manuscript version 7.3.25

A Deductive Argument for Strong Emergence Based on The Halting Problem

Maxwell Murialdo†, Esteban Céspedes*, Arturo Cifuentes‡

Abstract
A recent result from theoretical computer science provides for the verification of answers to the Halting Problem, even when there is no plausible means by which to derive those answers using a bottom-up approach. We argue that this result has profound implications for the existence of strongly emergent phenomena. In this work we develop a computer science-based framework for thinking about strong emergence and in doing so demonstrate the plausibility of strongly emergent phenomena existing in our universe. We identify six sufficient criteria for strong emergence and detail the actuality of five of the six criteria. Finally, we argue for the plausibility of the sixth criterion by analogy and a case study of Boltzmann brains (with additional case studies provided in the appendices.)

keywords: emergence, reductionism, computation, physics

† Lawrence Livermore National Laboratory, Livermore, CA, USA.
ORCID: 0000-0003-3611-6796

* Corresponding author. ecespedes@ucm.cl
a) Philosophy Department, Catholic University of the Maule, Talca, Chile
b) Valparaíso Complex Systems Institute, Valparaíso, Chile
c) Center for Studies in Philosophy of Science, Logic and Epistemology, Valparaíso University, Chile
ORCID: 0000-0002-5329-5434

‡ Clapes UC, Catholic University of Chile, Santiago, Chile
ORCID: 0000-0001-9689-3939

1. Introduction

Emergent properties are normally understood as properties of a system that arise from changes in the complexity of the system's constituents and are in some relevant sense irreducible to those constituents (cf. El-Hani and Pereira, 2000; Kim, 2006). One of the first classical formulations of the notion of emergence can be

found in the work of John Stuart Mill (1843), who offered a conceptual analysis for effects of joint causes that are not described as the algebraic sum of those causes. Based on Mill's analysis, George Henry Lewes (1875) distinguished between resultant and emergent effects, in what apparently involved the first use of the concept of emergence in a technical philosophical sense. A further development was made later by the so-called British emergentists Samuel Alexander (1920), Conwy Lloyd Morgan (1923), and Charles Dunbar Broad (1925), who deepened the notion of emergence, focusing on irreducibility and on its relevance to understand mental and biological processes. Given the remarkable progress of areas such as particle physics and molecular biology during the 20th century, among other reasons, the reductionist agenda gained force and emergentism began to be conceived as a less serious philosophical point of view. But later, with the development of complexity science, involving fields such as chaos theory and network theory, the concept of the emergent property not only resurfaced as useful, but also as necessary (cf. Holland, 2014). Some of the paradigmatic features of an emergent property, conceived in various approaches, are its relational character with regard to the basic parts of a system, its novelty, its irreducibility, and its holisticness (cf. Kim, 2006; Humphreys, 2016). But where it succeeds, reductionism has neither lost methodological nor explanatory force. Now, is it possible to reconcile the concept of emergence with the empirical success of reductive explanations?

A distinction that allows one to explore ways in which the notion of emergence could be compatible with reduction is the distinction between weak and strong emergence. According to the account developed by Mark Bedau (1997), a state of a system is weakly emergent with regard to the system's constituents, if that state can be predicted from the constituents, but only by simulation, i.e., only by reproducing all intermediate steps of the inferential chain between their representations. In a computational sense, weak emergence implies predictability but also incompressibility (cf. Humphreys, 2016). By contrast, strongly emergent phenomena are not deducible at all from the constituents of the system. While Bedau has argued that cases of strong emergence seem mysterious and hard to characterize scientifically, David Chalmers (2006) claims that consciousness is a clear instance of it that can be understood in relation to the physical domain. Looking for a naturalistic account of strong emergence, Rani Lill Anjum and Stephen Mumford (2017) have proposed to characterize emergent properties in terms of causal powers, arguing that strong emergence implies causal transformations not only at the high-level, but also changes in the causal powers of the system's constituents. Other recent views consist in proposing algorithms for characterizing strong emergence that are consistent with the observable behavior and evolution of biological systems (Hao et al., 2021).

