Functionalism Fit for Physics

Eleanor Knox and David Wallace

October 5, 2023

Abstract

We put the recent flurry of interest in functionalism in philosophy of physics into context by considering functionalism's roots in philosophy of mind. There we identify two types of functionalism, which we call 'causal-role' and 'constitutive' functionalism: the former is a defeasible reductive hypothesis, while the latter, when true, is analytically so, and is not itself reductive. We argue through case studies that it is the constitutive notion of functionalism that is the better fit to physics.

1 Introduction: the roots of functionalism

> Functionalism is the idea that handsome is as handsome does, that matter only matters because of what matter can do. Functionalism in this broadest sense is so ubiquitous in science that it is tantamount to a reigning presumption of all of science.
>
> Daniel Dennett[1]

If Dennett is correct, functionalism ought to be ubiquitous in physics too, and so it comes as no surprise that philosophers of physics have found uses for functionalism. *Spacetime functionalists* mostly agree on a few basic facts - that spacetime is whatever fills some particular functional role, and that spacetime functionalism might be helpful in understanding how to think about the relation between apparently non-spatiotemporal theories of quantum gravity.[2] They disagree as to what the functional role of spacetime is, and (importantly for our purposes here), where to look for the realizers of this role. Lam & Wüthrich (2018) see spacetime functionalism as an instance of Kim's functional reduction, and look for realizers of the role in the putatively non-spatiotemporal theories of quantum gravity.[3] Knox (2019) applies her function-

[1](Dennett, 2005, p.17)

[2]Chalmers' spatiotemporal functionalism, which is aimed at recovering our experience rather than at interpreting physical theory, is an exception here.

[3]Putatively, because, as several have mentioned[refs], this kind of reductive functionalism tries to identify spacetime in a theory of quantum gravity and hence establish that it was spatiotemporal after all.

alism to general relativity and other theories without an explicitly reductionist agenda.

Functionalism about physics appears earlier in the context of quantum mechanics. Wallace (2003, 2012) appeals to functionalism in response to a challenge: How can distinct macroscopic worlds emerge from the wavefunction in Everettian quantum mechanics when the wavefunction describes superpositions of particle states, rather than multiplicity of particles? Albert (2015, Ch.6) uses functionalism to answer a similar kind of question: How can three-dimensional objects be recovered in 'wavefunction realism', a position that holds that the fundamental space is 3N dimensional configuration space? In each case one proposes a fundamental theory whose ontology looks radically different from some piece of structure we need to recover at a higher level. Functionalism is used to explain how we can possibly recover some more familiar structure. Returning to spacetime functionalism in the context of quantum gravity, one might propose a similar central question: How can *spacetime* be the kind of thing that one recovers from a non-spatiotemporal picture?

But while (following Dennett) the slogans here are easy — for example *spacetime is as spacetime does* (Lam & Wüthrich, 2018) — the details are harder, and more contested. Insofar as philosophers of physics have developed a philosophical account of functionalism, they have tended to assume that functionalism is functional *reduction*, and have cashed that idea out in terms drawn from David Lewis (1970). Probably the clearest such statment is due to Butterfield & Gomes (2020): they demand that allusions to functionalism be accompanied by a clearer philosophical account of the kind of functionalism proposed, and offer their own, admirably clear, Lewisian account.

We agree with the need for an account. But a Lewisian approach, for all its clarity, does not seem to us to be the most useful account for physics, precisely because it equates functionalism with functional reduction. We'll argue here for an alternative: functionalism in physics is not itself a reductive program, though it is often accompanied by one.

To understand this claim, and more generally to untangle the web of positions on functionalism in physics, it's helpful to study functionalism in its original habitat: philosophy of mind. In section 2, we do just this, and distinguish two quite different notions of functionalism: 'causal role' and 'constitutive functionalism'. In section 3 we turn to physics: we describe the case of fluid dynamics, and look at how the distinction between causal-role and constitutive functionalism applies there. In section 4 we look in more detail at the reductive picture proposed by the two views and relate this to an account of the relation between semantic and syntactic views of theories proposed by Wallace. Casual-role functionalism appears to give a route to Nagelian reduction of theories conceived syntactically. Constitutive functionalism maps onto a semantic conception of theories, and helps to explain how we move from mathematical models to linguistic descriptions that admit of a syntactic form. In section 5 we examine a standard case of reduction in physics - that of the recovery of Newtonian gravitation in the weak field limit of general relativity, and argue that causal-role functionalism is a poor fit for this. In section 6, we

look at the prospects for rehabilitating causal-role functionalism in a semantic context, and argue that they are lacking. Our conclusion (section 7) is that it is constitutive and not causal-role functionalism that finds a natural home in physics.

The physics we appeal to is in all cases well established and we do not attempt to quote original sources. See, e.g., (Thorne & Blandford, 2017) for a technical reference for the classical-mechanical ideas we discuss, and (Balescu, 1997; Zwanzig, 2001) for the statistical-mechanical reductions.

2 Two kinds of functionalism

In its original philosophy-of-mind context, 'Functionalism' is ambiguous between two quite distinct ideas, both traceable back to Ryle and Wittgenstein's logical behaviorism (Ryle, 1949; Wittgenstein, 1953).[4] To logical behaviorists, mental concepts like pain, or the belief that snow is white, are logically reducible to behavioral dispositions: to be in pain is *inter alia* to moan, grimace, struggle to concentrate; to believe that snow is white is inter alia to assert that snow is white in appropriate contexts. Logical behaviorism foundered on (i) the increasing realization that no set of behavioral dispositions plausibly characterizes one mental state in isolation, but instead mental states give rise to behavior in some more holistic way; (ii) a rejection of Rylean hostility to empirical science, so that the behavioral characterization of the mental is not determinable purely by conceptual and linguistic analysis but is also informed by scientific data.

