Triviality Arguments Against Functionalism*

Peter Godfrey-Smith
Harvard University

Philosophical Studies (2008).

1. Introduction

Functionalism in the philosophy of mind holds that systems with mental properties have them in virtue of the systems' functional organization, and particular mental states such as pains and hopes are functionally individuated internal states. "Triviality arguments against functionalism," as I will call them, hold that the claim that some complex physical system exhibits a given functional organization, or is in a particular functional state, is either trivial or has much less content than is usually supposed.

I group in this category a family of arguments with different ambitions. An early version is attributed to Ian Hinckfuss, in discussion in the 1970s.[1] The "Hinckfuss' pail" argument is often described by saying that a bucket of water sitting in the sun has so much causal complexity that, via a suitable categorization of states, it can be taken to realize the functional organization of a human agent. Searle (1990) gave a triviality argument against computationalism about the mind, asserting that there would be some

* I am grateful to David Chalmers, Alan Hájek, Peter Koellner, William Lycan, Susanna Rinard, Nick Shea, and a referee for the journal for very helpful comments on earlier drafts. The paper also benefitted from audience comments during presentations at the Australian National University and Oxford University.

[1] According to Cleland (2002), Hinckfuss' argument was presented in a 1978 discussion of computation at the Australasian Association of Philosophy. Lycan (personal communication) says the discussion was during presentation of an early version of Lycan (1981) at the conference, a paper which then appeared with a presentation of the Hinckfuss argument. Lycan treats the argument as something different from a triviality argument in my sense, however; Lycan says the bucket of water *might*, by chance, come to realize a human's functional organization over some interval.

legitimate way of interpreting his wall as running the program Wordstar, or any other computer program. Neither Hinckfuss nor Searle gave proofs of these claims.

Putnam (1988) gave two arguments. The more important one claims that functionalism collapses into logical behaviorism. Once we know that a system has the input-output properties associated with some particular functional organization, we know that it can be interpreted as having that functional organization as well. Putnam did offer proofs. Chalmers (1996) criticized Putnam's arguments but developed several modifications of them. He takes some of these arguments to have surprising conclusions, but denies that they endanger computationalism or functionalism about the mind. Copeland (1996) gave a formal treatment of Searle's argument.

So some of these arguments take functionalism as their target, while others focus on computationalism. Some are given with proofs, while others rely on intuitions. Some conclude that a functional or computational description is entirely trivial, when applied to a sufficiently complex system, while others conclude only that functional descriptions collapse into behavioral descriptions.

Given the threat that such arguments pose, it is surprising how little they have been discussed, especially as the mainstream functionalist literature does not use accounts of the realization of functional structures that make it *clear* that triviality problems do not arise. Many accounts of realization used by functionalists are so schematic that it is uncertain how these problems are handled.

Here I examine the problems posed by such arguments for mainstream functionalism in philosophy of mind. The discussion covers functionalist accounts of both "folk" and scientific psychological properties, but the analysis of computation is regarded as a separate question.[2] I also do not consider teleological versions of functionalism that understand function in a rich biological sense.[3] The discussion is restricted to what might be called "dry" functionalism, of the kind seen in Fodor (1981),

---

[2] Although it is difficult to say exactly what computationalism about the mind is committed to, it is intended to be a stronger claim than functionalism (Smith 2002, Piccinini 2004). Computationalism is supposed to involve a claim about particular characteristics of the functionally characterized operations that comprise cognition.

[3] Lycan (1981) treats this as part of the answer to Hinckfuss' pail.

Stich (1983), Braddon-Mitchell and Jackson (1996), Crane (1995), and many others. The arguments are also presented within a larger framework associated with mainstream functionalism; I employ a picture of physical systems and a set of idealizations associated with that project.

After an initial discussion of functional characterization, I present three arguments (sections 2 to 4). Two are modifications of arguments seen in Putnam and Chalmers. They are about the "surplus content" that functional description has over behavioral description. The third argument, which is new, is a more precise treatment of the stronger claims associated with Hinckfuss and Searle. I then argue (in section 5) that while the threat raised by these problems has been underestimated, it is not fatal. Avoiding the problems is possible, but has consequences and costs. These include the revison of popular claims about the "autonomy" of functional description.

2. Realization of an FSA

There are two ways of talking in a precise way about functionally characterized states and their realization. One, associated with the computationalist style of functionalism, specifies a functional profile as a set of relations between abstract entities, and understands realization in terms of a mapping between abstract and physical structures. The other, deriving from Lewis (1972), uses a Ramsey sentence or something similar. According to the second approach, an interlocking set of hypotheses specify a set of roles, and various objects may be *occupants* of those roles. The tools used in this paper are borrowed from discussions of computation, but the arguments are supposed to work within either approach. (In the final section I discuss the relations between the two.)

My starting point is a framework in which functional roles are described using "state transitions." Initially we look at cases where the system is taken to be in a single total functional state at each time. A *finite state automaton* (FSA) is a finite set of inputs, outputs, and inner states that are related by rules that map each combination of present inner state and input to a new inner state and output.[4] This is one way of formalizing the

[4] Technically, this is a "Mealy machine," not a "Moore machine," as the outputs are associated with transitions rather than states.

kind of functional characterization often envisaged in the days of "machine functionalism" (Putnam 1960, Block and Fodor 1972). I will use the following example of an FSA, which corresponds to a simple coke machine.

Coke Machine State Transitions

$(S_1, I_1) \rightarrow (S_2, O_1)$ $\quad$ $(S_1, I_2) \rightarrow (S_3, O_1)$

$(S_2, I_1) \rightarrow (S_3, O_1)$ $\quad$ $(S_2, I_2) \rightarrow (S_1, O_2)$

$(S_3, I_1) \rightarrow (S_1, O_2)$ $\quad$ $(S_3, I_2) \rightarrow (S_1, O_3)$

I begin with the following account of the realization of this kind of structure, which modifies a formulation due to Chalmers.

> **FSA**: A physical system realizes a given FSA during a time interval iff there is mapping M from states of the physical system onto states of the FSA, and from inputs and outputs of the physical system onto inputs and outputs of the FSA, such that: for every state-transition (S, I) → (S', O) of the FSA, if the physical system were to be in state P and received input I* such that M(P) = S and M(I*) = I during this time interval, then it would transition to state P' and would emit output O* such that M(P') = S' and M(O*) = O.

This will be called a "simple mapping" criterion for realization.

The FSA itself treats inputs and outputs abstractly. It merely distinguishes two inputs and three outputs. Any system with this number of possible inputs and outputs could, if appropriately organized, realize the FSA. When specifications of this kind are used in philosophy of mind, it is natural to require that a system's inputs and outputs be of a specific kind. Just as a coke machine has to be able to accept money and give out cokes,

---

Some early discussions of functionalism focused on Turing machines. I take Turing machines themselves to be an unpromising model for the mind, though important for *in principle* discussions of the mechanization of intelligence. The CSA framework, discussed below, can be used to represent Turing machines, as Chalmers (1996) notes.

an intelligent agent, perhaps, has to be able to track and act on the world in particular ways. If this is right, then an FSA understood as a mathematical object only specifies the formal backbone of a functional structure in the sense relevant to philosophy of mind (Block 1978).

So where necessary, I will distinguish between an FSA in a *broad* sense and in a *narrow* sense. An FSA in the broad sense includes specification of particular inputs and outputs; the FSA in the narrow sense is just the formal backbone, with inputs and outputs treated abstractly. Specifying the coke machine FSA in the broad sense includes giving both the state transitions above, and the following specification of inputs and outputs (complete with anachronistic mid-twentieth century pricing).