In this work we analyze strong emergence from a digital physics perspective. Digital physics is the notion that the universe is a vast digital computer (or at least can be fully modeled as one). Many distinct variants of this notion have been proposed since Konrad Zuse (1969) introduced the idea in his 1969 book *Rechnender Raum*, and Edward Fredkin coined the term "digital physics" (Fredkin, 2003). Whitworth (2008) argued that some discoveries in modern physics such as quantum minima, quantum equivalence, the big bang, and the maximal speed of light for the

transmission of causal information, etc. point towards digital physics conclusions. More specifically, Seth Lloyd (2007) has argued that the universe itself is a giant quantum computer whose computed output is reality in real time. From this perspective, fundamental particles are simply bits (or qubits) of information. Any collision or interaction between fundamental particles is therefore considered data processing or computation (e.g., a bit flip) acting on the information. Stephen Wolfram has also contributed widely to this domain (Wolfram, 1985), and digital physics remains an area of active research (Beraldo-de-Araújo and Baravelle, 2017). Along a different vein, Nick Bostrom (2003) has argued for a version of digital physics known as the simulation hypothesis. This hypothesis posits that as technology advances, future generations will have a vast amount of compute power at their disposal, enough to render a great number of full universe simulations. Each full universe simulation might, for example simulate a past state of the original universe or a counterfactual state of the original universe. Furthermore, Bostrom hypothesizes that each simulated universe could be home to many conscious, simulated human minds who are unaware that they are living within a simulation. From this, a popular (though controversial) argument proceeds: if the total number of minds of the simulated variety significantly outnumbers the quantity of non-simulated minds, then by the principle of indifference any individual should expect (with high probability) that they are actually a simulated mind living in a simulation.

For the purposes of this work, we assume a digital physics backdrop with the following general (non-specific to a particular notion of digital physics) characteristics:

1) We assume that the entire universe either is a deterministic Turing machine or is a simulation running on a deterministic Turing machine.

2) We assume that the universe can be treated as a computationally bounded system. This is to say that the total number of computational steps that can be carried out over the lifespan of the universe is a finite number.[1]

3) We assume that the system is informationally closed (to non-random information). No non-random information that was not present at the inception of the universe is allowed to enter into the system from an external source at a later time. Information can only propagate by transforming prior information contained within the system (since the inception of the system) or from random sources. For example, if the universe is a simulation running on a computer, no one is allowed to be typing at the keyboard attached to the computer after the inception of the universe. (We briefly suspend this assumption to discuss a set of one or more alien agents we

1 In the case that the universe itself is a giant computer, see Comment 5 in Section 3. In the case that the universe is a simulation running on an external computer, we bound the system by segregating out the universe (simulation) and the computer it is running on as a single system. We can then draw an arbitrary box around the simulation and the hardware it is running on and call this a computationally bounded system.

denote as "Bob" in Section 4, but this is only used to illustrate a point and the input of information from a source external to the system is not key to the argument. The role of Bob is then pulled internal to the system in Section 6 to ensure that this assumption holds.)

4) We assume that there is no source of infinite computational resources that could be accessed.

The digital physics backdrop that we assume is tantamount to a specific form of physicalism (monist perspective) wherein the notion of *physical* is defined as something constituted of digital information, implemented and operating within a physical substrate.[2] Under this backdrop, every future state of the system can (and must) be fully computed from a prior state of the system (in addition to any random information that is injected into the system). More specifically, the state of the system at any specific timestep, N_t can be computed from a prior state N_{t-1} plus any injected random information, R_I. (This assumes certain rules of information transformation particular to the system, which we might loosely call the fundamental laws of physics.) Furthermore, the very act of computing the state of the next timestep of the system on the hardware of the system causes the system to enter into that state. Under this digital physics backdrop, the system can and must compute each future state in full using only its available computational resources. The system is informationally closed (to non-random information).

Under general physicalism principles, this system would likely be presumed to be causally deterministic and closed to outside/additional influences (once information about the prior state, N_{t-1}, and random information, R_I, are considered). Each prior state (along with random information) defines the next state in a self-propagating fashion. In this work, however, we describe counterexample cases wherein the informationally closed system may not be able to fully compute its next state (given limited internal computational resources). In these counterexample cases, the input information and rules of information transformation are known, but the computational power necessary to perform the information transformation is lacking. Therefore, in order for the system to compute and propagate itself into the state of its next timestep, the system would effectively have to reach outside of itself and borrow computational resources from elsewhere. However, in Section 4 and Appendix A we describe how the quantity of computational resources that would have to be borrowed to solve specific problems can effectively be made arbitrarily large (while still finite). Therefore, for any finitely sized computational resource that the system might borrow from, we can point to the fact that for appropriately selected problems, the resource will still (and always) be too small. Moreover, by assumption 4 above, it is unreasonable for the system to borrow computational resources from an infinite computational resource. This in turn creates a conundrum—how can the system compute and propagate itself into its next state (in a bottom-up

