only on theories with limited regimes of applicability. An instance of the latter case is Section §1.4's example of the Newtonian force theory and Hans' gorce and morce theory. Such theories are extendable: they break down beyond some regime of applicability inside their natural domain D at world W, as argued in detail in that Section. But, as discussed in footnote 25, in the case of extendable theories we are not always justified in believing internal interpretations, because an extension might always modify that interpretation[37], as in the example of Rufus vs. Kadee: a coherent interpretation is always dependent on how the theory is extended. In such cases, in which the theories are extendable and we do not know whether they are robust, we are not justified in believing that the theories are in a true relation of physical equivalence. We cannot a priori decide on the physical equivalence of such theories. The effective field theory perspective in §1.4 suggested that such theories are generically *not* physically equivalent.

In the light of duality, Glymour's position is best restated as saying that we are not justified in taking extendable theories with different formulations as physically equivalent, since they may have different ontological commitments. On the other hand, physical equivalence for *unextendable* theories makes sense. Such theories are highly constrained, valid for all values of the parameters, and cannot be coupled to any adjacent physics while preserving their theoretical equivalence.

This situation may be seen as an important contribution of the search for a theory of quantum gravity to discussions of physical equivalence. It also underlines the importance for philosophy of unextendable theories. Extendable theories abound: we have discussed Newtonian mechanics and effective QFTs. There is also general relativity, with its singularities due to gravitational collapse. Unextendable theories are rare but useful, and we do have some good examples of them: several conformal field theories, topological quantum field theories (some of which play a role in gauge/gravity dualities: Chern-Simons theories, various versions of Yang-Mills theory, WZW models), and topological string theories (which describe subsectors of string theories). They should be important case studies for the philosophical topics of theoretical and physical equivalence.

Finally, note that unextendable theories need not be 'finished' theories. Some of the examples above (especially some two-dimensional conformal field theories) are indeed understood with rigorous mathematics; but other examples (such as Chern-Simons theory and Yang-Mills theory: see also footnote 36), though expected to be unextendable for good mathematical reasons, are still "theory fragments", in the sense of Huggett and Wüthrich (2013: p. 284): the programme of "interpreting a theory 'from above', of explicating the

[36] We are proceeding on the plausible assumption that at least some of the known gauge/gravity dualities are exact. There is circumstancial evidence for this but as yet no mathematical proof.

[37] Unless one can show that the interpretation is robust: cf. footnote 25.

empirical significance of a theory, is both 'philosophical' in the sense that it requires the analysis of concepts, and crucial to every previous advance in fundamental physics... As such, it must be pursued by the study of theory fragments, toy models and false theories capturing some promising ideas, asking how empirical spacetime relates to them". So also for the theories mentioned above: completely rigorous mathematical proofs are still lacking, even if these fragments are robust enough that they already contributed to a Fields medal (for E. Witten: in 1990).[38]

4 Comparing with recent work on dualities

Let us take stock of the distinctions we have made so far: by discussing how they relate to extant philosophical discussions of dualities in the literature.

Our main contrast was between theoretical and physical equivalence: dualities are cases of theoretical equivalence, and as such they do not always have to relate to physically equivalent theories. This is, of course, not a new theme in the philosophy of physics literature: and the recent philosophy of dualities, in particular, has addressed it. In §4.1, I will describe recent work on dualities and state its limitations when it comes to clarifying the contrast just mentioned. In §4.2, I will discuss more closely Huggett's (2016) recent work on T-duality. In §4.3, I will discuss some aspects of Rickles' (2016) account of dualities as gauge symmetries, and of Fraser's (2016) comparison of Euclidean field theory and QFT.

4.1 Some recent work on dualities

Recent work on dualities engaging with the constrast between theoretical and physical equivalence includes, for example, Matsubara (2013: p. 485). He stresses "the importance of distinguishing between the mathematical formalism itself and the physical interpretation". Similarly, Dieks et al. (2015: §3.3.2) use the contrast of internal vs. external viewpoints to distinguish cases in which duality leads to an *identification* of two theories, on the internal viewpoint (cases of physical equivalence) from cases where, despite the perfect *formal* agreement of the two theories, they are *in*equivalent: on the external viewpoint.[39]

In a forthcoming special issue on dualities edited by Castellani and Rickles, several authors engage with the distinction between formal and physical equivalence[40]: De Haro (2016: §2.4) further develops the internal and external points of view, introducing the requirement of internal consistency and the concept of a 'theory of the whole world' for the internal viewpoint. Huggett's (2016: §1) "first interpretive decision: either the T-duals agree on the physical world or they do not" is the question whether two theories, which

[38] Atiyah (1990) mentions, in the medal citation, Witten's contribution to topological quantum field theory.

