The Movie Graph Argument Revisited

Russell Standish
School of Mathematics and Statistics
University of New South Wales

August 11, 2014

Abstract

In this paper, we reexamine the *Movie Graph Argument*, which demonstrates a basic incompatibility between computationalism and materialism. We discover that the incompatibility is only manifest in singular classical-like universes. If we accept that we live in a Multiverse, then the incompatibility goes away, but in that case another line of argument shows that with computationalism, fundamental, or primitive materiality has no causal influence on what is observed, which must must be derivable from basic arithmetic properties.

1 Introduction

Computationalism is the idea that our minds are computational processes, and nothing but. In particular, an appropriate program running on a computer will instantiate consciousness just as well as brains made of neurons. Bruno Marchal, who developed the Movie Graph Argument[2] is fond of introducing this concept via a parable:

> You have just discovered you have terminal brain cancer, and the doctor proposes replacing your brain by electronic computer running an artificial intelligence program initialised by the synaptic weights read out from your old brain prior to its destruction.

Would you say “yes” to the doctor? Do you think you will survive the transplant? If not, then what if the doctor proposes replacing your brain with a detailed emulation, including chemical and electrical properties, of all of the atoms making up your brain?

If you say yes at any point, you are affirming computationalism.

However, computationalism implies a number of surprising consequences. Because it is easy to copy a computer program, it should be possible to be cloned into a doppelgänger, whose memories are identical to one’s own up to the point of being cloned. The computer program can be transferred over the internet, allowing teleporting.

Moreover, it is possible to reimplement the exact same program in different ways, for example by coding the program using different programming languages. The detailed use of a machine's registers, and instructions executed will differ dramatically in each case, yet the conscious experience will be identical. Furthermore the same sequence of register states can be activated even though the computer is merely replaying a recording, rather than actively computing something. The Movie Graph Argument seeks to parlay this into an absurdity, where there is no active physical difference between a conscious computation, and the mindless replaying of a recording.

2 Universal Dovetailer Argument

The *universal dovetailer* is a computer program invented by Marchal [3] that effectively executes all possible computer programs, on all possible inputs, albeit with exponential slowdown. It works by executaing the first step of the first program, then the first step of the second program, the second step of the first, second of the second, first of the third, and so on, zig-zagging between executing the next step and starting a new program.

Clearly, if computationalism is valid, then all possible experiences are instantiated by the dovetailer. Each experience will occur with a certain measure within the dovetailer, a measure moreover that is independent of the specific universal machine, or the specific universal dovetailer used (see appendix). What we experience will be drawn randomly from those experiences, hence typically one with high measure in the dovetailer.

One of the consequences of the universal dovetailer argument is that you cannot tell which computer program is you. For every program that instantiates your current conscious state, there are an infinite number of possible continuations of that program, corresponding to different possible futures. This leads to an irreducible indeterminism — even an omniscient god cannot know what you will experience next. The question is ill-posed — which of the possible future "yous" is the real one?

This indeterminism, which Marchal calls *first person indeterminancy* (FPI)[1] is related strongly to how quantum indeterminism appears within the deterministic Many Worlds Interpretation of quantum mechanics.

This indeterminism also implies we are made up of all computations having the same initial history that computes our current conscious state. This will feature in the definition of the *computational supervenience thesis* in §7.

However, the Church-Turing thesis implies that it doesn't matter what physical computer the dovetailer is run on. It could equally be a contraption of gears and cogs, like Babbage's difference engine, or pebbles on a lattice with a child moving them according to the rules of the *Game of Life*, as an electronic computer we know today. Our experienced physical world must therefore be independent of any such such primitive physical substrate, which then serves no explanatory purpose whatsoever. All physical experiences are grounded in the properties of the Universal Dovetailer running on a Universal Machine.

The *Universal Dovetailer Argument*[1] steps 1–7 presents this radical conclusion that physics cannot be ontologically primitive if computationalism is true, in a series of steps that build gradually upon the reader's intuition.

The wrinkle is to suppose that the universe doesn't have sufficient resources to run a universal dovetailer. Whilst a universal dovetailer can be coded and run on the physical computers we have today, in practice only a short initial portion of the UD can be run. What if the universe goes into a heat death before any conscious program is started?