2 Note that the arguments in this work can in principle be extended to apply to more general physicalist, monist perspectives. However, we leave this extension as future work.

fashion)? This conundrum points towards the plausibility of strong emergence and information that is discoverable and verifiable (and defines a future state of the system) even if it is not derivable from the bottom up within the system. Here we define a bottom-up approach as a deductive sequence that starts from simple objects or axioms (that cannot be decomposed into subparts) and builds towards assemblies of parts.

In order to flesh out the foundations of this conundrum, we start by describing the existence of specific instantiations of the Halting Problem that would require computational resources greater than the system's bounded computational capacity in order to solve for the correct answer. We demonstrate that these types of problems are likely to exist and that their solutions are verifiable (though not derivable from the bottom up) with minimal computational resources.

Furthermore, we argue that if these solutions can be discovered and verified, then they can have causal consequences for the whole system. For example, a robot could be programmed to take in input information regarding the solution to one such specific halting problem, verify the solution, and if verified, physically move in a way that is programmed to be different than if the solution were incorrect and could not be verified. Under physicalism, we would expect the robot's actions to be fully determined solely by the information embodied in the robot's hardware and any software inputs (along with the rules of fundamental physics). However, in our counterexample cases, the robot's actions would additionally be codetermined by objective information that is discoverable and verifiable within the system but not derivable from the bottom up within the computationally bounded system. We will call this third input to the robot's behavior (e.g., the answer to the specific instantiation of the Halting Problem in question) "ethereal information".

Therefore, the state of the overall system at the next timestep is determined by the robot's hardware information, software information, *and* by ethereal information. In this counterexample case the overall system would be underdetermined without considering the ethereal information input, and therefore the next state of the system could not be accurately determined. The existence of ethereal information (of this type) would be indicative of strong emergence in a digital physics paradigm.

Historically, strong emergence has been dismissed as mysterious and hard to characterize scientifically (Bedau, 1997). In this work we provide a conceptual framework for understanding strong emergence against the backdrop of digital physics. We do not attempt to prove or disprove the existence of strong emergence, but rather present deductive arguments and case studies that highlight and clarify its plausible existence.

In this work we delineate strong emergence in a new manner that is slightly different from some other historical notions of strong emergence. Whereas in the past some have loosely defined strong emergence as *that which could not be predicted by Laplace's Demon*, an entity with infinite capacities (cf. Davies, 2004; Gompert et al., 2022; Collier, 2011), in this work we are interested in *that which could not be predicted given an arbitrarily large, though finite, computational capacity*. We contend that our version of strong emergence is strong enough to be philosophically

interesting. This is especially true within a digital physics backdrop. While this version of strong emergence might only include future states that are logically entailed from prior states, random information, and the fundamental rules of information transformation, these future states cannot be predicted from within the computationally bounded system. Therefore, a conundrum arises as to how one state can cause a subsequent state without the computational power to predict that state.

In order to cement this framework, we detail a set of six criteria that are sufficient for the existence of strong emergence and a deductive argument composed from these criteria. (We do not argue that these criteria are necessary, only that they are sufficient.) We go on to describe the existence of at least five of these six criteria and argue for the plausibility of the final criterion by case studies. If the six criteria do in fact exist within our universe, then they either indicate the existence of strong emergence or the implausibility of the digital physics paradigm that we have described.

2. Six proposed criteria that together imply strong emergence

First, we will outline the criteria that we believe imply strong emergence. We argue that anything that meets the following six criteria should be considered strongly emergent (i.e., the criteria are sufficient, but not necessary, see Section 5). Specifically, these criteria can be met by a hypothetical question-and-answer pairing (to be discussed further). The question-and-answer format was selected as it lends itself to rigorous analysis by considering the mathematics of a closed system. Herein we define a closed system as one that is computationally bounded (i.e., constraints impose a finite limit on the number of computational steps that can be taken within the system) and does not allow for the introduction of non-random information from an external source.[3] The six criteria that we believe to be sufficient for a question-and-answer pairing to exhibit strong emergence are:

Criterion 1. The question-and-answer pairing can be formulated and posed in a way that is compatible with and containable within the constraints of the closed system.