The (always schematic) result of this is that mental states become terms in some systematic psychological theory (call it T), which makes predictions about an agent's bodily movements, speech acts and the like contingent on a full characterization of that agent's appropriate mental states (usually including at least beliefs, desires, and memories). T is functionalist in the sense that it posits a series of functional relations between one mental state and another, and between the functional network of mental states and the agent's actions.

But there are now two quite different attitudes to T available. To causal-role functionalists like Fodor (1975) and Lewis (1972; 1980), mental states are to be thought of as unobservables akin to the unobservables of physics. The empirical success of T is to be explained by the fact that the unobservable entities posited by T actually exist, and that their influence on action is to be understood causally: my belief that there is wine in the glass is some posited neurological excitation; it is causally brought about by the impact of the light from the glass on my retina; in turn that belief, combined with my desire to drink wine (another posited neurological excitation), causally brings about my sipping from the glass. It is logically possible that there are no such identi-

[4]The distinction we draw here is not one often found in contemporary overviews of the subject. But earlier work makes the distinction: Dennett discusses it extensively in Chapter 10 of *The Intentional Stance* (1987), and Block (1978) alludes to it, if only to dismiss one side of the debate as not really functionalist.

fiable neurological excitations, in which case I actually have no such beliefs or desires; but the (presumed) empirical success of T licenses a compelling inference-to-the-best-explanation argument for the existence of those excitations, even as we remain ignorant of their nature.

The alternative attitude might be called constitutive functionalism. To a constitutive functionalist, what it is for an agent to have mental property X is nothing more or less than for mental property X to be part of that ascription of mental properties to an agent that, collectively, best matches their behavioral-dispositions-according-to-T to their actual behavioral dispositions. On this approach, from the fact that I drink the wine (and from other facts about my behavior, like the fact that I ordered it at the bar) it will follow analytically that I believed there to be wine in the glass and that I wanted to drink it. Constitutive functionalism, unlike causal-role functionalism, makes no hypothesis about the sub-personal level and cannot be falsified by any discoveries about that level. Probably the best known concrete proposal for constitutive functionalism is Dennett's (1987) *intentional stance*: one takes the intentional stance towards a system if one attributes to it a set of beliefs and desires such that its actions can be predicted on the assumption that it acts rationally to fulfil its desires on the basis of its beliefs (supplemented by some assumptions that those beliefs are updated appropriately when the system receives appropriate information); a system actually has given beliefs and desires if the intentional stance is predictive of the system on that assignment of beliefs and desires.

Causal-role functionalism is inherently reductive: it attempts to ground behavior at the psychological level in facts at the sub-personal, neurological level. Constitutive functionalism is not reductive, but has an associated reductive project: we want to understand how, in sub-personal terms, a physical system comes to be accurately described at the behavioral level by T (by the intentional stance, say). But the success conditions on this reductive project are simply that T somehow comes to describe the system at that level. It might do so via the causal-role functionalist approach: maybe the intentional stance applies to a system because it really has some set of neurological subsystems that can be understood as individually encoding beliefs and desires, and interacting computationally in some way that mimics the intentional-level description. But the mere success of the intentional stance licenses no inference to that conclusion, (and, Dennett argues persuasively,[5] there are good theoretical reasons for skepticism).

Butterfield and Gomes advocate causal-role functionalism about physics. Wüthrich and Lam also seem to have this kind of functionalism in mind.[6] We'll argue here that constitutive functionalism is more helpful for the understanding of physics and its inter-theoretic relations. For one thing, it better serves a project that we call "explication" in physics - that is, the project of taking a mathematically-formulated theory and offering an English language descrip-

[5]Dennett (1987), pp.65-68 and throughout.

[6]In what follows, we'll continue to call this kind of functionalism causal-role functionalism, although we'll also note that the roles proposed in the philosophy of physics literature often appeal more to dynamics than to traditional causation.

tion of it. For another, it fits better with actual examples of reduction in physics, which rarely involve the process described by the causal role functionalist.

3 Causal-role and constitutive functionalism applied to physics

Consider now a physics-based example: the macroscopic concept of liquid. Our pre-theoretic (or perhaps folk-physics) concept of 'liquid' presumably includes things like 'flows through holes', 'is not compressible', 'has no fixed shape' and the like; fluid dynamics refines and corrects this to something like 'liquids are characterized by the fact that they obey the incompressible Navier-Stokes (N-S) equation'. That is, the equation below, written here for a fluid with velocity $\mathbf{u}$, pressure p, density ρ, viscosity ν, and external (e.g., gravitational) bulk force $\mathbf{g}$:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} - \nu \nabla^2 \mathbf{u} = -\frac{1}{\rho}\nabla p + \mathbf{g} \tag{1}$$

There are plenty of physical systems — water, molten brass, molasses, milk — for which the N-S equation is indeed powerfully predictive. This might be considered a functional definition of 'liquid', and — perhaps more closely analogous to mental states — of the various parameters in the N-S equation, like density and viscosity, which for the most part are not observable directly but only via their effects on the bulk behavior of the liquid.

Both the causal-role and constitutive approaches to functionalism are available here. A causal-role functionalist will say that viscosity is a theory term, hopefully to be identified with some microscopically-characterized feature of the system of atoms and molecules comprising the liquid; arguably, 'liquid' itself will likewise be so characterised. In principle no such identification might be possible, in which case the liquid does not really have a viscosity; in practice (says the causal-role functionalist) the very success of the N-S equation licenses a compelling inference to the best explanation, leaving us highly confident that there will indeed be such an identification.

Let us spell out causal-role functionalism in a bit more detail. The causal-role functionalist seeks their reduction in the following way:[7]

1. Choose a theoretical term to functionalize in a higher level theory.

2. Pick out its functional role in terms of its relation to other terms. This involves dividing the language into terms to be given a functional treatment and those taken to be understood, but can involve simultaneous unique definition. (This project is hard, involved, and not guaranteed to succeed!)