Coke Machine Inputs and Outputs:
I_1 = 5 cents; I_2 = 10 cents; O_1 = null; O_2 = coke; O_3 = coke & 5 cents.

This causes a complication in discussions of realization. If realization is understood in terms of a mapping that preserves relations, there must be abstract objects with mathematical relations between them on one side, and physical objects with causal or other physical relations between them on the other. But a mapping criterion for realization can then only be applied directly to the narrow sense FSA. Showing the realization of a broad sense FSA then involves two steps. One is showing a mapping between the narrow sense FSA and the physical system. The other is showing that the inputs and outputs of the physical system are of the right kind.[5]

In general, broad sense FSAs will be more important than narrow sense FSAs below, and the symbols "I" and "O" will be used for inputs and outputs in a concrete sense except where otherwise indicated.

[5] If a functionalist does *not* see the functional roles relevant to philosophy of mind as involving specific kinds of inputs and outputs, perhaps because of cases of humans with unusual interfaces with the world, then the mapping approach can be used on its own. This makes Hinckfuss-type arguments more threatening. This issue will be discussed in section 3.

The first triviality argument aims to show that any sufficiently complex system with the input-output dispositions associated with a given FSA (broad sense) is also a realization of the FSA. My presentation uses graphical methods, and modifies arguments due to Putnam (1988) and Chalmers (1996). I represent FSAs using what I will call *contingency trees*. The tree for the coke machine is represented, over three rounds of input and output, in Figure 1.

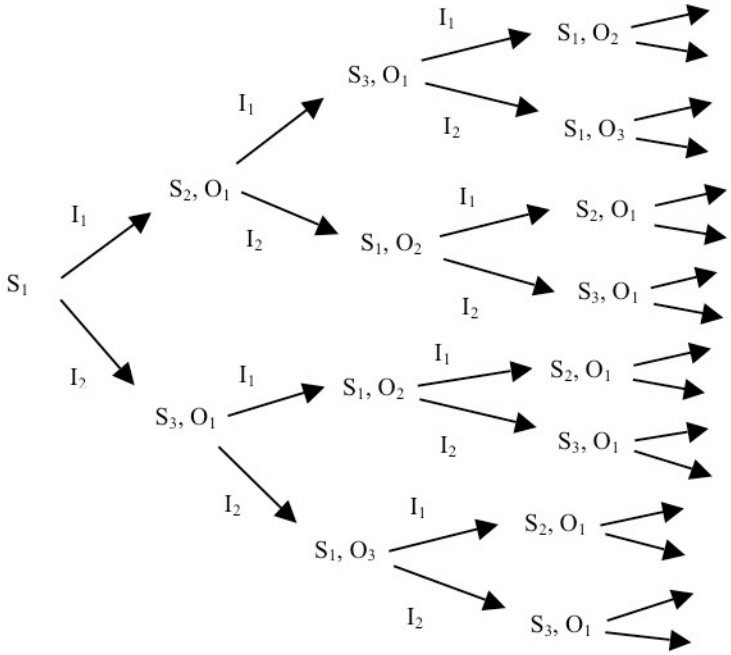

Figure 1: Contingency tree for the coke machine FSA

The same kind of tree can be used to represent the dispositional properties of a physical object. If we specify a set of inputs and outputs, any physical system starting in a particular state will have *some* set of dispositions in response to those inputs (if it can interact with them at all), and it may (or may not) emit outputs of the relevant kind. Such a tree will usually only represent some of the object's dispositional properties, though we can imagine a "full" tree for an object over an interval, representing all the influences the system might receive from the rest of the world, and all the ways it would affect its environment in response.

Suppose we have a physical object whose input-output dispositions over some time interval are the same as those associated with a particular FSA. The object goes through some particular sequence of states, emitting the right outputs in response to a series of inputs, and was also disposed to emit suitable outputs if its input history had been different. Assume also that these dispositions are the product of the internal structure of the object, as opposed to being mediated by an external controller. The contingency trees for the FSA and the physical object can then be superimposed on each other. (Imagine doing this with transparent slides.) The only differences are found in the inner states represented at each point. Provided that the physical system has sufficient overall complexity, there will be a mapping between physical states and FSA states such that the physical system is a realization of the FSA.

The relevant sense of "sufficient overall complexity" is as follows. The system's total physical state at every point on the tree is unique. The system starts in a particular state, P_1, and each input it may receive will not only send it into a new unique state, but also affect the system's physical response to later inputs, perhaps in only microphysical ways. As a consequence, every sequence of inputs will send the system down a path comprising physical states that cannot be reached by any other sequence of inputs.[6] Putnam offers reasons why actual-world physical systems should satisfy a principle of this kind. Instead I treat this as a condition in my claims about the realization of functional structures. *If* a physical system has this kind of overall complexity and has the same input-output properties as a given FSA, then it is a realization of that FSA. Below all the physical states labeled as P_i are assumed to be distinct from each other.

From here it is simple to show the existence of a mapping of the relevant kind. Using the two superimposed trees, we map each FSA inner state to a coarse-grained physical state that is specified with a disjunction of all the unique physical states whose positions on the physical tree correspond to occurrences of that inner state in the FSA tree.

---

[6] The uniqueness claim here is intended to apply to *intrinsic* properties of the system, to avoid it collapsing into triviality. I assume an account of intrinsicness along the lines of Langton and Lewis (1998).

More formally, we inspect the FSA tree and note all locations at which a particular state, perhaps S_1, appears. Each location can be identified independently of the inner state that appears there, by its place in the space of possible input sequences. Call the set of locations at which S_1 appears $\sum_1$. We then inspect the physical tree, and note all the physical states that appear in locations in $\sum_1$. This set of physical states is described with a disjunction, labeled Q_1. We do the same for the other FSA states. We can then construct yet another contingency tree for the physical system, where each node on the tree is characterized by an output and a disjunction of physical states created by the method outlined above. The information in this tree is a weakening or coarse-graining of that contained in the original physical tree. Each Q_i will appear in all and only the locations occupied by S_i in the FSA tree. As the transition properties of any FSA state are expressed entirely by its set of locations in the tree, the Q_i's can be seen to have the same transition properties as the S_i's that they are mapped to.

The claim that that the Q_i's have the same "transition properties" as the S_i's is subject to a qualification discussed below. Before addressing those complications I will work through the coke machine example to illustrate the procedure.

The contingency tree for the coke machine FSA was given in Figure 1. A complex physical system, arbitrarily chosen, with the same input-output properties, will have the following contingency tree (again assuming an interval over which three inputs are received.)

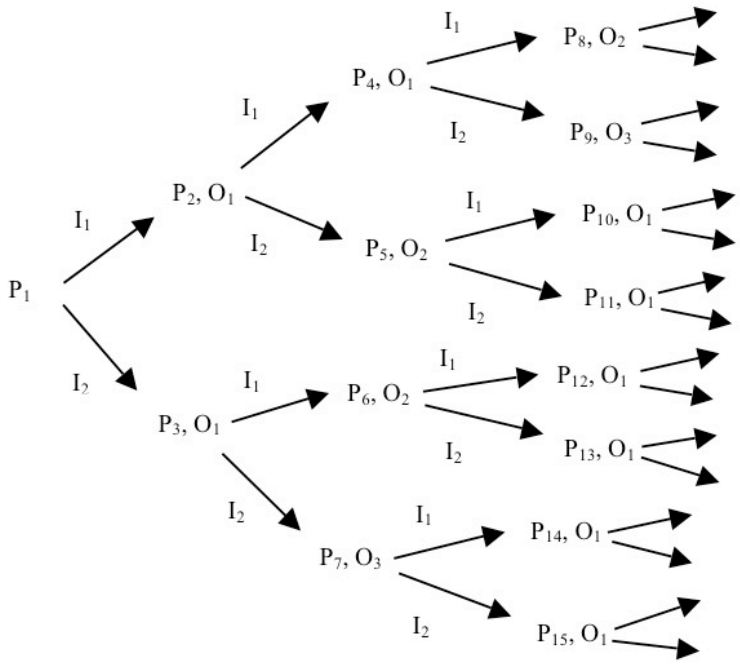

Figure 2: Contingency tree for a physical system that has the behavioral properties of the coke machine

Behavioral criteria are used to "align" the FSA with the physical system at the start of the interval. This aligment may yield any FSA inner state as the initial state, though here I assume the initial state is S_1. We superimpose the trees and map each formal state to a disjunction of physical states as follows.