[39] Dieks et al. (2016: §3.3.2) spoke of external and internal 'viewpoints'. In this paper, I have promoted these to the status of 'interpretations': for I have given an account, in §1.1, of what bare theories and interpreted theories are. I will continue to use 'interpretation' in what follows.

[40] I discuss these works in chronological order of appearance.

are related by T-duality, are physically equivalent. Fraser (2016: §3) also distinguishes predictive, formal, and physical equivalence, and compares Euclidean field theories to QFTs: two theories can be predictively and formally equivalent yet fail to be physically equivalent. Also Rickles (2016: §2) distinguishes interpretive, formulational and theoretical underdetermination—with the resulting change in meaning because of the word 'underdetermination' rather than equivalence. But Rickles, too, states that "dual theories are simply examples of theoretically equivalent descriptions of the same underlying physical content" (2016: Abstract).[41] Castellani (2016: §3.3), from the perspective of the elementary particle vs. soliton distinction, contrasts representational/functional democracy with ontological democracy, where this 'democracy' means, roughly, the absence of a hierarchy or fundamentality relation. De Haro et al. (2016: §3.2) emphasise "that usually there are many *token* systems of the type treated by a theory", therefore recognising "that there are cases of two disjoint parts of reality... that: match exactly, 'are isomorphic', in the taxonomy used by some theory... but are otherwise different." And they also discuss 'theories of the universe' as one case in which the idea of distinct but isomorphic existences falls by the wayside, thus making room for the internal viewpoint and justifying the verdict of physical equivalence where the latter is appropriate.

Clearly, the distinction between theoretical and physical equivalence is central to the above authors. But much as I admire them, I submit that the extant accounts are insufficient, for two reasons, which I state here and develop in §4.2 and §4.3:

(i) Despite the fact that some of these authors distinguish theoretical and physical equivalence, they do not provide a satisfactory explanation of, or clear criteria for, when what I have called 'theoretical equivalence' amounts to 'physical equivalence', and when it doesn't: they fail to give an analogue of my scheme of contrasts (a)-(d) in the Introduction, as I will argue in more detail in §§4.2-4.3.

(ii) A related problem is that some commentators, even though discussing the *external* interpretation, are too quick to dismiss it as a possible interpretation of specific dualities. As a consequence, they overlook important points: such as the difference between extendable and unextendable theories.

I will now select a few of these problems in the literature and explain how my scheme successfully deals with them. My criticism of the quoted authors will thus come down to saying that the extant accounts, *qua* accounts of dualities, are not sufficiently articulated to provide a clear difference-maker between theoretical and physical equivalence: other than in very specific cases, and using judgments that do not strictly follow from the given accounts. In their defence, one may add that it was not their purpose to look for such a clear-cut difference-maker!

[41] However, Rickles (2016) here uses the phrase 'theoretical equivalence', whereas most commentators use 'physical equivalence'. The reason for this may be in his footnote 14: "We must therefore remain purely in the realm of theoretical considerations", rather than making physical claims about the equivalence of the duality, which can be easily broken by 'external confounders', e.g. adding point particles to a theory of strings would break mirror symmetry.

4.2 Huggett on T-duality

In this subsection, I discuss Huggett (2016), whose focus seems to be closest to the ideas developed in this paper. Roughly, T-duality relates one kind of string theory in a space with a circle of radius R to another kind of string theory in a space with a circle of radius $1/R$. Given that the two theories are theoretically equivalent, in the sense used in this paper, but postulate different radii for space, the question then arises as to whether the two theories are also physically equivalent, and whether there is a fact of the matter about the radius of space.

Huggett's account gives a syntactic analysis of T-duality, which is an interesting alternative to mine, which as mentioned is closer to the semantic conception of theory (though I have explicitly compared with Glymour's account, which is syntactic, in §1.4).