3. Find the realizer of the functional role in a reducing theory.

[7] Here we follow Butterfield & Gomes (2020, p.4)

4. The resulting statement of identity provides a bridge law in the Nagelian sense.

The second step here sometimes goes by the name 'implicit definition'. As an approach to theoretical terms, it's related to a class of proposals made by Ramsey, Carnap and Lewis, and is sometimes called the Ramsey-Carnap-Lewis method.[8] The obvious way to implement it is to produce the 'Ramsey sentence' of the theory with respect to the particular functionalized term(s).

How might we implement the above procedure in the liquid example? If we are content that the N-S equation yields a functional definition of, say, viscosity and liquid, then the first two steps seem moderately straightforward, at least if we can write the theory in an appropriate form. But things get much trickier when we reach step 3. There is no obvious candidate entity in molecular dynamics to identify with liquid (a mereological sum of molecules is a poor candidate: liquids are continuous and smooth, mereological sums of atoms are not), let alone with viscosity (properties like this tend to be derived from the microphysics through complex and indirect means; 'temperature is mean kinetic energy' is a very special case, and at any rate applies only to dilute gases). The derivation of fluid dynamics from molecular dynamics is quite well understood but also highly mathematised: as a rough sketch, one (a) starts with the overall, highly spiky density function of the molecules; (b) Fourier transforms it into a sum of smooth periodic functions; (c) looks for conditions in which the comparatively-small number of long-wavelength functions in that decomposition have dynamics approximately autonomous from the details of the shorter-wavelength functions; (d) seeks further approximations under which that autonomous dynamics has the N-S form; and finally (e) reads off the viscosity and other coefficients. This process does not — at least, does not in any remotely simple sense — give rise to the sort of identification of microphysical features with functionally-characterized features that the causal-role functionalist seeks. Indeed, it profligately mixes ontological categories (an object, the liquid, is derived from certain collective properties — low wavelength modes — of the underlying molecular distribution) in a way that seems systematically hostile to implicit definition.

By contrast, for the constitutive functionalist, all this is unproblematic. Recall that, on their account, behaviour in the higher-level account is sufficient for the truth of claims about the functionalised entity or property. As such, they are committed to something like the first two steps of the casual-role functionalist's reduction, but not the third or fourth. We say 'something like' these steps, because explicit Ramsification is not the best strategy for the constitutive functionalist. Rather than taking a theory already expressed in predicate terms, the constitutive functionalist is interested in applying predicate terms - belief, desire, viscosity, or density, to a system.

Their account proceeds in studied ignorance of the microphysics: never mind why, in microscopic terms, the system obeys the N-S equations, it suffices

[8]For discussion of the method, and historical references, see (Raatikainen, 2021).

to observe that it does. Nothing more is required for it to be a liquid; nothing more is required for it to have viscosity v than for that to be the value that the viscosity must be set to for the N-S equation to reproduce the actual behavior of the system. It is constitutive of a system being a liquid, and its having viscosity v, that these facts obtain. In themselves they comprise no kind of reduction, but of course a serious reductive question remains: what, in microphysical terms, explains the fact that the system obeys the N-S equation? But that question has a quite satisfactory answer in statistical mechanics: the one we sketched above, in fact. Once we combine (a) the constitutive-functional analysis that tells us what it is to be a liquid of a certain viscosity is to satisfy certain equations, and (b) the derivation that certain collective degrees of freedom of microphysically-characterized systems indeed do satisfy those equations, no residual reductive work remains.

If a more piecemeal reduction is available in specific cases, all well and good. For instance, in certain rather special systems temperature can approximately be identified as mean kinetic energy and not as some more holistically-characterized feature of the system; somewhat more generally, if both higher-level and lower-level theory have spatiotemporally local dynamics, there is a (defeasible) expectation that the reduction relation can be spatially localized to some degree, with higher-level properties of some region being determined by lower-level properties of approximately the same region. But there is no reason for a general expectation — let alone a formal requirement — that any such piecemeal reduction can be obtained (expectations of localizability, for instance, plausibly fail in quantum gravity).

To further illustrate the distinction between the two views, consider the possibility of 'content zombies': beings lacking content-bearing states but behaving as if they had them. ('Qualia zombies' — systems lacking qualitative experience or self-consciousness but behaving as if they had them, as discussed in (e.g.) (Chalmers, 1996) — are not relevant to the physics analogy we wish to make here.) Content zombies are systems behaviorally identical to humans but lacking genuine mental states. For the causal-role functionalist, zombies are at least logically possible, albeit perhaps not scientifically respectable: all it would take would be a physical system accurately described by theory T, but without anything playing the causal role of the belief and desire states in T. (An absurdly large lookup table, perhaps, or — the classic example of a content zombie in the philosophy of mind literature — Searle's 'Chinese room' (1980).) For the constitutive functionalist, zombies are conceptually (not just physically) impossible: a system accurately described by a certain belief/desire assignment just is a system with those beliefs and desires.

What, then, is the causal-role functionalist to make of liquids, given that no implicit definition of viscosity is available? They appear forced to the view that the apparent liquids we observe are really zombie liquids: they behave as if they were liquids with certain viscosity, but they are not really liquids; they do not really have viscosities. And indeed one sometimes sees exactly this claimed: since liquids are continuous on all scales, and since nothing in nature is continuous on all scales, there are not really liquids, but only systems

that behave like liquids on appropriate scales. But, of course, similar reasoning will apply to the atomic scale below the liquid; if one considers quantum field theory, implicit definition of particles will look just as unlikely. And quantum field theory is hardly the last word. One might reasonably expect the reduction of quantum field theory to whatever more fundamental theory underlies it to proceed in a manner just as hostile to implicit definition. The proponent of causal-role functionalism is forced to admit that the vast majority of the systems we encounter are zombie-systems — an uncomfortable conclusion for a position that is usually motivated by realist considerations.