Let:

$Q_1 = P_1 \vee P_5 \vee P_6 \vee P_7 \vee P_8 \vee P_9$

$Q_2 = P_2 \vee P_{10} \vee P_{12} \vee P_{14}$

$Q_3 = P_3 \vee P_4 \vee P_{11} \vee P_{13} \vee P_{15}$

Then for all $i \in \{1, 2, 3\}$, $M(S_i) = Q_i$

The Q_i's can be used to generate a coarser-grained physical contingency tree, a fragment of which can be seen in Figure 3.

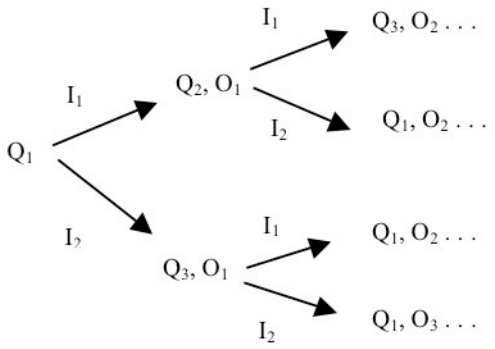

Figure 3: Coarser-grained contingency tree for the physical system in Figure 2.

This procedure can be used to show both "good" and "bad" realizations of an FSA by a physical system. What is the difference? A first response might be that the "good" cases are those where the physical states disjoined to produce a Q that maps to some formal state are *similar* in some non-trivial sense. In a real-world system, we need not suppose that these physical states are *identical*. But in natural realizations of an FSA, the system should be doing something physically similar when it reaches the various states that each map to some S_i. In the coke machine case, S_3 is especially useful to think about in this connection. The FSA has two different ways of reaching S_3 within a single cycle, depending on the inputs received. Do the physical states the system reaches via these two different paths have anything in common? If not, the FSA representation of that device seems at least somewhat misleading. Explanations of the behavior of the machine in terms of the role of S_3 in its functional economy would imply a spurious unity across the processes in which that state is involved. This response to the problem, which I think is indeed along the right lines, will be fleshed out in more detail in section 5.

Before moving on I will discuss some complications and possible objections. These involve the question of whether procedure above really shows that the system satisfies the conditionals used in the criterion for realization. The problem has two aspects. First, as the treatment above assumes a particular time interval over which a restricted number of inputs can be received, each Q_i as defined above will contain "terminal" physical states, states obtaining at the end of the interval. The analysis given does not suppose that we know how the system will respond to inputs when in these states. So when it is claimed that the system is causally disposed to transition from one Q to another, in a way that corresponds to the FSA state transitions, the simplest way of understanding these dynamical relations between the Q states is not available. That simplest option would require that, if the coarse-grained tree has it that if the system is in Q_i and receives input 1 it will transition to Q_j, then it must be the case that for *every* state disjoined in Q_i, if the system is in that state and receives input 1 then it will transition to some state in Q_j. This does not apply to terminal states in Q_i. So the analysis above requires that conditionals describing how the system is disposed to move from one Q state to another describe the system's dispositions over the interval even when the antecedent Q's include terminal states.

I will argue that these conditionals are true, under an interpretation that is appropriate for functionalism. The point can be be made by imagining a long interval and a system undergoing a definite aging process over the interval, though the point applies generally. The Q's are then very long disjunctions, containing physical states that would naturally encountered when the system is at various different ages. Each Q includes a terminal state, one that would only be reached (if it is reached at all) at the end of the interval. Suppose we have an FSA conditional that says if the system is in S_1 and receives I_1, it will transition to S_2. Then if the system realizes the FSA over the interval, Q_1 maps to S_1, and Q_2 maps to S_2, it must be true that if the system were to be in Q_1 at, for example, the start of the interval, then if it received I_1 it would have to transition to Q_2. What would it involve for the system to be in Q_1 at the start of the interval? It would involve the system being in an "age-appropriate" physical state within the disjunction. In possible-worlds jargon, these are the nearest Q_1-worlds. If the system would transition

appropriately from that physical state at that stage in the time interval, the conditional is satisfied. To assess whether a system of this kind realizes an FSA over a specific time interval we do not have to know how the system would behave if it were, at the beginning of the interval, in a physical state that could only arise late in the interval. So if functionalism is intended to capture systems that physically develop or age, the relevant interpretation of the conditionals is one in which antecedents and consequents are assessed in a way that respects the temporal location of physical states in the history of the system.[7]

There is also a second problem. Although a contingency tree of the kind above specifies various non-actual paths as well as the actual one, some states of an FSA might not be reachable from the state the system was in at the start of the interval. So it will be impossible to assess whether the system respects transitions involving those FSA states. Consequently, showing the realization of an FSA in those cases requires describing the physical system with a number of different contingency trees, each beginning with different initial states. This does not create a problem for the methods used above. Provided that the physical states at all the places on all the trees are distinct from each other, the disjunctions can be constructed in the same way as before.

Before moving on I will compare this handling of the problems above with that in Chalmers (1996). Chalmers uses a more elaborate specification of an arbitrary physical realizer for an FSA. He assumes that the physical device keeps a record of its "input history," the complete sequence of inputs it receives from some initial state. Chalmers also assumes the system has a "dial" that can be permanently set to a particular value at the start of any run. Such a device has the capacity to enter an indefinitely large number of distinct inner states, each specifiable in advance in a way that pairs it with a particular input history and a particular initial state. The "dial" feature enables us to describe how a

---

[7] As a referee pointed out, this has the consequence that a system might have the dispositions to transition (given suitable input) from S_1 to S_2 at one time step and (also given suitable input) from S_2 to S_3 at that same time-step, without being disposed to transition from S_1 to S_2 and then to S_3, given those inputs in series over multiple time-steps. But this result is appropriate, as it may well be that one consequence of receiving either input at the first time-step is to disable the system with respect to further transitions.

given input history could be experienced from many different initial states. If the system has the right behavioral dispositions as it traverses all of these possible physical states, indefinitely long disjunctions can be constructed for use in mappings of the kind discussed above. Counterfactuals can be specified about what would have happened if the system had begun the time interval in a different state, including a state that was, in the actual world, terminal.

Chalmers' method achieves a more straightforward treatment of the conditionals linking Q states, by making richer assumptions about both the internal structure and behavioral dispositions of the physical realizer. On the behavioral side, Chalmers imagines a realizing system that exhibits the right behavioral responses over indefinitely long series of inputs. I assumed only that a system has the right behavioral dispositions over an interval which begins with either the actual initial state or one of a set of relevant alternatives. I keep Chalmers' procedure on the table, as an alternative to the simpler argument given above. But I do not think either the behavioral or structural assumptions Chalmers makes are necessary for the construction of a triviality argument that raises problems for functionalism. If a system has a recorder of its input history and a dial, of the kind Chalmers describes, then we can know in advance that the system must enter unique states during the time interval under consideration, will do so for both actual and non-actual input histories, and would do so from relevantly different initial conditions. But a system need not have those features in order for it to be *true* that it is disposed to enter unique physical states under all those circumstances. All that the extra features give us is the possibility of labeling the states compactly and in advance. As long as a physical system *has* such a range of states (and the behavioral asssumptions are met), there is the possibility in principle of constructing a mapping that shows it is a realization of the FSA.