Huggett (2016: §2) starts off his philosophical discussion reporting the philosophical consensus: "Commentators have been pretty uniform in taking the stance that the T-duals should indeed be taken as giving the same physical description... I argue below that they are correct." Huggett makes his first interpretive decision (that the T-duals are physically equivalent) right at the beginning, so as to then move on to his second interpretive fork: whether target space (the space postulated by the theory) and phenomenal space (a space constructed using an experiment with photons) are the same. And *after* he has decided that phenomenal space is not identical with target space, he goes back to arguing, in §2.2, that "duality is considerably stronger than 'empirical equivalence'." He uses the example of harmonic oscillator duality (which was apparently first contrasted with dualities in string theory in Zwiebach cf. (2004: pp. 377-378), also cf. Matsubara (2013: pp. 478-479)) to illustrate how "systems which have dual descriptions... are not necessarily physically equivalent". For "if the theories describe literal, concrete, physical oscillators in our world, then the two systems are not the same... and are indeed readily distinguishable, by measuring the masses, for instance: the mass on the first spring is m, that on the second $1/k$... But the case is disanalogous to string theory, if that is taken as a theory of *everything*. What happens if a duality applies to a 'total' theory, in the sense that it is the complete physical description of a world, so that there is nothing outside the theory?" (§2.2).

Huggett's position here is, in essence, the same as that in Dieks et al. (2015: §3.3.3) and De Haro (2016: §2.4) in the context of gauge/gravity dualities but also more generally for exact dualities: in some cases, there are ways of discriminating the duals, because the quantities receive their interpretation externally, from *outside* the theory: there is an *external interpretation* (in the harmonic oscillator case: measuring the masses independently). But theories of the universe lack such an external viewpoint, and so the theory must be interpreted *internally*: in which case—as suggested in §1.3.3—no meaningful distinction can be made between two dual theories.

And yet such accounts are not entirely satisfactory. For "a theory of *everything*" (or, better: a theory of the whole world, cf. De Haro (2016: §2.4.1)) is a slightly vague notion. Leaving aside the fact that one must require the bare theories to have correct rules for forming propositions and must be mathematically consistent (as I discussed in §1.1.1 and §3.2). The requirement that the theory be "a theory of *everything*" is too strong, since we do not need the theory to describe absolutely *all* the facts, not even all the physical

facts, of a world, in order for it to admit an internal interpretation. It is sufficient that the theory describes all the *relevant* physical facts, i.e. it is enough that it be unextendable relative to a domain D of W (which, in particular, contains the notion of *completeness*: cf. §1.3.3). The internal interpretation[42] does not hinge on a strong physicalist thesis, though it is compatible with it! My construal, in §1.1.2, of a physical interpretation I_T as a pair of surjective maps, which (as per the discussion in §1.3.1 and §1.3.3) can be defined for an unextendable theory on a domain D of W, clarifies and makes this point precise.

This lack of clarity regarding when an internal interpretation obtains is an example of the first limitation, (i), listed in §4.1: the lack of a clear-cut criterion, in extant accounts, which would discern theoretical and physical equivalence. I have argued that the key concept needed in order to analyse this difference correctly, so that it could be applied to other cases, starts from the notion of *unextendability*, and is articulated together with the contrasts (a)-(d) listed in the Introduction. A bare theory, which is a triple T, admits an internal interpretation if it is unextendable: i.e. it is complete[43] relative to W (it has a well-defined syntax which is consistent and encompasses all empirical data in $D \subset W$) and it cannot be extended (cf. especially §1.3.3).

As to problem (ii) pointed out in §4.1's list, there are two points to make:

First and most important: on my analysis, there *are* significant cases in which the external interpretation is to be taken seriously—cases for which it may, in fact, be the *only* coherent interpretation: and my notion of extendability explains when this is in fact the case: for one such example (and there are more!), see the discussion of black holes in the next subsection. So my construal of physical equivalence allows for cases—also cases of dualities *in string theory*!—where one and the same duality can be interpreted either internally or externally, depending on the context in which the theory is used. On the other accounts, it is not clear whether, and how, this works. Thus, in my view, the accounts of Huggett (2016), Fraser (2016), and Rickles (2016), for all their merits in recognizing the importance of the *internal* interpretation, do not engage sufficiently seriously with the *external* interpretation of dualities. Castellani (2016), and especially McKenzie (2016), rightly deal with this question with more caution.