By contrast, while the constitutive functionalist is often viewed (correctly) as a pragmatist, they are at least not committed to a vast error-theory. Insofar as there are systems well-described by our physics, those systems contain the kinds of entities and properties a thoughtful description or explication of the physics suggests. This thoughtful description itself involves a flexible functionalism, one which is tolerant of borderline or tricky cases. As every parent or child who has ever played with the starch and water mixture known as oobleck[9] knows, systems can behave as liquids in some circumstances and not in others: 'non-Newtonian fluids' have variable viscosity under stress. They will obey the N-S equations under some stress regimes, and in others, will cease to behave like a liquid at all. And indeed, that is how we talk about them — oobleck is a liquid when it's dripping from a spoon, and a solid when hit by the same spoon — that is the whole point of the science experiment.

4 Functionalism and inter-theoretic relations

To make this talk of theory description, and its relation to reduction, precise, it will be helpful to lay out a schema for understanding the mathematised models of physics. Our proposal is based on Wallace's 'mathematics-first structural realism' (2021), but one needn't subscribe to every aspect of his programme to accept the basic layout.[10] What is essential to us here is:

1. Theories in physics consist of classes of mathematical models. While these are interpreted theories in the sense that they have (at least) target systems and representational capacities, they are not *linguistically* interpreted: they do not provide a full linguistic description of the system in terms of, say, particles, forces, liquids, viscosities and so on.

   Further linguistic description of a theory, while common in the scientific literature, tends to be partial, heuristic, and not always consistent. Trying to regiment these partial heuristics into a full linguistic description

[9] See, e.g., (Zabawski, 2009).

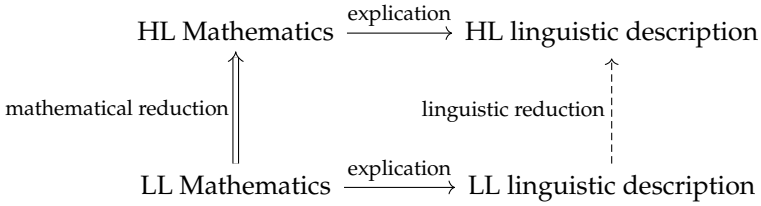
[10] Wallace's full version of structural realism depends on seeing mathematical models as prior to predicate descriptions, and seeing the predicate description as underdetermined by the success of the mathematical models. Although what we say here is entirely compatible with this (and might be seen as an argument in favour of the view), one needn't adopt his particular realist attitude for this schema to be helpful.

of a theory, specifying its ontology and ideology in the way familiar to philosophers, is a substantive additional step. The result of doing this (which we shall here call explication in order to avoid confusion with interpretation in the model-theoretic sense) is a (more-or-less) fully fleshed out language-based description of the system which can be formalized in predicate logic. (Wallace calls such a description a *predicate precisification*.)

2. This project of explication is not itself reductive.

3. In addition, there are two possible styles of reduction: one which relates the mathematical models of the theory, and another which relates the full language-based description.[11]

One can illustrate this schema with a simple diagram:



Here, the left-hand side describes the theory (conceived as a class of mathematical models) and the right-hand side describes a full linguistic/predicate description. One can then ask questions about the four arrows of the diagram. The left-hand vertical arrow represents the reductive relation between models. The right-hand vertical arrow represents the reductive relation between predicate descriptions/ontologies. The horizontal arrows represent the interpretative process by which we extract a predicate description from the mathematical models.

We can now see that constitutive and causal-role functionalism articulate and emphasize different relations. Causal-role functionalists like Lewis and Fodor start with a predicate description of the higher-level theory, and use functional reduction to establish a Nagelian reduction to the lower-level theory. A constitutive functionalist like Dennett is concerned with the horizontal interpretation step, and not — or at least, not *qua* functionalist — with reduction.

Consider some system well-modelled by the Navier-Stokes equations. This could be a familiar application, like the flow of water or oil through an pipeline, or an unfamilar one, like a phenomenological model of the fluid dynamics of a distant planet or star constructed largely in ignorance of the microphysics. The

[11]Guo (2020) draws a closely-related distinction between what she calls 'theory-first' and 'ontology-first' reduction.

constitutive functionalist sees the mathematical model as describing a liquid of viscosity v — insofar as the model is correct (including insofar as approximations apply), that description is secure. The absence of a theory of the microstructure of oil, or of any theory as to the liquid contained beneath the planet's crust, need not threaten our linguistic claims. At the same time, if we *do* have a theory of the liquid's microstructure (as in the case of the oil), we will seek a reduction along the left-hand side of the diagram above. And that reduction will look like the sketch in section 2 — it will involve Fourier-transforming the density function, separating out long-wavelength modes, and eventually showing that, under a series of approximations, the discrete microstructure can model the N-S equations. The micro-theory itself will also be explicated along constitutive functionalist lines - for example, notions of particle and force in the micro-theory might be given a functionalist gloss. While it's not impossible that there might also be some reduction on the right-hand side of the diagram (a relation that maps the predicates of one description to the predicates of the other) the absence of such a reduction is neither surprising, nor problematic. And the absence of such a reduction does not threaten the description of the higher-level theory as involving a liquid with viscosity v.

The causal-role functionalist, in contrast, applies their functional reduction along the right-hand side of the diagram. Their starting point is the theory as described in language, and they seek to locate bridge laws for a Nagelian reduction: Butterfield and Gomes, for example, see this as the central aim of functionalism:

> [functionalism provides] bridge laws that are mandatory, not optional: they are statements of identity (or co-extension) that are conclusions of a deductive argument, rather than contingent guesses or verbal stipulations; and once we infer them, we have a reduction in a Nagelian sense. (Butterfield & Gomes, 2020, p.1)

But if the arguments of section 2 are right, the prospects for a right-hand side reduction are dim. How should the causal-role functionalist respond? If they don't wish to fall into the kind of far-reaching error theory suggested in the last section, their options seem to be (i) to argue that a bridge law exists for viscosity, but we have yet to find it (ii) to argue that the liquid case is not the standard case, and many other cases allow for functional reduction or (iii) to claim that the schema above is misleading, and that what we've called 'reduction as model-instantiation' can actually be thought of as functional reduction.