3. Realization of a CSA

The preceding argument does not constitute a threat to contemporary functionalism. The only position threatened so far is a form of the now-outdated "machine functionalism." An FSA is in a single internal state at any time, and it is usually thought that a viable view in philosophy of mind must recognize that a cognitive system is in more than one

mental state at a time. Behavior is the consequence of interactions between several simultaneously present mental states (beliefs, desires, moods...), along with sensory input. Chalmers' 1996 discussion includes a sketch of a Putnam-style triviality argument for this case. I will present an argument of the same kind in a different form.

Chalmers introduces the "CSA" formalism (for *combinatorial state automaton*). The total inner state of a system is now represented as a vector, or list, of substates. Transitions have the form: $(<C_{11}, C_{21}, C_{31}>, I_1) \rightarrow (<C_{12}, C_{21}, C_{31}>, O_1)$. Though Chalmers uses this category in an analysis of computation, I take the CSA concept to be a good description of the kind of structure that the mainstream functionalist literature often has in mind.

The criterion for realization of a CSA is different from that for an FSA, as something in the physical system must be mapped to every *substate* that figures in the state transitions. An initial account of realization might be given as follows:

> **CSA**: A physical system realizes a given CSA during a time interval iff there is a mapping from states P of the physical system onto substates C of the CSA, and from inputs and outputs of the physical system onto inputs and outputs of the CSA, such that: for every state-transition $(<C_1, C_2,...C_n>, I) \rightarrow (<C'_1, C'_2,...C_n'>, O)$ of the CSA, if the physical system were to be in a combination of states $<P_1, P_2...P_n>$ that map to $<C_1, C_2,...C_n>$ during this time period, and received input I* that maps to CSA input I, then it would transition to a combination of substates $<P_1', P_2'...P_n'>$ that map respectively to $<C'_1, C'_2,...C_n'>$, and would emit an output O* that maps to CSA output O.

Again we can distinguish between narrow and broad sense CSAs. The narrow sense treats inputs and outputs abstractly. (The **CSA** criterion above uses the symbols "I" and "O" in this abstract sense.) A broad sense CSA specifies what the system's inputs and outputs should actually be. So showing that a physical system realizes a broad sense CSA involves both showing a mapping between the physical structure and the narrow sense CSA, and showing that the inputs and outputs are of the right kind.

The natural-looking way to map physical features of the system to CSA substates is to take different regions within the physical system to implement different substates. We might treat substate C_{12} as mapping to a particular state, 2, that region 1 can be in. But the criterion for realization given above also allows other mappings. Without some extra constraint, a CSA is realized by any sufficiently complex physical system with the right input-output properties. As before, "sufficiently complex" means that the system is in a unique total physical state at every time-step, a state dependent on the history of inputs received.

The argument is similar to that in the previous section. We first represent the CSA as a contingency tree, as in Figure 4.

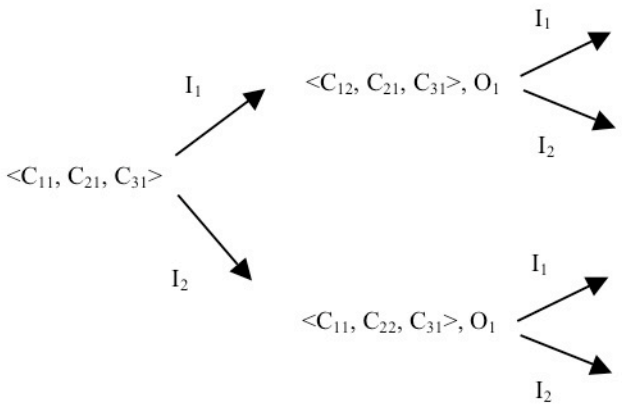

Figure 4: Contingency tree for a CSA

It is possible, as we saw earlier, to represent any physical system with the same input-output properties as a tree also. If we superimpose the contingency trees for the physical system and the CSA, they will differ only in the internal states at each node. On the physical tree we have total physical states. On the CSA tree we have vectors of substates.

To construct a mapping of the right kind, we represent each physical state of the system as a conjunction of disjunctions of total physical states. We also "coarse-grain" the physical contingency tree, as in the previous section. The aim is to show that a

particular disjunction of physical states can occupy locations in the physical tree that correspond exactly to the locations occupied by a particular substate in the CSA tree. Vectors of CSA substates correspond to conjunctions of disjunctions of physical states.

The disjunctions are constructed as follows. Consider the first state of the first CSA substate variable, C_{11}. We look at all the locations on the CSA tree where the system is in a total state that includes C_{11}. Call this set of locations on the CSA tree $\sum_{11}$. We then look at the physical tree, and note the total physical state of the system at all locations in $\sum_{11}$. These physical states are collected and labeled with a disjunction, Q_{11}.

Then we begin construction of a coarse-grained physical tree, where the physical state of the system is specified at each location with a conjunction of disjunctions. We make Q_{11} the first member of the conjunctive specification of the system's state in the coarse-grained tree at every location in $\sum_{11}$. So Q_{11} will appear as the first conjunct at every location in $\sum_{11}$, and nowhere else. This is the same as the set of locations where C_{11} appears as the first member of the CSA state vector. The procedure can be repeated for C_{12}, C_{21}, and all the other CSA substates. This shows that CSA substates can be mapped one-to-one to disjunctions of physical states such that all relations between states and inputs and outputs on the tree are preserved.

This argument assumes that issues arising with terminal states and alternate initial states are handled in the same way described in the previous section. As before, it is also possible to make richer assumptions about the physical realizing system, as Chalmers does, which ensure that an indefinite number of unique physical states which the system might enter can be specified in advance.

The obvious initial response to this problem is to add, as a constraint on realization, a feature that was mentioned earlier. The CSA substates operate in explanations as independent parts of the system, which combine to give rise to the behavior of the whole. So it seems natural, as a next step, to require that the physical states which map to CSA substates be states of independent components of the physical system. In the final section I will argue that this is indeed the right next move, but it does not solve the entire problem. It needs to be combined with a similarity contraint of the kind discussed in the FSA case. Before looking at solutions, however, I will discuss one

more triviality argument. This one does not assume a realizing system with particular input-output properties. It uses the framework above, derived from Putnam and Chalmers, to make a stronger challenge to functionalism of the kind associated with Hinckfuss and Searle.

4. A New Triviality Argument

An initial outline of the argument is as follows. Any sufficiently complex physical system can be made into a behavioral duplicate of an intelligent agent, via a change to the "transducer layer" of that system. Changing the transducer layer of a physical system alone ought not alter whether or not it has at least some basic mental properties. But drawing on earlier arguments, we know that if a complex physical system can be given the *behavioral* profile of an intelligent agent, it is thereby made to realize the *functional* profile of that agent, if realization is understood using a simple mapping criterion. As a change to transducer layer should not alter whether a system has mental properties, every complex physical system *already* has the functional features that give agents like you and I mental properties. A change to transducer layer may change *which* mental properties we have, but ought not to change whether we have mental properties at all. Consequently, functionalism combined with a simple mapping account of realization collapses into triviality. Every complex physical system has functional properties sufficient to give it some mental properties of the kind found in paradigm human agents.