The second point relates to the logic of Hugget's text (and I am grateful to him for clarifications[44]).

In (2016), Huggett *first* makes the decision about physical equivalence, i.e. *before* he decides about other important matters such as the sameness of phenomenal and target spaces. But physical equivalence is an *endpoint* of the analysis of bare theories, rather than *the starting point*, in my account. Physical equivalence is the endpoint of the analysis of bare theories, and the *starting point* of their *interpretation*. In order to move from duality to physical equivalence, one first needs to: (a) distinguish bare vs. interpreted theories, (b) decide on extendability vs. unextendability, and (c) decide on the availability of external vs. internal interpretations. In other words: more conceptual work is needed before we are

[42] As mentioned in the Introduction (b) and in footnote 25, the internal interpretation can be upheld under slightly weaker conditions, namely when the theory can be extended to a larger domain but the interpretation is *robust* against such extensions, i.e. none of the potentially relevant extensions change the internal interpretation.

[43] Completeness was defined at the end of §1.1.1.

[44] Huggett, personal communication.

entitled to conclude physical equivalence—on pain of missing relevant cases of physical *in*equivalence!

The difference seems to me to be substantive. Huggett's analysis is thus concerned with cases of physical equivalence only. And our differences do not lie his first assuming physical equivalence and arguing for it later, after the second interpretive fork—which can simply be seen as a "deferred proof of a lemma". The difference is that Huggett unproblematically (though not uncritically!) endorses the consensus: the commentators "have been pretty uniform in taking the stance that the T-duals should indeed be taken as giving the same physical description... I argue below that they are correct." Thus while Huggett does have a discussion of why the two duals are physically equivalent, he does not give any sign of envisaging situations in which one T-dual theory *can, in fact,* be the correct description, and the other theory simply an auxiliary tool.

4.3 Rickles and Fraser

As mentioned in §4.1, Rickles (2016) distinguishes three cases of underdetermination, duality qualifying as a case of 'formal' underdetermination. On Rickles' analysis, "dual theories are simply examples of theoretically equivalent descriptions of the same underlying physical content." (2016: Abstract). One interesting aspect of Rickles' account is his assertion, which goes further than the other authors discussed, that theoretical equivalence of dual theories should be understood as a case of *gauge-type* symmetry for *all* cases of duality. Also, Rickles engages with the literature on equivalence by Ben-Menahem, Earman, Norton, and others.

My account differs from Rickles' in this obvious sense, that it is stated formally as an isomorphism between triples, rather as a gauge-type symmetry.[45] But the most important difference is in points (i)-(ii) mentioned in §4: like Huggett, also Rickles does not seem to contemplate serious cases of duality in which the theories could fail to be physically equivalent, perhaps because of the lack of a detailed analysis of physical equivalence along the lines of the scheme of contrasts (a)-(d) in the Introduction.

If one's account moves too quickly from duality to physical equivalence, it may render the external, and multifaceted, uses of dualities unintelligible. Rickles (2016: §2) gives an interesting example of "*applied* dualities, in which, e.g. the gauge-gravity duality is applied to a real-world manipulable system, such as condensed matter systems. In such cases we simply *know* that we are not dealing with microscopic black holes in higher dimensions." How we "simply *know*" this fact is left unexplained by Rickles (though the word 'manipulable' might be relevant). I have no doubt that Rickles *knows*, and justifiably so, that we are not dealing with black holes in higher dimensions. But I submit that such knowledge conflicts with his own construal of duality (at least, without further elaboration), according to which duality is a gauge-type symmetry: "I hold that this can be generalized to *all* cases in which one has a duality symmetry: they can always be promoted to gauge-type symmetries because they just *are* gauge-type symmetries." (§2, penultimate paragraph).