Section 5 will argue that one cannot simultaneously gloss the right-hand side's reduction as functional reduction and stay true to the aims of the causal-role functionalism *à la* Lewis - semantic functional reduction is not a viable or interesting route for the causal-role functionalist. Knock-down arguments against (i) and (ii) are trickier for obvious logical reasons. But the next section will give reason to think that, even in the friendliest of cases, functional reduction neither matches our actual reductive practice, nor offers helpful insights when applied.

Before moving on, it's worth saying more about explication and constitutive functionalism. After all, it might seem surprising to think of the identification of v with viscosity as an additional step - it's just what the variable represents! Once embedded in the N-S equations, a term like viscosity 'gets its meaning' from its role in the equations. How is there space for giving different descriptions of this mathematical model? There are two things to note here: First, there is usually more than one way of formulating some piece of theory. Explication is, in part, a matter of deciding which formulation's variables we would like to emphasise.[12] Second, it's no coincidence that we connect the variable v with a word that predates the N-S equations by centuries. It's a crucial part of explication that we connect antecedent notions, like the thickness and stickiness of a liquid, and earlier theories of viscosity (like Newton's), with the formal equations - it's this that actually makes an interpretation comprehensible in a way that pure mathematics may not be. Theoretical terms like 'liquid', 'particle', 'force', 'spacetime' or 'viscosity' used in contemporary theories retain their ties to our previous notions, even while they are given a more precise functional definition. A concept like viscosity is refined and precisified in the Navier-Stokes equations - which, for example, allow us to understand viscosities in Non-Newtonian fluids - but offers no extra understanding unless it remains associated with the ideas that came before it.

This balancing act between antecedentally understood functional roles and theory-specific ones is not unique to the constitutive functionalist. The 'Canberra plan metaphysics' associated with causal-role functionalism usually starts with the ultimate antecedent notions: folk-theoretic platitudes.[13] One then looks to theory to tell us what realizes our folk-theoretic role. For example, Menzies (1996) analyzed the role of causation and suggested that something like energy-momentum transfer might realize the role. Thus the Canberra plan effectively treats folk-theory as another theoretical level to be submitted to functional reduction. By contrast, the constitutive functionalist sees the relationship between folk-theoretic term and successive theoretical terms as one of gradual conceptual change, rather than metaphysical reduction.

5 Reduction: a gravitational example

Does the causal-role functionalist's reduction have a place in physics — perhaps alongside reduction as model instantiation? Perhaps the Navier-Stokes example is an unusually demanding one, requiring, as it does, a connection between continuum and discrete dynamics.[14] We'll turn here to a much more

[12]For Wallace, this multiplicity of formulation means that the predicate precisifications that result from interpretation are pragmatically determined and there is no independent matter of fact about which of these offers a true description. He argues that the picture of reduction portrayed here pushes us in that direction. But nothing we've said so far precludes taking a more realist stance committed to the truth of a predicate precisification.

[13]See (Braddon-Mitchell & Nola, 2009) for a collection that applies the Canberra plan to a range of applications.

[14]Not that this makes it especially unusual in physics!

tractable example of reduction, one which manages to avoid the pitfalls of quantum mechanics and statistical mechanics: the reduction of Newtonian gravitation (NG) to general relativity (GR).

Let us recall the key features of causal-role functional reduction. The causal-role functionalist would have us start with the Newtonian theory, and functionalise key terms - perhaps the gravitational field or potential. We then look in the reducing theory (GR) for what instantiates that role, and the resulting identity is a Nagelian bridge law. It is also important to the causal-role functionalist that we place restrictions on the allowable connections - in order for bridge laws to be "statements of identity (or co-extension)" one must identify *properties* of the system at one level with *properties* at another. Consider the (somewhat fictionalised; the identity applies only to dilute gases) classic philosophers' example of temperature and mean kinetic energy. The identity here is intended to be ontological, not merely mathematical.

How well does the reduction of NG to GR fit into this framework? Textbook accounts of the relation between general relativity and Newtonian gravitation [15] proceed in something like the following way:

Start with the general class of GR models. Then take the 'weak-field limit': that is, restrict ourselves to the class of metrics that in some reference frame take the form:

$$g_{ab} \approx \eta_{ab} + h_{ab} \tag{2}$$

where η_{ab} is the flat Minkowski metric and h_{ab} is 'small' - that is, all components are $<< 1$ in some inertial reference frame of the metric η_{ab}. Substituting this into the Einstein field equations, retaining only terms linear in h_{ab}, and gauge fixing h_{ab} gives us the weak field limit of the GR field equations:

$$\partial^c \partial_c h_{ab} = -16\pi T_{ab} \tag{3}$$

To get the Newtonian limit, we need to move to a reference frame in which all source velocities are small, Along with the assumption that stresses are small, this gives a stress-energy tensor that can be approximated as:

$$T_{ab} \approx \rho t_a t_b \tag{4}$$

In these coordinates, equation 3 becomes:

$$\nabla^2 h_{00} = -16\pi\rho \tag{5}$$

One can then see that the Poisson equation

$$\nabla^2 \phi = 4\pi\rho \tag{6}$$

will be satisfied by the system if the following holds:

$$\phi = -\frac{1}{4} h_{00} \tag{7}$$

[15] E.g. (Wald, 1984, Ch. 4), on which this presentation is based.

Recapping the above: We start with the full class of GR models. We then make some (physically motivated) approximations and assumptions, and reduce that class of models to a very special subset - those for which there exist coordinates in which they take a very particular form. We then proceed to show that, given some further approximations the general relativistic field equations can be made to take the form of the Poisson equation if we assume a particular mathematical relationship between the Newtonian potential and a component of the metric.

How well does this fit with the causal-role functionalist's notion of reduction? Presentations of causal-role functionalism, such as (Kim, 2007, p.102) or (Butterfield & Gomes, 2020) present it as giving a methodology for functionalism. The causal-role functionalist starts by functionalising key terms in the higher-level theory before finding their realisers in the lower level theory. In the case above, we start not with Newtonian gravitation, but with GR, the lower-level theory. We started with a large class of models in the lower-level theory, and then restricted and massaged these until we demonstrated that, in appropriate physical conditions, models of GR could instantiate models of Newtonian gravitation. This process, we hold, is typical of reduction in physics. Indeed, it explains why physicists usually talk of lower-level theories 'reducing' to higher level theories, rather than the other way round: reduction, for the physicist, is a matter of reducing the class of models until we reveal that a higher level theory holds in some special circumstances.