To distinguish between these cases, Marchal calls a universe capable of running a universal dovetailer fully a *robust* universe. Whilst such a universe is necessarily infinite, we can, for the purposes of this argument, consider robust universes to be ones that can run enough of the universal dovetailer for programs instantiating all possible human experiences of consciousnesses within a human life time be executed. This is still an immense universe, but no longer an infinite one.

Since all our possibe experiences will be instantiated, and observed, our phenomenal physics depends only on the properties of the universal machine, not on any underlying physical sustrate.

A non-robust universe is incapable of generating consciousness by running the universal dovetailer. Conscious entities can only appear by a correct program being instantiated at a primitive physical level, whilst other experiences and entities are not so instantiated. The primitive physical world then potentially has a causative effect on phenomenal physics by allowing some experiences to be experienced, and not others.

The Movie Graph Argument (MGA) was developed to show that even in non-robust universes, primitive physics plays no explanatory role. But before presenting the MGA, we must first introduce the notion of *supervenience*, and also discuss a rather similar argument by Tim Maudlin, pointing at the inconsistency of computationalism with materialism.

3 Supervenience

Supervenience is an attempt to capture the dependence of some phenomenon on its substrate[5]. Loosely speaking, a phenomenon *supervenes* on a substrate if a change in the phenomenon ncessarily entails a change in the substrate. For example, consider the phenomenon of speech. Situations where the words "hello" and "hi" will necessarily involve different motions of the air molecules, so we can say that speech supervenes on molecular motion.

In the case of consciousness, it is widely believed that consciousness supervenes on our brains, as it is observed that different brain states invariably correspond to different conscious experiences.

Now consider the scenario of a class of school children, one of whom is named Alice, and another Bob. Does Alice's consciousness supervene on the class? Well, yes, as we observe that any change in Alice's consciousness must correspond to a physical change in the classroom, concentrated in Alice's brain.

But we can ask a slight different question — does consciousness supervene on the class. In this case, we'd have to answer no, because both Alice's conscious states and Bob's, not to mention the teacher's and other students are all present in the class. A difference in conscious state does not correspond to a physical difference. We can express the same conundrum using the speech case, exploiting the so-called "cocktail party" effect. Alice says "hello", and Bob says "hi" simultaneously — but which word we hear depends on who we're actively listening to. The words no longer supervene on the air molecules, but on the state of the listener.

To see how this applies to the universal dovetailer, recall that the universal dovetailer instantiates all possible experiences. A different experience does not entail a difference in the universal dovetailer. So counterintuitively, consciousness cannot supervene on the universal dovetailer, even though it can supervene on a non-dovetailing computation being executed by the dovetailer.

4 Computational Supervenience and Counterfactual Equivalence

The basic idea of the computational supervenience thesis is that a conscious state supervenes on a computation. Of course there are many running programs that perform the same computation. The most trivial example of these being programs that perform the same steps up to some time t, but then diverge after that time, which is the source of the first person indeterminism. But it is also true that two distinct programs will pass through the same sequence of machine states, without being computationally equivalent. The simplest example of such a difference might be if program A executes the "or" instruction on registers x and y, and B executes the "and" instruction.

If it so happens that both x and y both contain the same value (both true or both false), then the resultant machine state is identical with each program. Yet the two programs are quite different, as if the two registers had different values, the resulting machine state would be quite different. We call this "if it had been different" a *counterfactual*. In this case, programs A and B are not counterfactually equivalent.

It seems plausible that counterfactual inequivalence is needed as part of the definition of what could differ between computations supporting different conscious experiences. This is supported by the intuition that a mere playback of a recording of a conscious machine (eg reanimating a dead brain by passing recorded EEG sgnals through the neurons) is not sufficient to instantiate a consciousness. A recording playback (when done perfectly) will pass through the same sequence of machine states as the original conscious computation, yet it need not be conscious because it is not *counterfactually equivalent* to the original computation (we would say it is not *counterfactually correct*).