I now discuss the premises in more detail. I begin with the idea that any functionally characterized physical system whose operations link it with its environment can be broken down into (what I will call) a *transducer layer* and a *control system.* The transducer layer is the interface between the system and its environment. On the input side, the transducer layer responds to physical impacts in a form that the rest of the system can use in further processing. Parts of the normal human transducer layer include the retina, which responds to electromagnetic radiation with neural firings, and hair cells in the inner ear, which respond to physical vibrations with neural firings. On the output side, muscle fibers are part of the transducer layer, as they respond to motor neuron firings with contractions.

The term *control system* will be used for everything that is functionally important in a system other than the transducer layer. This includes the basis of memory, the manipulation of representations, learning, planning, and so on.

When described in abstract terms, the role of each part of the transducer layer is mapping one physical variable to another. Its role could be represented with a look-up table. This simplicity is essential to transducer layers as I conceive them. If we are looking at a peripheral part of a system, and we find a role for memory or learning, then we have not looked peripherally enough to find the transducer layer. The simplest cases to think about are those where the transducer layer remains fixed in its input-output properties over time. Systems like ourselves, however, may be plastic with respect to these features. In that case the argument should be applied to a system over an interval in which the transducer layer remains fixed.

Often it may be unclear where the border between transducer layer and control system is. But locating the transducer layer is a consequence of locating the divide between system and environment. If there is a problem with the idea of a boundary between a functionally characterized system and its environment, that is a problem for mainstream functionalism itself. The assumptions I make about transducer layers are made within functionalism.

Earlier I took functional profiles to include a "concrete" specification of inputs and outputs. That treatment of inputs and outputs was favorable to functionalism; any system with the functional profile of a human would have to receive sensory input in human form and give outputs in such form. A bucket of water cannot possibly have the same functional profile as a human agent, as it does not have the right input-output properties. But we now look at the possibility of taking a functionally characterized system and *changing* its transducer layer, while keeping the control system intact. This is done by changing the physical devices that interface with external objects. We might alter the hair cells in the ear so they are not moved by vibrations, but by magnetic fields. We might have muscle fibers moving a mouse on a computer screen. Altering transducer layers has important therapeutic possibilities for people with sensory and motor disabilities.

When the transducer layer of an intelligent system is altered, what are the consequences for its psychological properties? Here I do not mean the changes that will result to its history of experience. Rather, I imagine looking at a snapshot of the system, or assuming preservation of the formal structure of input over an interval, and asking which changes to its mental states are logically implied by a change to the transducer layer.

There may be many psychological changes implied, but it is natural to think there are *some* mental features of an agent that depend only on the properties of the control system, and are unaffected by the properties of the transducer layer. The mental properties that are unaffected will not include the truth-conditions of its propositional attitudes. They may not include how the world seems to the agent. They may include such features as being able to learn by reinforcement, or being able to reason hypothetically. But they may also be as basic as having mental states *at all*.

Such a commitment is implicit in therapeutic work on human transducer layers. When we try to equip a disabled person with novel transducer capacities, the aim is to give better environmental interfaces to a control system that we take to have many of its mental properties – at least to have *a* mental life – independently of the features of its transducer layer. There is not taken to be a risk that altering the transducer layer will rob the system of *all* of its mental properties. A similar commitment is implicit in many science fiction stories, especially of the paranoid style of "The Matrix." This commitment could be false, of course. But a more theoretical argument can also be given. From a functionalist point of view, what the transducer layer does is quite simple. Plants and single-celled organisms have transducer layers that are in some ways similar to ours. These organisms have much less complex control systems than ours, however, and it is here that the cognitive differences between plants and humans seem to lie. The processes of reasoning, decision-making, and learning studied by psychology are much more dependent on control system features than on what happens to be transduced when the system interacts with the physical world.

This argument should not be pushed too far. It is false to think we could work our way into a system from its periphery, and argue that *whenever* we encounter a simple

input-output device, removing or changing that device cannot imply wholesale changes to whether system has mental states. Every part of an intelligent system, viewed from sufficiently close up, can be described as having a simple input-output character. But for my purposes, only a weak application of these ideas is needed. I only consider changes to a very "thin" transducer layer that is the boundary between system and environment. Further, the premise needed is that *some* mental features of a system with non-marginal mental properties are not altered by changes to this transducer layer. The qualification "non-marginal" excludes organisms like worms whose control system is so simple that a change in their transducer layer alters a large proportion of their total functional properties.[8]

I now introduce a more contentious premise. The parts of a transducer layer can be seen as input-output devices, as noted above. They map one set of physical magnitudes to another. The new premise is that there are no constraints on the character of these mappings, in a *bona fide* transducer layer. An input device might map magnitudes one to one, or many to one. The same applies at the output end. And if the mapping is many to one, the "many" need not be a clustered, natural-looking collection, such as a continuous range of values of some variable. A transducer layer input device is just a device that takes *some* set of physical stimuli and maps them to some magnitude that the rest of the system can use. This operation, again, might be represented with a look-up table. An output device, similarly, takes *some* set of values of an internal variable, and maps them to an output. The boundary between control system and transducer layer is not automatically pushed "outwards" if we find that a transducer device is grouping an apparently dissimilar set of inputs and treating them as equivalent. So a "mere change to transducer layer" can include changes to the formal properties of the mapping, as well as which physical magnitudes are involved.

This claim is certainly suspicious. If a transducer layer input device is mapping many disparate frequencies of EM radiation to a single rate of neural firing, it might be

[8] We should probably also stipulate, as Susanna Rinard pointed out, that when a transducer layer is changed, the general *kind* of interface it has with the control system is preserved. Some transducer layers may interface lethally with some control systems.

argued that this must be such a complex device that it cannot be regarded as a "mere" transducer, hence something that can be changed while leaving some mental properties of the system intact. But if these assumptions are granted, they can be used to strengthen existing triviality arguments so that they do not merely show that functionalism collapses into behaviorism, but something more troubling.

Consider an actual human agent, A, with non-marginal mental properties. If functionalism is true, this agent has its mental properties in virtue of its functional organization. This functional organization will be labeled S, and I assume it is specified in the form of a CSA. Then we take a complex physical system, B, that has interactions with its environment. It is "complex" in the sense used earlier; at every instant it is in a different maximal physical state. Following Hinckfuss' example, B might be a large bucket of sea water, isolated from its environment except in ways an agent can control. There will be a possible transducer layer (described below) that can be added to B that will give it the input-output profile associated with S, the functional organization that makes A an agent with mental states. Call B with its modified transducer layer B_L. But if B_L would have the input-output properties associated with S, then, by earlier arguments in this paper, it would also be a realization of S in the functional sense, assuming an simple mapping account of realization.

The actual bucket of water, B, does not have this special transducer layer. Its transducer layer is the water/air surface. But if a system has non-marginal mental properties, a mere change to its transducer layer should not alter this fact. Two functionally similar systems that differ only in physical make-up and transducer layer must either both have, or both not have, non-marginal mental properties. So if a the bucket of water lacks only the right transducer layer to be a functional duplicate of A, then it must already have some non-marginal mental properties.

The part of this argument that needs to be outlined in detail is the claim that if B is a complex physical system, there is a possible transducer layer that can be given to it to yield a system with the input-output profile associated with S. The key to showing this is to note (or require) that all B's physical outputs, as well as inner states, are unique. This makes it possible to apply the same techniques used in the previous sections. To show

how the procedure works, I will return to my earlier example, and discuss how to turn a bucket of sea water into a coke machine.