Criterion 2. The answer to the question has a non-random, objective (observer independent) truth value that corresponds to an object or construct within the closed system or to the system itself.[4]

3 Of course, we don't require the system to be absolutely closed, but that there must be a set of conditions, properties, or regularities (possibly laws) delimiting the information that is accessible within the system.

4 For our purposes, "truth" or "truth value" simply indicates that the information is non-random and can be verified in an objective sense (e.g., a fact, a law, a pattern, a regularity, etc.). For the purposes of this work, additional warrant indicating that the answer is true in

Criterion 3. The answer to the question cannot be derived using a bottom-up approach from within the closed system under consideration (although it might be derivable in an alternate system).[5]

Criterion 4. The evidence necessary to argue for the truth value of the answer (increase its a posteriori credence, conditioned on an event) is compatible with and containable within the constraints of the closed system. Therefore, the answer is hypothetically *knowable* within the system.

Criterion 5. A gain in knowledge regarding the truth value of the answer can be induced to have physical, causal consequences within the system.

Criterion 6. The answer and its supporting evidence are *discoverable* within the closed system. (Defined in more detail in Section 5.)

We will explain each criterion and its importance in Sections 4 and 5.

3. Background and assumptions

In this section we comment on various assumptions and background information used in this work.

Comment 1: Basic assumptions

In this work we assume the validity of logic and mathematics. Without these starting assumptions, deductive and mathematical arguments of this sort could not proceed. Furthermore, for simplicity we assume the validity of modern physics (e.g., the standard model and general relativity). For our argument, it is not necessary that all of modern physics be complete or fully correct, but rather that specific points (as enumerated below) be correct so that we are able to make use of the results from "MIP* = RE" (Ji et al., 2021). "MIP* = RE" is the title of a 2021 paper published in *Communications of the ACM*. This paper proves that solutions to the Halting

some deeper sense (beyond, for example, being a non-random, objectively verifiable regularity) is not necessary (as our arguments do not rely on such).

5 Here we use the term bottom-up derivation to denote the process of acting upon input information (present at the inception of the closed system) in ways allowed by the system's fundamental rules of information transformation in order to generate new informational states. This process can then be iteratively applied to a new set of input information that additionally includes the new informational states from the prior steps.

Problem can be verified even if they cannot be derived in a bottom-up fashion (given finite computational resources).

For portions of this work relying on "MIP* = RE," we assume that

1) finite-dimensional quantum mechanics is a good model of physics,
2) the provers (Bob) cannot communicate directly amongst themselves, and
3) the verifier (Alice) has access to truly random coin flips which are unknown to the provers.

Comment 2: Gaining knowledge

In this work we use the notion of "gaining knowledge" in the following context. Per our use of information theory, "gaining knowledge" about the truth value of an answer to a question is not simply a matter of being able to write down or enumerate *all possible* answers in a list. "Gaining knowledge" comes from being able to discern (at least in part) the cases that possess truth value from the cases that are pure gibberish. This requires a process of verification.[6] Per Comment 1, we have assumed the validity of mathematics which provides one logical framework for the verification process. An alternate verification process can involve direct experiential verification by acquaintance. For the purposes of this paper, we will assume a factual notion of truth. To gain knowledge is to approach that truth in a probabilistic and verifiable fashion[7].

As an example, consider the somewhat misleading claim that "I know John's six-digit ATM pin number" because I have in my back pocket a list of all possible six-digit numbers. Surely John's pin number is on that list somewhere. However, I do not know where it is on the list. From an information theory standpoint, this is an equivalent way of saying that in reality I know *no* explicit information about John's ATM pin number (beyond the obvious a priori constraints). If I were to learn the first three digits of John's six-digit ATM pin number then I will have gained knowledge about John's ATM pin number, even without learning all six digits.

In order to truly gain knowledge (or information), I must be able to separate or discern (at least in part) the needles (true answers) from the haystack (set of all possible answers). To gain knowledge (in our usage) is to increase the a posteriori credence that a particular answer (from the entire set of possible answers) is the true answer, conditioned on some event. This is in line with the Shannon entropy notion of gaining information.