An account of dualities as being gauge-type symmetries does not explain the black

[45]For more on the relation between duality and gauge symmetry, see De Haro et al. (2016, 2016a).

hole case, at least not without further elaboration: for an explanation is needed for why the 'condensed matter-gauge' is real and the 'black hole-gauge' is not. In its turn, the characterisation of physical equivalence as 'gauge-type' does not work towards clarifying the notion of 'gauge' either: a notion about which there is still much confusion in the physics literature (on this point, cf. §2, and in particular footnote 1, of De Haro et al. (2016): one needs to disentangle redundant from nonredundant gauge symmetries, whether they be local or global; as well as the work done in §1.1-§1.2 of this paper). And explaining the difference between these two 'gauges' *is* possible once we admit the idea of an external interpretation.[46]

Next I turn to Fraser (2016). My main point will revolve around the example of Euclidean field theory (EFT) and QFT being not even a case of theoretical equivalence, let alone physical equivalence. Further elucidating the distinction between theoretical and physical equivalence indeed seems to be the main aim of Fraser's (2016) example of equivalence between EFT and QFT. She maintains that EFT and QFT are *theoretically equivalent*[47] but not physically equivalent. She contrasts this with dualities in string theory, which, she says, are cases of *physical equivalence*. The example would seem to be well chosen indeed: and an elaborate case study might be able to clarify what it is that blocks physical equivalence in one case but not in the other. Fraser's mastery of QFT is indisputable: but, unfortunately, she does not give a detailed account of how these two cases are supposed to differ (for Fraser's own account of this, see two paragraphs below).

And, more importantly, the example *itself* is deceptive if taken as a difference-maker for theoretical and physical equivalence: for it is in fact *not* a case of theoretical equivalence! As Fraser admits in the first three sections of her paper, EFT and QFT are *not* isomorphic: there is a map from EFT to QFT but not the other way around. So "the 'vice versa' in the criteria [of theoretical equivalence] is not satisfied, and the theoretical relations are entailments rather than equivalences... That the relations are entailments rather than equivalences is not one of the points of comparison between the EFT-QFT case and string theory that I want to emphasize" (§3, paragraph 3).

One may, of course, choose to downplay the role of equivalence, and to focus instead on a one-way entailment, if one is just interested in a *generic* contrast between the EFT-QFT case and dualities in string theory, as indeed Fraser is. But entailment relations will *not* give us the difference-maker we need in order to distinguish theoretical from physical equivalence, for the very reason that these are *not cases of theoretical equivalence*.[48]

[46] In this case, the *extendability* of the two models rules out the internal interpretation. The black hole case was discussed in detail in De Haro (2016: §3.2.3, 3.5).

[47] Formally equivalent, as Fraser calls it. For Fraser, 'formal equivalence' entails: (i) the existence a 'translation manual' between the two theories, mapping the physically significant states, and the quantities, between the two theories; (ii) the map preserves the evaluations of all physical quantities which are physically significant. From her discussion in Section 4.3, it seems that Fraser is also assuming equivariance (though not identity!) of the dynamics with respect to duality. In the case of an invertible map, these conditions thus together establish an isomorphism between the two theories, in the way I have discussed it. For the case that the map is not invertible, this is what Fraser calls 'one-way entailment': see the next paragraph.

[48] One might argue that something might still be learned from the contrast between *one-way theoretical entailment* and *one-way physical entailment*. But, unfortunately, here also Fraser proposes no clear difference-makers.

As I have argued, the contrast of external vs. internal interpretations *is* such a difference-maker. Such an account is thus needed, if we wish to explicate what is special about those dualities which amount to physical equivalence, and what it is that the other dualities lack.

The internal interpretation of dualities is also not restricted to a context of discovery, as some accounts might suggest. Thus Fraser says: "This difference between the analytic continuation and T-duality cases occurs in the context of discovery" (Fraser (2016: §4.1)): "in the case of T-duality, each of the dual theories is only partially physically interpreted prior to introduction of the duality transformations. *Moreover, it is the fact that each theory is only partially physically interpreted that makes possible the ultimate judgment that the theories are physically equivalent*" (her emphasis, 2016: §4.1). But the key sufficient condition for claiming physical equivalence, in my view, is to secure an internal interpretation, and this depends on an objective criterion—unextendability—not on a historical or psychological one, such as the contingent fact that a theory happens to be only partially interpretated. Thus I submit that my analysis applies in the context of discovery *as well as* the context of justification. Of course, gravity theories in five dimensions, and conformal field theories in four dimensions, were known—and interpreted!—long before gauge/gravity dualities were discovered. The internal interpretation can thus *correct*, where appropriate, an extant external interpretation.