What of term-wise implicit definition? We seek to compare whole equations, not single out a particular term - the crucial move in the above is establishing that the GR field equations approximately instantiate the Poisson equation under specific circumstances. Nonetheless, the causal-role functionalist might insist that equation (7) constitutes a bridge law, and that this was derived, in some sense, by looking at the functional role of the gravitational potential. Is this correct? Recall that for Butterfield and Gomes, bridge laws are "mandatory, not optional: they are statements of identity (or co-extension) that are conclusions of a deductive argument". Does equation (7) express an identity or co-extension? What it says is no more or less than that in some very particular reference frame, under some very specific circumstances, a particular component of the metric can be expressed in a particular form. It is not clear where co-extension or identity fits in here.

It is also unclear what level of approximation is allowable within causal-role functionalism. The original philosophy of mind context lacks the mathematical structure to ask questions about coarse grainings and approximations, but it is natural to see the realisation relation as one of strict ontological identity, which makes approximation hard to incorporate. The example above involves a number of substantial approximations, not least those of equations 2 and 4.

More importantly, the derivation above cannot be taken as capturing the essence of the inter-theoretic reduction, simply because important cases are not caught by it. Consider, for instance, a system of neutron stars and/or black holes where (the astrophysically typical case) the distance between the black holes is much larger than either's Schwarzchild radius. It's obvious here that

the weak field limit does not apply on all scales, and yet systems like these are well-modelled by Newtonian gravitation, on the coarse-grained scale at which we can treat the black holes as Newtonian point masses. In this case, equation (7) relates the Newtonian potential to a component of a sufficiently coarse-grained metric. The success criterion here is simply derivation of the Newtonian equations under some approximation salient to the problem at hand; there is no particular reason to expect a single overarching framework that encompasses every situation where this occurs.[16]

This level of coarse-graining and approximation should give the causal-role functionalist pause, and yet, this example is far from cherry-picked. Indeed, this is a particularly clear and clean case - most examples involve probabilities, quantum mechanics and statistical techniques, and we've managed to evade those here. But even here we what we've called 'reduction by model instantiation' fits the standard physics much better than functional reduction.

Where does constitutive functionalism fit in here? Its aims are not reductive, so while it's compatible with the picture above, it doesn't play a role in the reduction described. If it were to play a role, it would be in asking whether terms like 'gravitational field' can be applied in the gravitational theory above. There is a literature going back to Einstein that looks to identify the gravitational field within general relativity. At various points, it has been suggested that the Christoffel symbols, the metric field itself, or the curvature represent a gravitational field in general relativity.[17] None of these proposals has been particularly helpful in establishing the reduction of Newtonian gravitation, although some of Einstein's early proposals, connected to the equivalence principle, played some role in his (bumpy) road to general relativity. In our view, one should view these proposals and the debates surrounding them as attempts to apply constitutive functionalism to general relativity itself. These attempts have been unsuccessful precisely because there is nothing in general relativity that behaves sufficiently like a gravitational field to deserve the title — by constitutive functionalist lights, gravitational fields aren't part of the interpretation of general relativity.[18]

6 Variations on Lewis

Our initial presentation of causal-role functionalism hewed close both to the original philosophy of mind context and to Lewis's presentation. Here it is clear that functional role is a causal matter, that theories are to be thought of

[16] In one sense this is a form of multiple realization; however, normally this term is applied where a single higher-level theory is realized by fundamentally different lower-level theories, rather than by the same lower-level theory in different-but-related ways.

[17] See (Lehmkuhl, 2008).

[18] It is not obvious to us — given the symmetries of Newtonian gravitation, and the possibility of reformulating it in Newton-Cartan terms (see, e.g., (Malament, 1995; Knox, 2014; Wallace, 2020) and references therein) — that 'gravitational field' is any more coherent in non-relativistic than relativistic gravity; exploring this further lies beyond the scope of this paper.

syntactically, and that functional roles are identified via Ramsification of sentences formulated in the predicate calculus, and there is no mention of, and no obvious space for, any notion of approximation. Many contemporary philosophers of physics would doubt that causation cuts deep enough to be relied upon in physics contexts, agree that physical theory is rarely neatly expressible in syntactic terms, and recognize that reduction almost invariably involves some degree of approximation. When such philosophers of physics advocate Lewisian functional reduction for physics they are advocating a variation on a Lewisian theme that they nonetheless take to be in the spirit of causal-role functionalism.[19]

What kinds of variation can we allow here? This is to some extent a matter of taste: your ability to hear the original tune in a variation may differ from your neighbour's. To our ears, one harmless variation moves away from the causal aspect of causal-role functionalism. Even those not inclined towards a Russelian view of causation will acknowledge that the roles played by theoretical terms in physics theories are not always causal. In a physics context, causal-role functionalism might be better named 'dynamical-role functionalism' in order to capture the non-causal aspect of theoretical roles. Taken alone, this move stays close to the Lewisian spirit, and indeed to Lewis's view on theoretical terms (Lewis, 1970), which requires a division of theoretical language into antecedentally understood terms ('O-terms') and terms to be functionalised ('T-terms') but does not require that the O-terms concern causal relations.

Beyond that, things get more controversial. But we have attempted here to draw a key distinction between causal-role and constitutive functionalism, and this distinction remains salient in the physics context. The core of causal-role functionalism as the claim that not just anything can successfully realize a functional role. It is central to causal-role functionalism that it is a defeasible scientific hypothesis in any given concrete example: it must be logically possible (as with the absurdly-large lookup table, for instance) that the functional hypothesis gets the empirical data right but is still wrong because after all nothing realizes the hypothesized physical structure.