---

6 The verification process can be used to ensure that the information is non-random.

7 For the present purposes, we prefer the idea of "gaining knowledge" instead of the one of "knowledge state". The latter is stronger than what we need and implies constraints regarding justification that we do not have to discuss here. Furthermore, we may assume that the factual character of the truths that are approachable in the process of gaining knowledge depends on how they are described and is therefore not independent from epistemic conditions.

Comment 3: The Halting Problem

In this work we repeatedly refer to the Halting Problem. The Halting Problem is a famous question in computer science that was one of the first decision problems that was proven to be unsolvable (Turing, 1936). In short, the question asks whether a specific computer program that is run on a specific deterministic Turing machine (computer) will halt after a finite number of steps (halt) or will continue running forever (not halt). One of the two options (either halt or not halt) must be correct. However, determining which option is correct is not simple (and sometimes not possible). One might naively try to write an analysis algorithm that will take in the specifics of the program and the specifics of the computer and within a bounded finite number of analysis steps always correctly determine whether the program will halt or not halt. Alan Turing proved that no such algorithm can exist that will work for all possible programs.

Furthermore, throughout this work we refer to "specific", "specifically posed", or "specific instantiations" of the Halting Problem. By this we indicate an appropriate pairing of a particular input program, *x*, with a particular computer, *M* that conforms to the criterion laid out in Appendix A. In short, this pairing should be selected so that the question as to whether this combined system (*x* run on *M*) halts or does not halt cannot be answered (from the bottom up, within the closed system of our idealized universe). Additional details are provided in Appendix A.

Comment 4: Examples of closed systems

In this work we will go through and explain each of the six sufficient criteria and its importance. In this process we will consider two distinct examples of closed systems in which each criterion might be applied: Possible World One (w_1), and Possible World Two (w_2). These examples are included for explanatory purposes. w_1 is an idealized version of our universe (possibly surrounded by other distinct parallel universes that are themselves closed systems). This idealized version of our universe is used to illustrate concepts and not to make pronouncements on the true physics of the universe we inhabit. In w_1, the total number of computational steps that can be taken is finite due to fundamental physics bounds, and this idealized universe comports with modern physics.

w_2 is an idealized simulated universe contained within a computer simulation on a standard laptop. This simulated universe contains conscious simulated minds and has a finite computational resource based on the capacity of the laptop. w_2 also comports with modern physics. Both w_1 and w_2 examples are selected with the intent to provide analogies that help to cement understanding and intuition regarding the six sufficient criteria listed above. Note that we cannot rule out the possibility that the real world which we inhabit is identical to w_1 or w_2.

Comment 5: Reasons to believe that our universe might be computationally bounded (a finite, closed system)

In this work, for our arguments, we assume that our universe is computationally bounded (a closed system). We assume that the number of computational steps that can be carried out within our closed system (idealized universe) is finite. While this is a reasonable assumption to make (philosophically speaking), its veracity in our particular universe is a question for physics, not philosophy, that is beyond the scope of this work. Nonetheless, a number of discoveries in modern physics suggest the physics-derived reasonableness of this assumption. Fundamental, physics-based boundaries on how many computational steps can be taken in our universe can be approximated by combining known physical constants and laws. For example, the speed of light puts a physical constraint on the maximum volume of space over which causal interactions can occur (within a given time). It is hypothesized that this volume is in turn quantized and cannot be subdivided smaller than a Planck's length (Planck, 1899). Together these constants set a maximum number of bits that can exist in the universe (using the Beckenstein bound (Beckenstein, 1981)). Bremermann's limit in turn sets a maximum rate of computation for a system with finite mass-energy (Bremermann, 1962). Finally, the predicted heat death of the universe sets a time limit for this maximally grand computation. (Alternatively, if the universe were to end in a "Big Crunch", this too would set a time limit for computation.) The exact maximum number of computations that can occur within our universe doesn't matter for the arguments herein. What does matter is that for our idealized universe the figure is finite, and that for every finite number there exists an infinite quantity of larger numbers. Accordingly, for any finitely sized system, there exist specifically posed programs for which we cannot determine, using a bottom-up approach, whether or not the program will halt (see Appendix A). We might not be able to specify which programs fall into this special category, but by trying a broad swath of differently seeded programs we are likely to run across some that do fall into this special category. We leave any attempts to calculate or constrain the likelihood of encountering such programs for future work.

4. Criteria 1-5 explained

In this section we dive into the specifics of criteria 1-5.