5 Maudlin's Olympia Argument

Tim Maudlin presented an interesting argument that computationalism is incompatible with materialism[4], which defines as a form of physical supervenience — that consciousness supervenes on physical activity. To summarise his argument, he transforms the physical process performing a conscious computation into one replaying a recording of the process. In a nod to Hoffman's tale *Der Sandmann*, Maudlin calls the former machine Klara and the latter Olympia. The machinery passes through the exact same sequence of states in both cases, but clearly in the second case the computation is utterly trivial — reading the machine state from a recording. The unstated assumption is that Olympia is too simple to be conscious.

It might be objected that Olympia is not counterfactually correct. Klara performs the calculation, and so would produce a different result if some of the intermediate results differed. Olympia, on the other hand, knows ahead of time what the states of the registers are. If the registers were different, it would still produce the same sequence of states as specified in the recording.

To counter this objection, Maudlin introduces a baroque construction of attaching a copy of Klara to each and every state of the sequence. Each Klara has been advanced to the point in calculation corresponding to the step to which it is attached. If the intermediate result differs (not that it will) from that of the recording at some step, the attached Klara will take over the computation from that point, thus preserving counterfactual correctness. Yet, these Klaras are physically inert, as the counterfactual states never occur. Maudlin's point is that if counterfactual correctness is relevant, then a simple switch connecting the physically inert Klaras to Olympia suffices to switch consciousness on and off.

6 Multiverse objection to Maudlin's argument

If the Many Worlds Interpretation of quantum mechanics is literally true, then we must also consider that the counterfactuals will occur in alternate universes. In Maudlin's setup, the program either has fixed inputs, corresponding to a specific history, or has no inputs, perhaps corresponding to a dreaming state. If the situation is one of fixed inputs, then it is easy to see that the quantum multiverse must also contain versions of the same program with differing inputs. If it is the no input situation, then counterfactual states must also occur in any physical implementation due to the possibility of error or noise in the implementation.

So if counterfactual situations are physically realised somewhere in the multiverse, then one can no longer claim that the attached Klaras in Maudlin's thought experiment are physically inert.

Nevertheless, this objection is not a valid objection to the use of the Maudlin's argument, nor the MGA for step 8 of the UDA to obtain the incompatibility of computationalism and materialism for non-robust universes. The reason is that a multiverse is a physical quantum computer, and at least all possible human experiences are experienced somewhere in the Multiverse, if not all possible conscious experiences. Thus the Multiverse, even if finite in size, should be considered a robust universe.

7 The Movie Graph Argument

Marchal originally presented his argument in French as *l'argument du graphe filmé*[2], or the filmed graph argument, which got rendered into English as the Movie Graph Argument. The idea is that the conscious computation is implemented as a graph (or network) of stateful objects (eg abstract neurons) embedded in a glass plate. This allows a movie camera to record a movie of the operation of the artificial brain. Then by parts, he severs some of the network links between neurons, but by projecting the movie back onto the network, is able to excite those neurons as though they were still connected. The result, like Maudlin's argument, is a physically identical process (at the state level) that is computationally not identical. The process is eviscerated still further until we have to conclude that consciousness is supported or not depending on when an external observer chooses to look at the apparatus.

Marchal defines the following supervenience theses:

Computational Supervenience Thesis (CST) That a conscious experience supervenes on all counterfactually correct computations passing through some sequence of machine states.

Physical Supervenience Thesis (PST) That a conscious experience supervenes on a physical process.

The requirement that the computations be counterfactually correct means two distinct conscious experiences may produce the same sequence of machine states, but they must differ in their behaviour for some counterfactual machine state. For example, the program "x and y" and "x or y" will produce the same sequence of states if registers x and y both held the value 1, but differ on the counterfactual case of x=1 and y=0.

In particular, the simple playback of a recording of sequence states (such as Maudlin's Olympia example) will not be counterfactually correct, thus ensuring the CST does fall into an absurdity.

It is quite conceivable that two different programs supporting two different conscious experiences will pass through identical sequences of states, differing only counterfactually. Whilst in a quantum Multiverse, the physical processes will behave differently, due to the different behaviour in the counterfactual branches of the Multiverse, in the non-robust universe, we are considering only a single universe. These two distinct experiences correspond to the exact same physical process, contradicting the PST.