The transducer layer that has to be given to the bucket of water to make it into a coke machine includes an input device and an output device. At the input end, we need the device to accept 5c and 10c coins. This is no problem; they can be dropped into the bucket. We do have to assume a stock of very physically similar 5c and 10c coins, and a uniform method of dropping. Each coin sends the bucket of water into a new unique physical state, and also generates a unique output. Here, the outputs are the effects of ripples in the water on air molecules at the surface. At each moment, the effects of the water surface on these air molecules are unique products of the prior state of the water and the particular impact of a coin. (A distracting feature of the combination of the Hinckfuss example and the coke machine is the possibility of tracking the displacement of the water by each coin. For generality, assume this easy option is not available.)

The coke machine builder would next draw a new kind of contingency tree for the bucket with its added input device. In Figure 5, outputs O^P_i are the unique physical effects of the water surface on the layer of neighboring air molecules.

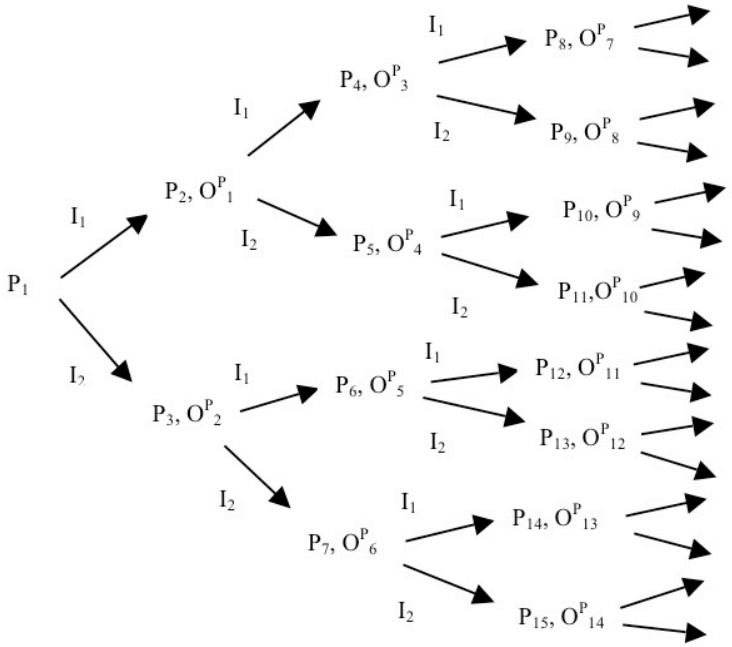

Figure 6: Contingency tree for the bucket of water

All that then has to be done is collect the O^P_i's that should map onto each of the three desired outputs of the coke machine. Those are O_1 (null output), O_2 (emit coke), and O_3 (emit coke and 5c change). So physical outputs O^P_1, O^P_2, and O^P_3 (and some others), should map to O_1, physical outputs O^P_4, O^P_7 (etc.) should map to O_2, and physical outputs O^P_6 and O^P_8 should map to O_3. All the designer has to do to generate coke machine behavior over the interval is build a transducer device that does nothing when it detects O^P_1 (etc.), emits a coke when it detects O^P_4 (etc.), and emits a coke and change in response to O^P_6 (etc.). This, again, is an input-output device with no memory or internal processing. It is as if a designer had enormous knowledge of the physical dispositions of the bucket of water, and very fine-grained ways of building input-output devices, but no way of building the memory needed for a coke machine. So the designer uses the water's complexity as a natural memory. The designer builds a suitable input device, notes the exact physical paths taken by the water surface in response to each sequence of inputs, and the system's physical outputs at appropriate stages are used to control another transducer device which gives out cokes and change. This process can continue for as long an interval as is covered by the designer's physical knowledge.

The coke machine is, of course, a very simple FSA whose realizations do not have mental states. But this process could be applied in principle to any FSA or CSA. If a normal human's functional organization over some interval is represented by a CSA, then our designer could build a transducer device that perturbs the bucket of water in specific ways in response to every possible sequence of inputs that a human might receive, and another transducer device that maps the water's responses to appropriate human behaviors. So a bucket of sea water could act as the control system for a humanoid robot, provided that our designer was extraordinarily knowledgeable about the object's contingency tree and skilled in the building of input and (especially) output transducer devices.

The argument of this section can then be summarized as follows.[9] It is presented as a valid argument with an unacceptable conclusion, so at least one premise must be denied.

1. For any sufficiently complex system B, there is a possible system that differs internally from B only in its transducer layer, and that has the input-output properties of a human agent with non-marginal mental properties.
2. Any sufficiently complex system with the input-output properties of a human agent is a functional duplicate of that agent. (Simple mapping criterion for realization)
3. Functional duplicates share all their mental properties (Functionalism)
4. Two systems that differ only in their transducer layers must either both have, or both lack, non-marginal mental properties.

Therefore:

5. Any sufficiently complex system has non-marginal mental properties.

Even more briefly: if human agent A has mental properties in virtue of his or her functional organization, then any complex system B has some mental properties too. Agent A is functionally identical to an imagined B_L, which is B with a modified transducer layer. But if B_L *would* have mental states, B must *actually* have them, because B and B_L differ only in transducer layers.

It might be thought that one conclusion to be drawn is that the kind of functionalism discussed in these sections leads to panpsychism, a position which might be accepted. But the position implied would not be an "ordinary" form of panpsychism. As complex systems would realize the functional profile of many different intelligent agents, the position implied would be a doctrine of super-position of multiple divergent minds on the substrate provided by each complex physical system.

---

[9] I am indebted to an anonymous referee for suggesting a simplified summary of the argument, which I have adapted here.

5. Consequences for Functionalism

The criteria for realization discussed in this paper are clearly too weak. Even if the near-complete collapse into triviality discussed in section 4 is avoided, by rejecting some premises concerning transducer layers, the earlier results show that changes are needed to make functionalism viable. Otherwise, many standard functionalist thought-experiments involving systems with the behavioral properties associated with mental processing, but inappropriate internal organization, become counter-examples to functionalism too (Block 1981, Braddon-Mitchell and Jackson 1996). So in this section I discuss responses. I emphasize that although the problems may look like a manifestation of familiar difficulties that arise from the introduction of disjunctive predicates, the "fixes" that suggest themselves have consequences for functionalist projects in philosophy and cognitive science. One way to put the point is as follows. The criteria for realization discussed above look weak because of the existential quantifiers; all that is required that a system have *some* physical states that map onto a given structure, or contain *some* states that are related in such a way that they occupy a given set of roles. But this weakness is often something that functionalism *seeks*, because of the message of multiple realizability, and the alleged "autonomy" of high-level descriptions of complex systems.

I first discuss ways of strengthening the criteria for FSA and CSA realization. I then revisit the question of whether matters look different when a Ramsey sentence formulation of functionalism is used.

In the case of FSAs, what seems needed is an extra constraint on the sets of total physical states that are disjoined and mapped to each formal state. The obvious requirement is for some substantive *similarity* between the members of each set. What is needed is not some measure of overall similarity, in a metaphysical sense, but similarity in relation to the causal properties of the system, under lower-level or physical description. Many similarities will be irrelevant, those involving features (eg., color) that have no role in the causal economy of the system.[10] This approach will probably yield a gradient distinction between better and worse realizations, rather than an absolute

[10] I am indebted to Nick Shea for comments substantially improving this part of the argument.

constraint. Once we accept that a system will be in unique states at every point on a contingency tree, it is unlikely that a non-arbitrary absolute standard will be recovered.

Another response that some may have had to the problem in the FSA case is to question not the status of the disjunctions of physical states themselves (the Q_i's), but the conditionals relating them. The simple mapping criterion did not require that the physical states be linked by causal relations, as opposed to dependence relations in a broader sense, but perhaps this is a natural strengthening of the conditions. I see this approach as essentially similar to the one above, but expressed in less promising terms. Familiar ways of talking about causation make it appear there should be a binary distinction between connections that are causal and those that are not, but the situation we are facing, as argued above and below, is one characterized by distinctions of degree. Various degrees of resistance to a causal interpretation of the conditionals will be responses to questionably low levels of similarity across the states collected into coarse-grained categories.