4.1 Criterion 1

The first criterion is that the format of the question and its possible answers must be compatible with and containable within the closed system. Informational formats that cannot exist within the closed system might exist elsewhere, but are likely neither approachable, accessible, nor interesting to anyone within the closed system. If

we take w_1 as an example of a closed system, an example of a forbidden format would be physical instantiations of higher dimensional objects that cannot exist in the limited dimensions of this universe. For example, higher dimensional beings in some other universe might have caught a 37-dimensional butterfly. However, there is no way for this high-dimensional butterfly to exist within the confines of our three-spatial-dimensional idealized universe (even if the beings had a method of transferring objects between universes). Alternatively, if we take w_2 as an example of a closed system, then there could exist prohibited formats of information, such as special characters that are not recognizable within the syntax of w_2, or analog information that cannot truly be digitized. Luckily, plenty of supported information formats *do* exist within each closed system. These formats allow us to ask interesting questions like "is there a largest prime number?" This interesting question both makes sense and is easy to pose in the allowed language formats of either closed system (w_1 or w_2). The possible answers "yes" or "no" also can be readily enumerated in either closed system. This is therefore a valid question for these closed systems.

4.2 Criterion 2

The second criterion demands that the question have a non-random, objective truth value in the context of the closed system. This mandates that the question makes sense and has at least one correct answer (and by proxy at least one incorrect answer) in an objectively verifiable (observer independent) sense. Questions like, "Is the smallest prime number a good number?" or "What is the density of the smallest prime number?" would not satisfy this condition. These questions are ill-posed under mathematics, so they do not have objectively true answers. However, the question "is there a largest prime number?" does have an objectively true answer— "no" (there are infinite prime numbers and therefore no largest prime number exists). Note that here (once again) we are only concerned with pragmatic definitions of "objective truth" that suffice for the arguments at hand, without delving into deeper metaphysical questions regarding a general definition of "what is truth" or "what is objective". For the purposes of this work, we define an answer as having "objective truth value" if it can be verified in a manner that is compatible with and logically follows from the fundamental rules of the system, and the verification will yield the same result regardless of which competent entity performs the verification.[8] Note that the answer must be non-random and enable some mechanism and basis by which verification can be performed. For example, the observable values of purely random quantum noise would not suffice for this criterion.

8 For the purposes of this work, we assume that the fundamental rules of the system (e.g., mathematics, logic, physical theories) are universally known and agreed upon.

4.3 Criterion 3

Criterion 3 requires that the aforementioned truth value cannot be derived using a bottom-up approach. To illustrate, the fact that there are infinite prime numbers was proved (derived using a bottom-up approach) circa 300 BC by Euclid in ancient Greece. This was achieved by starting with basic axioms of mathematics and processing on them collectively while following the laws of mathematics to generate simple theorems. These simple theorems could then be further processed on collectively (again following the laws of mathematics) to generate more complex theorems, and so on, until the desired answer is reached. This process illustrates a bottom-up derivation approach wherein each stratum of theorems follows from and builds on the strata that lie beneath.

Is it conceivable for a question to have an objectively true answer (within or about a closed system) when that answer cannot be derived using a bottom-up approach (within the closed system)? Gödel's incompleteness theorems prove that it is impossible to prove (or derive from the bottom up in a finite number of steps) all truths about a closed system, from within that closed system. This surprising result from 1931 proved that there are true statements within a formal system that cannot be derived from the bottom up within that formal system (Gödel, 1931).[9]

How can a true statement that is within and compatible with the framework of a formal system not be derivable from the bottom up within that formal system? One possible explanation lies in the clash between the finite and the infinite. When a true statement (or theorem) is derived from the bottom up, this process starts with a set of assumed fundamental axioms. Then, step by step, the logician combines and rearranges the fundamental axioms and their direct products, building up a ladder to the desired result (proved theorem), one rung at a time. Gödel proved that for some true results this process requires building a ladder with an infinite number of rungs. In a system of finite resources, achieving such a task that requires an infinite number of discrete steps is impossible. (For the purposes of this work we consider the clash between a system with finite computational resources and a specific halting problem task that can be induced to require any arbitrarily large (though finite) number of computational resources. This task, while technically finite, can therefore be placed an arbitrarily large distance out of reach for a specific system of finite computational resources. Accordingly, the impossibility of accomplishing the task in a bottom-up fashion from within the finite system mirrors the aforementioned clash between the finite and the infinite.)