8 Conclusion

Physical supervenience is simply not compatible with computational supervenience in a non-robust universe. In order for physical supervenience to be compatible with computational supervenience, we need to inhabit a Multiverse, which as noted above is a robust case.

The anthropic principle is the notion that our physical environment is compatible with our physical existence within that environment. This is a form of physical supervenience, and observed evidence consistently points to the anthropic principle as being true.

If we wish to assert we don't live in a Multiverse, then we need to abandon the computational supervenience thesis. But that would entail that a copy of a conscious program need not be conscious, contradicting the so called "Yes, doctor" axiom of computationalism. So we would also need to abandon computationalism.

However, if we embrace computationalism in a robust Multiversal reality, then we are led to the conclusion that the exact form and structure of any primitively executing computer can have no explanatory or causative role on empirical physics. Thus we are led to the *reversal*, physics is entirely determined by the properties of universal computation, which in turn is completely determined by any sufficiently rich system, such as integer arithmetic.

References

[1] Bruno Marchal. The origin of physical laws and sensations. Invited talk at SANE'2004, http://iridia.ulb.ac.be/˜marchal.

[2] Bruno Marchal. Informatique théorique et philosophie de l'esprit. In *Actes de 3ème colloque internatial de l'ARC*, pages 193–227, Toulouse, 1988.

[3] Bruno Marchal. Mechanism and personal identity. In M. de Glas and D. Gabbay, editors, *Proceedings of WOCFAI '91*, pages 335–345, Paris, 1991. Angkor.

[4] Tim Maudlin. Computation and consciousness. *J. Philosophy*, 86:407–432, 1989.

[5] Brian McLaughlin and Karen Bennett. Supervenience. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford, spring 2014 edition, 2014. http://plato.stanford.edu/archives/spr2014/entries/supervenience/.

A Measure over the universal dovetailer

Fix a universal machine U, and let $\mathcal{P}$ be the set of all programs of U, with a fixed enumeration p_i.

First we need to define the equivalence class of programs that perform an identical computation for their first k steps.

Definition 1 *$\alpha_{jk} \subset \mathcal{P}$ is a partitioning on $\mathcal{P}$, ie $\bigcup_j \alpha_{jk} = \mathcal{P}$ where $p \in \alpha_{jk}$ iff $\forall x \in \alpha_{jk}$, the first k steps of p are counterfactually equivalent to x*

Since it is a partitioning, we are interested in the probability measure $\mu(\alpha_{jk})$, where $\sum_j \mu(\alpha_{jk}) = 1$, which would correspond, under COMP, to a measure over observer moments.

Definition 2 *Let $u_i(\alpha_{jk})$ be the probability measure that program $p_i \in \mathcal{P}$ executes a program in α_{jk}, with $\sum_j u_i(\alpha_{jk}) = 1$.*

Remark 1 *For most programs $p_i, u_i(\alpha_{jk}) = \delta_{il}, \exists p_l \in \alpha_{jk}$, but for dovetailers and other interpreters, the distribution differs from the Kronecker delta. In particular, for a universal dovetailer p_i, $u_i(\alpha_{jk}) > 0, \forall j$.*

We have

$$\mu(\alpha_{jk}) = \sum_i \mu(\alpha_{ik}) \sum_{p_\ell \in \alpha_{ik}} u_\ell(\alpha_{jk}) \tag{1}$$

Writing this as a vector/matrix form with $\mu_i = [\mu(\alpha_{ik}]$ and $\mathbf{U}_{ij} = [\sum_{p_\ell \in \alpha_{ik}} u_\ell(\alpha_j k)]$, eq (1) can be written

$$\mu = \mathbf{U}\mu \tag{2}$$

The measure over initial sequences of programs is therefore an *eigenvector* of $\mathbf{U}$ corresponding to the eignevalue 1. We still need to establish that $\mathbf{U}$ has an eigenvalue of 1, and that the null space of $\mathbf{U} - 1$ is 1 dimensional, guaranteeing a unique measure μ, however it is clear that μ is independent of which program $p_i \in \mathcal{P}$ is primitively running on U. Hence we may substitute a different reference machine V, and run an interpreter of U executing a universal dovetailer, showing that the measure μ is also independent of the reference machine.