So I cautiously opt to supplement the original account of FSA realization with a distinction between (more) *natural* and (more) *unnatural* realizations, where naturalness derives from similarity in the physical states collected into coarse-grained categories for use in the mapping. The relevant notion of similarity, again, is not an overall or case-independent one, but similarity in relation to a lower-level description of how the device works. The coke machine once again can furnish examples. An extreme example of a machine that is a very unnatural realizer of the coke machine FSA would be one which responds to an initial insertion of 5c or 10c by activating one of two entirely different ensembles of machinery. If the first coin is a 5c, the left half of the machine is activated and the right side shuts down. If the first coin is a 10c, only the right side operates. Then there is nothing in common when "S_3" in the FSA is reached by its two different possible input paths.[11] This contrasts with a case where there are only microphysical differences, invisible to a macroscopic causal description, between the "S_3" state reached through insertion of two 5c coins and the state reached through one 10c coin.

[11] A member of an audience at a conference at Aarhus, 2005, suggested this example.

Here my treatment contrasts with Chalmers (1996), who accepts that an FSA is realized by *any* complex physical system with the right input-output properties. He treats this as a conclusion we seem forced to. Rather than holding to a yes-or-no account of realization with a weak standard, I opt for a gradient distinction between more and less natural realizations (within systems that have the right input-output profiles). This gives us a way to functionally differentiate the two coke machines in the preceding paragraph.

I now turn to the realization of a CSA. Here the key feature of "good" realizations initially seems easier to capture. In good cases, each CSA substate maps to the state of some *part* of the whole system. Substates and C_{11} and C_{12}, for example, should map to two different states of one component of the physical system, a component whose state should be independent of the state of the part of the system that maps to such states as C_{21} and C_{22}. Chalmers (1996) endorses a constraint of this kind. I agree that this is the first step that should be taken, but after fleshing out this idea I will argue that an appeal to lower-level similarity, as in the FSA case, is required as well.

The relevant sense of "independence" is a logical one. In a good realization, specifying some components of a state vector should not logically constrain the state of other components. The Q_{ij}'s used in the CSA triviality argument above do not pass this test. Given that the Q_{ij}'s are disjunctions of P_i's that are exclusive of each other, the instantiation of some combinations of Q_{ij}'s logically implies the instantiation of others. For example, if $(P_1 \text{ v } P_2)$, $(P_1 \text{ v } P_3)$, and $(P_1 \text{ v } P_4)$ are all Q_{ij}'s for a system, mapping to CSA substates that occur in vectors at different values of i, then the fact that $(P_1 \text{ v } P_2)$ and $(P_1 \text{ v } P_3)$ are both instantiated implies that $(P_1 \text{ v } P_4)$ is also instantiated. This contrasts with the case where all the substate variables map to "distinct existences," and the instantiation of a particular substate value has no logical implications about the instantiation of substate values in different regions. Then CSA analysis will, as it is intended to, make it possible to capture the idea of an causal economy of distinct internal factors, as well as interaction of the system with external influences.

In expressing the needed constraint, both an absolute standard and approximations to it are available. The absolute standard is complete independence of the physical states

mapped to each CSA substate. But this standard can be also approximated, as the logical "entanglement" of physical analogues of CSA substates comes in degrees.

This will only solve part of the problem. Even when we have determined that each variable in the state vector maps to a distinct part of the system, it is necessary to constrain which states *of that part* are collected together to be mapped to each C_{ij}. The problem that arose for FSA realization arises again, now at the level of parts rather than wholes. The problem can be illustrated with the simplest possible case. Suppose there is a CSA with two substate variables, each of which has two possible states. So the CSA includes only four possible states. Suppose also there is a physical device with a left half and a right half, which has behavioral properties that match the CSA. States of the first substate variable are mapped to the left hand side of the system, and states of the second are mapped to the right. Constraint on realization is supposed to come from the fact that C_{11}, for example, is a component in two state vectors, $<C_{11}, C_{21}>$ and $<C_{11}, C_{22}>$. These two total states of the system are supposed to have "something in common." The physical state mapped to C_{11} is supposed to play one role when it is part of a $<C_{11}, C_{21}>$ combination and another role when it is part of a $<C_{11}, C_{22}>$ combination. But if *any* physical states of the left hand side of the system can be combined in a disjunction and mapped to a CSA substate, then this combinatorial feature of the formalism exerts no constraint on realization. Any physical states of the left hand side can be regarded as different lower-level ways of being in a single coarse-grained physical state which is mapped to C_{11}.[12]

[12] Chalmers (personal communication) has argued that the combinatorial requirement is stronger than I acknowledge here. Instead, *each* of the physical states mapped to C_{11} have to produce the right behavior when combined with *each* of the physical states mapped to C_{21} and also C_{22}. This is a possible interpretation of the conditionals linking the coarse-grained physical states, but it is too strong an interpretation for functionalist purposes. Here the discussion at the end of section 2 is again relevant. In the case of a system that ages or undergoes other kinds of physical development, this stronger combinatorial requirement would require that the system behave appropriately when one part of it is in a physical state characteristic of early stages in life, and the other parts of the system are in physical states characteristic of late stages in life. This surely is not required for realization of a CSA. Often system when it is older will be realizing a different CSA altogether, of course, but it is surely possible to realize the same CSA while the physical parts of the system develop through time.

The picture emerging from these arguments is as follows. If something like CSAs formalize the kind of functional characterization seen in mainstream functionalism, then the realization of such a structure should be understood to involve not just a mapping between physical and formal states, but two other requirements. One is that substate variables map to independent parts of the realizing system. The other is that the parts' microstates grouped into coarse-grained categories be physically similar. Both of these requirements, but most obviously the second, come in degrees.

Once a similarity requirement is on the table, the question arises whether this could do the work of the independent-parts requirement as well. At least in the example used above, the "entangled" Q_{ij}s will score poorly on a test for lower-level similarity. I do not know whether this will be true of all cases. If it is, the appeal to lower-level similarity may be able to do all the work.

The philosophical consequences of these requirements are significant. Both are at tension with things that many functionalists have liked to say. Functionalists like to stress the "autonomy" of high-level functional description, and this has included rejecting the idea that the realizers of functional states should be distinct, localized parts of the system. It is common to say that there is no need for the realizers of psychological states to be localized to physical regions of the brain. Functional description is carried out at a "higher level of analysis." It is not merely a coarse-grained description of physical machinery, but a special kind of analysis that can posit entities not visible at all at a lower level (Fodor 1974).

The response to CSA triviality arguments above pulls against those ideas. The basic idea of "multiple realizability" of a given functional organization in different sorts of materials is not threatened. And spatial localization of the simplest kind is not necessary for the independence of parts discussed above. A single part of a system might be physically scattered – as the human immune system is scattered through the body. But even a scattered object of this kind is visible to lower-level description. So if an account of realization that requires that CSA substates map to states of distinct physical parts of the system is the best option, this puts pressure on familiar habits of functionalist thinking.

It is not only philosophers who often assume a relaxed standard for realization. The issues here connect to a divide seen in psychology and cognitive science. One style of work in these areas holds that an abstract characterization of psychological processes can be carried out without worrying about finding a simple match between psychological structure and the physico-chemical organization of the brain. Others think that close attention to brain structure is needed to anchor the explanatory posits of cognitive science (Churchland 1989). Some recent philosophical accounts of explanation in "mechanistic" sciences like cell biology and neuroscience have treated the localization of causal factors as an important desideratum (Machamer, Craver, and Darden 2000), though it has not been made very clear what localization of the relevant kind requires. Triviality arguments make these problems urgent ones.