While there might exist myriad forms of ontological gaps (uncrossable chasms encountered when attempting to derive something from the bottom up), the clash between the finite and the infinite provides an easily intelligible form. How can a gap become (ontologically) impenetrable when the process of going from each rung of the ladder to the next higher rung of the ladder is well established? One answer is to create a region on the ladder that has infinite rungs. Imagine, for example, a

9 Not all truths about, within, or pertaining to a system are truths that are necessarily accessible from within the system.

ladder that is 90 feet tall and split into 3 regions. In the lower thirty feet of the ladder there are 30 rungs to climb. In the upper 30 feet of the ladder there are also 30 rungs to climb. But in the middle 30 feet there are an infinite number of rungs to climb. No entity constrained by finite resources could climb all the way from the bottom to the top of the ladder (rung by rung). Nonetheless, a finitely constrained entity *could* make considerable progress by starting from the bottom (getting as far as 30 feet up) and then naively extrapolate that if they have made it this far, there must be no reason why they can't make it all the way up. This is a particular danger of the reductionist agenda (overextrapolation to untested regimes).

Shortly after Gödel published his famous theorems proving that (within specific formal systems) mathematical truths exist that cannot be derived from the bottom up in a finite number of steps, Alan Turing proved that in computer science, too, there exist problems whose answers cannot be derived from the bottom up in a finite number of steps within the system. Turing proved that no general algorithm can correctly solve the Halting Problem in all cases, using a bottom-up approach and a specified finite number of steps (Turing, 1936). The Halting Problem is a decision problem which seeks to find out whether programs fed in as inputs, *x*, to a given computer, *M*, will ever finish their computations, or whether they will simply keep computing forever (lost on an infinite wild goose chase). Using a clever and indirect tactic, Alan Turing proved that in general there is no bottom-up shortcut to know which case it will be: will a program halt, or not halt? The only effective general method to determine whether a program will halt is to start running the program and wait it out. The program might run for a second and then halt, or a million years and then halt. It might run for a billion years and then halt, or it might even run for an infinitely long period of time and never halt.[10] In his proof that the Halting Problem is generally undecidable, Turing showed that there is no general shortcut—we simply cannot rapidly know in general whether a program will halt or not, or how long the process will take if it does halt. This was one of the first yes/no decision problems that was proven to be impossible to solve.[11]

It is also important to note that the Halting Problem is typically generalized for a Turing machine with infinite memory. This, of course, is impossible to build within a finite system. However, for our purposes either a hypothetical Turing machine with infinite memory *or* a physical Turing machine with a large, but finite memory will suffice. This is because the number of physical states that a Turing machine can explore without repeating itself is 2^(number of memory bits). In this way even a relatively small Turing machine can be made to have more physically realizable states than the estimated number of computations that could be achieved in our idealized universe. It is therefore relatively straightforward to produce a real Turing machine for which a specific instantiation of the Halting Problem could not be solved from the bottom up within our idealized universe (see Appendix A). This is to say that for some specific program inputs, we could never determine (in a

10 Assuming a fixed computation rate

11 More details on the specific types of input/computer pairings that we consider of interest in this work can be found in Appendix A.

bottom-up fashion from within the confines of our computationally bounded universe) whether the program will halt or not halt.

4.4 Criterion 4

The fourth criterion is that the evidence necessary to argue for the truth value of the answer (increase its a posteriori credence, conditioned on an event) is compatible with and containable within the constraints of the closed system. Therefore, the answer is hypothetically *knowable* within the system. To elaborate, a truth that cannot be derived from the bottom up is one thing, but if it can't be known even in principle (and perhaps has no causal force) within our universe, does it matter? If, for example 37-dimensional butterflies *do exist* in some other universe but cannot interact with our universe in any way and even knowledge of their existence cannot be reasonably be argued for or against within our universe, one may question whether this truth has any importance. Are we just lost in Meinong's Jungle of abstract hypotheticals (Meinong, 1910)? For many years, this inaccessibility presented an impasse. However, in 2021 groundbreaking new work in computer science shattered this impasse. In a technical, 206-page paper entitled "MIP* = RE," Zhengfeng Ji, et al. for the first time proved that the answer to the Halting Problem is at least *knowable* (Ji et al., 2021).[12] That is to say that if Bob already had the answer to a specific halting problem (e.g., the answer was provided by an oracle), then Bob could share the answer with Alice in such a way that Alice could verify that the answer is correct. Furthermore, this verification process could be performed by Alice in a short, fixed period of time, despite the fact that Alice cannot possibly derive the answer herself (from the bottom up, within her closed system). This satisfies the fourth criterion, that the answer must be *knowable* within the closed system.