The articulation of the internal interpretation in terms of the notions of bare and interpreted theories and their unextendability, as in this paper, makes clear that a decision whether an internal interpretation exists is, first and foremost, a (complex) decision about *physics*. But to then be able to move to physical equivalence, we need to secure unextendability:[49] and, as I have argued, unextendability is highly constrained by both physics and by mathematics!

Envoi

Let me end by echoing an important remark: the discussion, in Section 3, of the philosophical significance of gauge/gravity dualities, returns us to the idea (echoed, for instance, in the remark by Huggett and Wüthrich, quoted at the end of Section 3) that philosophical analysis goes hand-in-hand with theory construction (cf. §3.1).

The analysis of gauge/gravity dualities shows specific features of this two-way street. On the one hand, concepts such as theoretical and physical equivalence help us to construct theories of quantum gravity, and so brings metaphysical analysis to bear on theory construction.

But also, on the other hand: theories of quantum gravity, in particular gauge/gravity dualities, help us achieve greater clarity about these two philosophical concepts (cf. §1.3), thereby exhibiting the virtues of a 'science first' approach to metaphysics.

[49] Or, at least, that they are sufficiently robust under extension: see footnote 42.

Acknowledgements

It is a pleasure to thank Elena Castellani, Adam Caulton, Dennis Dieks, Doreen Fraser, Nick Huggett, Yolanda Murillo, Huw Price, Dean Rickles, Bryan Roberts, Nicholas Teh: and, especially, Jeremy Butterfield, for insightful discussions and comments on this paper. I would also like to thank several audiences: the British Society for the Philosophy of Science annual conference, the Oxford philosophy of physics group, LSE's Sigma Club, the Munich Center for Mathematical Philosophy; *Carlofest*, Carlo Rovelli's 60th birthday conference in Marseille; *L'émergence dans les sciences de la matière* in Louvain-la-Neuve, and *Philosophy of Science in a Forest*, Doorn. This work was supported by the Tarner scholarship in Philosophy of Science and History of Ideas of Trinity College, Cambridge.

References

Aganagic, M. (2016). "String Theory and Math: Why This Marriage May Last. Mathematics and dualities of quantum physics". *Bulletin of the American Mathematical Society*, 53(1) pp. 93-115. arXiv:1508.06642 [hep-th].

Ammon, M. and Erdmenger, J. (2015). *Gauge/Gravity Duality. Foundations and Applications*. Cambridge: University Press.

Atiyah, M. (1990). "On the Work of Edward Witten". *Proceedings of the International Congress of Mathematicians*, pp. 31-35.

Atiyah, M. F. (2007) "Duality in Mathematics and Physics", lecture delivered at the Institut de Matemàtica de la Universitat de Barcelona: `http://www.imub.ub.es`.

Barrett, T. W. and Halvorson, H. (2015). "Glymour and Quine on theoretical equivalence". PhilSci 11341.

Bouatta, N. and Butterfield J. (2015). "On Emergence in Gauge Theories at the 't Hooft Limit", *European Journal for Philosophy of Science* **5**, 55-87.

Castellani, E. (2016). "Duality and 'particle' democracy", forthcoming in *Studies in History and Philosophy of Modern Physics*. doi: 10.1016/j.shpsb.2016.03.002.

Coffey, K. (2014). "Theoretical Equivalence as Interpretative Equivalence". *British Journal for the Philosophy of Science* 65, pp. 821-844.

De Haro, S., Skenderis, K., and Solodukhin, S. (2001). "Holographic reconstruction of spacetime and renormalization in the AdS/CFT correspondence", *Communications in Mathematical Physics*, 217(3), 595-622. doi: 10.1007/s002200100381 [hep-th/0002230].

De Haro, S. (2009). "Dual Gravitons in AdS(4) / CFT(3) and the Holographic Cotton

Tensor," *Journal of High Energy Physics* **0901** 042. doi: 10.1088/1126-6708/2009/01/042 [arXiv:0808.2054 [hep-th]].