By contrast, the constitutive functionalist makes no such stipulation. In the Dennettian tradition, functionalist statements are analytic: to function as if I have beliefs and desires is to have beliefs and desires, regardless of how these are in fact realised in the system at hand.

So the challenge for any variation of causal-role functionalism is to develop a notion of realization rich enough to do justice to physics, yet not so rich as to make realization analytic. As we will illustrate, this is not easy.

For instance, a routine feature of inter-theoretic relations in physics is that they are casual about ontological categories: objects at the higher level, for in-

[19]Butterfield and Gomes are the most vocal current advocates of a Lewisian approach. Their presentation is syntactic, but explicitly makes space for approximation. Others (here we draw on conversations with Nick Huggett and Henrique Gomes; for a recent statement in print, see (Lorenzetti, 2023)) hold that functional reduction can find a home in a semantic presentation of theories.

stance, are identified with functions over objects (or over properties, etc.) at a lower level. This is awkward within conventional causal-role functionalism, because a function over a class of concrete objects is not itself a concrete object. Even if we allow very permissive tools like unrestricted mereological composition, it won't give us *functions* of objects and properties in the lower-level ontology.

So that suggests a variation that stays within the syntactic view of theories, but allows the functional role to be filled by mathematical expressions - perhaps functions of lower-level variables. The identities offered via functional reduction now connect mathematical objects (like functions!), not just physical ones.

This makes the metaphysics of causal-role functionalism quite weird. Functions, as traditionally understood, are abstracta, and not the kind of thing normally allowed to stand in causal or dynamical relations with concrete objects. A hardline Quinean will be insouciant: we have to quantify over mathematical objects anyway, so they are in our ontology, so who cares whether we attach the label 'abstract' to them (Quine, 1948). But few philosophers of science are so relaxed about the abstract/concrete distinction.

But more importantly, this variation makes functional reduction all too easy. As Mark Wilson (1985) pointed out some time ago, we have reason to believe that our mathematical tools are flexible enough to achieve the appropriate connection between *any* compatible set of theoretical descriptions. And so the central idea of causal-role functionalism as a defeasible hypothesis drops away. Looking back to the philosophy of mind context, if causal-role functionalism is liberalized to this degree then even lookup tables will after all realize beliefs and desires: there is bound to be *some* mathematical function on the space of lookup table states that realizes their roles.

What else might we try? Instead of adding mathematics to the syntactic picture, perhaps the causal-role functionalist could embrace a semantic view of theories. (Among other advantages, this view is much friendlier to shifts of ontological categories, and to the employment of mathematical constructions - cf (Wallace, 2021).) Moving in this direction, however, creates other difficulties for the Lewisian view. Suppose we accept that theories are classes of mathematical models, and that relations between theories are then relations between the models in those classes. In particular, if (i) a lower-level theory T_L describes a certain system, and (ii) a higher-level theory T_H also describes that same system (perhaps more coarsely, or in more restricted circumstances), then that must force a mathematical relation between T_L and T_H: any fact about a model in T_H must be derivable from the totality of facts about some model of T_L, on pain of inconsistency between the two descriptions.

But now what is left for causal-role functionalism to do? Recall: the core idea of Lewisian reduction is to find some *articulated*, *term-by-term* relation between entities, properties or relations at one level and at another, something that constitutes a substantive scientific hypothesis and might be proven wrong. There is no obvious space in the semantic view of theories for this to happen — except by requiring the reduction to relate not just the mathematically-

characterized models but their respective linguistic explications, which returns us to a syntactic conception of reduction. On a semantic conception of theories, the bare claims that a system is described at one level by one theory and at another level by another, and the establishment of mathematical consistency between these claims, exhausts inter-theoretic reduction.

To illustrate, consider again the philosophy-of-mind context. Suppose that it is true that a certain creature is well described at one level by the intentional stance. Suppose also that the creature can be described at a finer level by some neurological, or perhaps microphysical, model. Basic consistency requires that any fact about the higher-level description is determinable by the totality of facts at the lower level, but nothing whatever follows about how, if at all, that determination can be articulated in terms of a finer-grained, term-by-term, relation between neurological concepts and behavioral ones, no matter whether the theories are characterized syntactically, semantically, or as a hybrid of the two.

In summary, the causal- (or dynamical-)role functionalist faces a dilemma once they accept the limitations of the syntactic view of theories as applied to physics. The whole essence of their view is that the success of a higher-level theory, once functionalized, implies a compelling, but defeasible, set of hypotheses about the lower level. But if they do not restrict the nature of the relations they hypothesize between the two levels, any such relation is analytic, and so fails to go beyond what the constitutive functionalist seeks, or to say anything in particular about what is happening at a lower level. And if they do restrict it, it is hard to see what restriction is available that does not flatly fail in paradigmatic physics examples.

7 Conclusions

Where does this leave functionalism in physics? Causal-role functionalism has a venerable pedigree in the philosophical literature, and the advantage of great logical clarity. It offers a recipe by which we might seek reduction in physics, and criteria by which we can judge which reductions are metaphysically acceptable. However, it does not appear to be a good match for actual physical practice. For one thing, the methodology it offers for reduction, in which we start with a higher-level theory and functionalise its terms, is not the one that appears in standard examples of textbooks. For another, it is closely tied to a syntactic view of theories, and there is reason to doubt both that this is the right general view of theories, and that reduction proceeds by relating theories conceived of syntactically. Attempts to tie causal-role functionalism to a semantic view of theories fail to retain the core commitments of the view.

In light of this, we would like to rehabilitate an alternative form of functionalism - that which we've here called constitutive functionalism. On this view, functionalist statements like "spacetime is what spacetime does" are true analytically - a commitment to functionalism about a concept alongside the fact that some system functions in the relevant way guarantees the application of

the concept. For this thesis, it does not matter what realises the relevant functional role, and there are no restrictions on how this realisation comes about.