Lastly, I return to the question of how these issues appear within a form of functionalism that uses Ramsey sentences – the "Ramsey-Lewis" approach to functionalism (Lewis 1972, Braddon-Mitchell and Jackson 1996).[13] In the Ramsey-Lewis approach, the notion of mapping does not appear explicitly. A set of hypotheses, given by folk theory or science, is seen as specifying a set of causal roles. Various sets of objects can act as *occupants* of these causal roles. The key relation is not one of mapping between two structures, but the satisfaction of a set of interlocking descriptions by a collection of objects.

I claimed earlier that the change in formalism does not make a difference to the status of triviality arguments. If a Ramsey sentence merely says that there exist a number of inner states of a system, such that when the system receives a given input and is in a given state, it enters some other state, and so on, then triviality problems can be raised as before. What sort of thing can count as the occupant of such a role? The word "occupant," with its real estate connotations, suggests something concrete, but this is not actually required by the framework. If we express our coke machine's workings as a Ramsey sentence, we still have to grapple with the question of why a heterogeneous disjunction of states is not a legitimate occupant of the S_3 role.

[13] Here again I include cases where the theory is folk-theoretic and cases where it is scientific.

This raises a puzzle, however. It can appear that the Ramsey sentence approach does not have triviality problems of the kind discussed in the computationally-oriented literature. (I have often heard this claim made in discussions.) I offer three diagnostic comments. The usual *examples* given to motivate the Ramsey-Lewis approach have special features that defuse triviality problems – in the case of those examples. The special features of these examples are not found in the cases in the philosophy of mind that the Ramsey-Lewis analysis is supposed to handle. However, the Ramsey-Lewis approach can be supplemented with additional constraints, of the kind seen earlier in this section.

(i) The usual examples given to motivate the Ramsey-Lewis approach have special features that defuse triviality problems.

Standard examples used in this tradition, such as Lewis' murder mystery case (1972) and the example of an automobile engine, have features built into them that strongly constrain the entities that can realize the roles. In these examples, causal roles tend to be specified with "thick" causal verbs, as opposed to statements about mere dependence. In addition, many of the functionally characterized entities – not just "peripheral" ones – have direct connections to entities that are not dependent on functional characterization.

In Lewis' 1972 murder mystery case, typical components of the causal roles in the theory given by the detective are "met the victim in Uganda" and "planted the bomb in the attic." In the car engine case, a typical role is "mixes fuel with air." Here functional roles are specified in terms of concrete causal relations that are antecedently understood (*planted* the bomb, *mixes* fuel with air). And in many cases, the functionally characterized entity has direct relations with entities that are not functionally characterized (met the victim *in Uganda*; mixes *fuel* with *air*.) These features impose constraints on realization in a way that is internal to the functional specification, as opposed to being imposed by an additional commentary. Similar results might be obtained by specifying the causal roles in thinner and more abstract terms, but then imposing restrictions on occupants exogenously (as in "all occupants have to be distinct physical objects," etc.).

(ii) The special features of these examples are not applicable in the cases that the Ramsey-Lewis analysis is supposed to help us with.

When the Ramsey-Lewis approach is applied to the mind, we then deal with a system in which these sorts of thickly specified causal relations are typically *not* available for use in functional specification. Instead, we are confronted with a system with a periphery and a rich internal structure. The states to be characterized have few direct links to entities not dependent on functional characterization, and the relations between most states will be treated as abstract dependence relations.

(iii) The Ramsey-Lewis approach can be supplemented with additional constraints of the kind discussed earlier.

We should not expect the Ramsey-Lewis approach to automatically avoid triviality problems, but this approach can, like the mapping approach, be supplemented with extra constraints. Both the constraints discussed earlier in this section may be used; we can exclude gerrymandered collections, and require that some kinds of functional roles be occupied by distinct parts whose states can vary independently.

At least some discussions within the Ramsey-Lewis approach may have always had these extra constraints in mind. Lewis, in particular (e.g., 1994), may have envisaged that the occupants of roles be recognizable as *bona fide* parts of the system, by a criterion independent of the utility of a particular functional description. The versions of functionalism most troubled by the arguments discussed here are those developed specifically to steer a path between the identity theory and behaviorism – seeking to avoid the reductionism and "chauvinism" of the identity theory while retaining the idea that mental states are inner causes of behavior. This is what generates the idea of an abstract "functional state" that can enter into causal explanations while retaining "autonomy" in relation to the system's physical make-up (Fodor 1981). Steering such a path is harder than has been supposed. This threat is that such views will either collapse into behaviorism, or start to draw on the sorts of physical features that were to be kept at arm's length.

In conclusion, triviality arguments are not fatal to functionalism, but avoiding them is not as easy as many functionalists assume. Dealing with the arguments requires a treatment of localization and the relations between levels that is at odds with much 20th century functionalist thinking. Further, if functional description is to be logically stronger than behavioral description, then whether a behaviorally appropriate system realizes a particular functional structure should be seen as a gradient matter.

* * *

References

Block, N. (1978). "Troubles With Functionalism." In C. W. Savage (Ed.), *Perception and Cognition: Issues in the Foundations of Psychology*. Minneapolis: University of Minnesota Press, pp. 261-325.

Block, N. (1981). "Psychologism and Behaviorism." *Philosophical Review* 90: 5-43.

Block, N. and J. A. Fodor (1972). "What Psychological States are Not." *Philosophical Review* 83: 159-181.

Braddon-Mitchell, D. and F. Jackson (1996). *The Philosophy of Mind and Cognition*. Oxford: Blackwell.

Chalmers, D. (1996). "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108: 309-33.

Copeland, J. (1996). "What is Computation?" *Synthese* 108: 335-359

Churchland, P. S. (1989). *Neurophilosophy*. Cambridge MA: MIT Press.

Cleland, C. (2002). "On Effective Procedures." *Minds and Machines* 12: 159-179.

Crane, T. (1995). *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*. London: Routledge.

Fodor, J. A. (1974). "Special Sciences (or the Disunity of Science as a Working Hypothesis)." *Synthese* 28: 97-115.

Fodor, J. A. (1981). *Representations*. Cambridge MA: MIT Press.

Langton, R. and D. Lewis (1998), “Defining ‘Intrinsic.’” *Philosophy and Phenomenological Research* 58: 333-45.

Lewis, D. (1972). "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy* 50, 249-258.

Lewis, D. (1994). "Reduction of Mind." In S. Guttenplan (ed), *A Companion to the Philosophy of Mind*, Oxford, Blackwell., pp. 413–431

Lycan, W. (1981). "Form, Function, and Feel." *Journal of Philosophy* 78: 24-50.

Machamer, P., C. Craver, and L. Darden (2000). "Thinking About Mechanisms." *Philosophy of Science* 67: 1-25.

Piccinini, G. (2004). "Functionalism, Computationalism, and Mental States." *Studies in the History and Philosophy of Science* 35: 811-833.

Putnam, H. (1960). "Minds and Machines." Reprinted in H. Putnam, *Mind, Language, and Reality. Philosophical Papers, Volume 2*. Cambridge University Press, Cambridge, pp. 362-385.

Putnam, H. (1988). *Reality and Representation*. Cambridge MA: MIT Press.

Searle, J. (1990). "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64: 21-37.

Smith, B. (2002). "The Foundations of Computing." In M. Scheutz (ed.), *Computationatism: New Directions*. Cambridge MA: MIT Press, pp. 23-58.

Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge MA: MIT Press.