De Haro, S. (2016). "Dualities and emergent gravity: Gauge/gravity duality", forthcoming in *Studies in History and Philosophy of Modern Physics*. doi: 10.1016/j.shpsb.2015.08.004.

De Haro, S. (2016a). "Invisibility of Diffeomorphisms". *Foundations of Physics*, submitted.

De Haro, S., Teh, N., Butterfield, J.N. (2016). "Comparing Dualities and Gauge Symmetries", forthcoming in *Studies in History and Philosophy of Modern Physics*. PhilSci 12009. doi: 10.1016/j.shpsb.2016.03.001.

De Haro, S., Teh, N., Butterfield, J.N., (2016a). "On the relation between dualities and gauge symmetries", *Philosophy of Science*, proceedings of the Philosophy of Science Association (PSA), in press.

De Haro, S., Mayerson, D., Butterfield, J.N. (2016b). "Conceptual Aspects of Gauge/Gravity Duality", forthcoming in *Foundations of Physics*.

Dieks, D., Dongen, J. van, Haro, S. de (2015), "Emergence in Holographic Scenarios for Gravity", *Studies in History and Philosophy of Modern Physics* 52(B), 203-216. doi: 10.1016/j.shpsb.2015.07.007.

Fraser, D. (2016). "Formal and physical equivalence in two cases in contemporary quantum physics", forthcoming in *Studies in History and Philosophy of Modern Physics*. doi: 10.1016/j.shpsb.2015.07.005.

Glymour, C. (1977). "The epistemology of geometry". *Noûs* 11(3), 227-251.

Huggett, N. and Wüthrich, C. (2013). "Emergent spacetime and empirical (in)coherence", *Studies in History and Philosophy of Modern Physics* 44(3), 276-285.

Huggett, N. (2016), "Target space $\neq$ space", PhilSci 11638; forthcoming in *Studies in the History and Philosophy of Modern Physics*. doi:10.1016/j.shpsb.2015.08.007.

Ismael, J. and Van Fraassen, B. C. (2003). "Symmetry as a guide to superfluous theoretical structure". In: *Symmetries in Physics: Philosophical Reflections*, K. Brading and E. Castellani (Eds.), 371-392.

Koch, J. (2003). "Frobenius Algebras and 2D Topological Quantum Field Theories". Cambridge: Cambridge University Press.

Matsubara, K. (2013). "Realism, underdetermination and string theory dualitites". *Synthese*, 190: 471-489.

McKenzie, K. (2016). "Relativities of fundamentality", forthcoming in *Studies in History and Philosophy of Modern Physics*, doi: 10.1016/j.shpsb.2015.08.001.

Papadimitriou, I. and K. Skenderis (2005). "AdS / CFT correspondence and geometry," *IRMA Lectures on Mathematical and Theoretical Physics,* 8, 73. doi:10.4171/013-1/4 [hep-th/0404176].

Polchinski, J. (2016). "Dualities of Fields and Strings". *Studies in History and Philosophy of Modern Physics*, to appear. arXiv:1412.5704 [hep-th].

Rickles, D. (2012). "AdS/CFT duality and the emergence of spacetime", *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 44(3), 312-320.

Rickles, D. (2016). "Dual theories: 'same but different' or different but same'?", forthcoming in *Studies in the History and Philosophy of Modern Physics*. doi: 10.1016/j.shpsb.2015.09.005.

Vafa, C. and E. Witten (1994). "A Strong coupling test of S-duality", *Nuclear Physics B* 431, 3. doi:10.1016/0550-3213(94)90097-3 [hep-th/9408074].

't Hooft, G. (1993). "Dimensional reduction in quantum gravity", in: Ali, A., J. Ellis and S. Randjbar-Daemi (Eds.), *Salamfestschrift*. Singapore: World Scientific.

Teh, N. and Tsementzis, D. (2016), "Theoretical Equivalence in Classical Mechanics and its relationship to Duality", forthcoming in *Studies in History and Philosophy of Modern Physics*. doi:10.1016/j.shpsb.2016.02.002

Weatherall, J.O. (2015). "Categories and the Foundations of Classical Field Theories". Forthcoming in *Categories for the Working Philosopher*, E. Landry (Ed). Oxford: Oxford University Press.

Zwiebach, B. (2009). *A first course in string theory*, Cambridge, UK: Cambridge University Press.