The move to constitutive functionalism may not offer a recipe for reduction, but it offers interesting new avenues for understanding our physical theories. There is much foundational work to be done in understanding how our theories fit together in such a way that one or another functional role is or is not satisfied. This kind of functionalism has a history in the philosophy of physics - it is this kind of functionalism that Wallace appeals to in his "Everett and Structure"Wallace (2003), and it also makes good sense of spacetime functionalist projects like Knox's (2014; 2019; 2011), which aim at understanding existing spacetime theories, rather than reducing them to underlying theories of quantum gravity. If causal-role functional reduction fails in familiar cases, it would be surprising if it succeeded in the alien domain of quantum gravity.

Acknowledgements

We would like to thank Henrique Gomes, Nick Huggett, Lorenzo Lorenzetti, Katie Richardson, and the Spring 2023 Fellows of the Pittsburgh Center for Philosophy of Science for useful discussions. The ideas in this paper were presented in Oxford, Irvine, Pittsburgh, Birmingham and London: many thanks to all these audiences for their contribution. Additional thanks go to Katie Richardson for the title of the paper. This work was completed during Knox's Fellowship at the Pittsburgh Center for Philosophy of Science: she would like to thank the Center for the opportunity.

References

Albert, David Z. 2015. *After Physics*. Cambridge, MA: Harvard University Press.

Balescu, Radu. 1997. *Statistical Dynamics: Matter out of Equilibrium*. London: World Scientific.

Block, Ned. 1978. Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science*, **9**, 261–325.

Braddon-Mitchell, David, & Nola, Robert (eds). 2009. *Conceptual Analysis and Philosophical Naturalism*. Boston: The MIT Press.

Butterfield, Jeremy, & Gomes, Henrique. 2020. Functionalism as a species of reduction. *arXiv preprint, arXiv:2008.13366*.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Dennett, Daniel C. 1987. *The Intentional Stance*. Boston: MIT press.

Dennett, Daniel C. 2005. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Boston: MIT press.

Fodor, Jerry A. 1975. *The Language of Thought*. Hassocks: Harvester Press.

Guo, Bixin. 2020. *Two Approaches to Reduction: A Case Study from Statistical Mechanics*. *Philosophy of Science*, forthcoming: preprint at http://philsci-archive.pitt.edu/20433/.

Kim, Jaegwon. 2007. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Knox, Eleanor. 2011. Newton-Cartan Theory and Teleparallel Gravity: the force of a formulation. *Studies in History and Philosophy of Modern Physics*, **42**, 264–275.

Knox, Eleanor. 2014. *Spacetime Structuralism or Spacetime Functionalism?* Unpublished, available at http://philsci-archive.pitt.edu/22630/.

Knox, Eleanor. 2019. Physical relativity from a functionalist perspective. *Studies in History and Philosophy of Modern Physics*, **67**, 118–124.

Lam, Vincent, & Wüthrich, Christian. 2018. Spacetime is as spacetime does. *Studies in History and Philosophy of Modern Physics*, **64**, 39–51.

Lehmkuhl, Dennis. 2008. Is spacetime a gravitational field? *Pages 83–110 of:* Dieks, Dennis (ed), *The Ontology of Spacetime II*. Elsevier.

Lewis, D. 1980. Mad pain and Martian pain. *Pages 216–222 of:* Block, Ned (ed), *Readings in the Philosophy of Psychology*. Cambridge, MA: Harvard University Press. Reprinted with an additional postscript in D. Lewis, *Philosophical Papers, Volume I* (Oxford: Oxford University Press, 1983), pp.122–132.

Lewis, David. 1970. How to define theoretical terms. *The Journal of Philosophy*, **67**(13), 427–446.

Lewis, David. 1972. Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, **50**(3), 249–258.

Lorenzetti, Lorenzo. 2023. *Two kinds of functional reductionism in physics*. Preprint: http://philsci-archive.pitt.edu/21563/.

Malament, D.. 1995. Is Newtonian Cosmology Really Inconsistent? *Philosophy of Science*, **62**, 489–510.

Menzies, Peter. 1996. Probabilistic causation and the pre-emption problem. *Mind*, **105**, 85–117.

Quine, Willard V.O. 1948. On What There Is. *The Review of Metaphysics*, **2**(5), 21–38.

Raatikainen, Patu. 2021. Troubles with the Canberra Plan. *Synthese*, **199**, 4039–4060.

Ryle, Gilbert. 1949. *The Concept of Mind*. Chicago: University of Chicago Press.

Searle, John. 1980. Minds, Brains and Programs. *Behavioral and Brain Sciences*, **3**, 417–457.

Thorne, Kip S., & Blandford, Roger D. 2017. *Modern Classical Physics: Optics, Fluids, Plasmas, Electricity, Relativity, and Statistical Physics*. Princeton: Princeton University Press.

Wald, Robert M. 1984. *General relativity*. Chicago: Chicago University Press.

Wallace, David. 2003. Everett and structure. *Studies in History and Philosophy of Modern Physics*, **34**, 87–105.

Wallace, David. 2012. *The emergent multiverse: Quantum theory according to the Everett interpretation*. Oxford: Oxford University Press.

Wallace, David. 2020. Fundamental and emergent geometry in Newtonian physics. *The British Journal for the Philosophy of Science*, **71**, 1–32.

Wallace, David. 2021. Stating structural realism: mathematics-first approaches to physics and metaphysics. *Pages 345–378 of:* Hawthorne, John (ed), *Philosophical Perspectives Volume 36: Metaphysics*. Wiley-Blackwell.

Wilson, Mark. 1985. What is this Thing Called Pain?-The Philosophy of Science Behind the Contemporary Debate. *Pacific Philosophical Quarterly*, **66**, 227–267.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. 3rd edn. Oxford: Blackwell. Translation by Elizabeth Anscombe.

Zabawski, Evan. 2009. Oobleck the dilatant. *Tribology and Lubrication Technology*, **65**, 6.

Zwanzig, Robert. 2001. *Nonequilibrium Statistical Mechanics*. Oxford: Oxford University Press.