This radical new approach ("MIP*=RE") that makes answers to the Halting Problem knowable relies on "interactive proofs". Unlike standard, static proofs that build up theorems stratum by stratum, these proofs require interaction between an honest verifier and one or more untrusted provers. In general, the provers are assumed to possess abilities that the verifiers do not possess (e.g., much greater computational power, or access to an oracle). Thus, while the provers can obtain the answer to the problem, the verifier cannot replicate the procedure. Therefore, if the provers are to convince the verifier that they have actually solved for the correct answer (and are therefore not lying), they must provide incontrovertible evidence in a roundabout fashion known as an interactive proof.

---

12 We assume a broad notion of "knowable" here. The answer could be known by acquaintance, but not by description. It might be observable, but not inferentially explainable according to the known laws of the system. This idea of knowledge is one of the roots of empirical science. Of course, in a question, the sought answer is formulated in relation to a description. However, we assume that the exact content of the answer could be known by acquaintance. The interesting task, then, is to show afterwards that such a non-descriptive knowledge fits the initial description of the sought answer.

As a rough analogy for an interactive proof, we can imagine a game played between Alice, an honest verifier who is blind, and Bob, an untrusted prover who has good vision (Hartnett, 2020). Alice possesses two marbles which are identical in size, shape, density, and surface smoothness. Bob points out that while identical to the touch, these marbles are visually distinct (e.g., different colors). However, Alice won't automatically trust Bob's assertion and lacks the ability to verify the assertion directly herself. Instead, she devises a game to put Bob's assertion to the test. She shows one marble in her right hand and designates it as Marble #1. She shows the other marble in her left hand and designates it as Marble #2. Now she puts both marbles behind her back and swaps them a secret number of times (Bob cannot see behind her back). Now she again presents the marbles to Bob to test whether he can still identify Marble #1 from Marble #2. Note that Alice knows which marble is which by virtue of knowing how many times she switched the marbles behind her back. If Bob were to correctly distinguish between Marble #1 and Marble #2 just once, it could simply have been a lucky guess. But Alice can repeat the game as many times as she likes. If Bob continues to correctly distinguish between the marbles, time and time again, proof of his assertion (that they are visually distinct) mounts. The probability of Bob getting the correct answer in every game by chance alone is one in 2^N, where N is the number of games he has played. By playing the game over and over, Alice can effectively "prove" to within an arbitrarily tiny uncertainty (for example only one chance in a trillion) that Bob is not lying.[13] This analogy illustrates the nature of an interactive proof.[14]

Zhengfeng Ji et al. took the idea of an interactive proof and supercharged it. They asked what types of interactive proofs could be constructed if there were more than one prover working in tandem (with still a single verifier). Moreover, what if the multiple provers could not communicate with one another directly, but only shared links provided by quantum entanglement? In fact, "MIP" stands for "Multiple Independent Provers", and the "*" indicates that they share quantum entangled states. "RE" stands for the class of recursively enumerable languages. To great surprise, the researchers found that these types of interactive proofs could verify answers to the Halting Problem.

13 Note that in theory it would take infinite steps to bring Alice's credence that Bob is telling the truth fully up to 1. However, there is a rapid convergence towards 1 as the number of steps is increased. This is fundamentally distinct from insoluble halting problems that cannot be solved in a finite number of steps. In the case of the Halting Problem there is no convergence towards an answer. Therefore, a large finite number of steps could be taken without having any improved knowledge on whether the system will or won't ultimately halt.

14 Note that an "interactive proof" can provide compelling evidence of a fact (conditioned on events) in accordance with Bayes' Theorem (built on prior assumptions of the validity of mathematics). The "interactive proof" is a verification method. Fortunately, the iterative tests of the interactive proof can be extended to impose an arbitrarily high bar to pass, so as to overcome any finite level of a priori skepticism (i.e., any prior greater than zero). Nonetheless, despite the name "interactive proof", this verification method does not prove a statement beyond a shadow of a doubt, but instead it provides compelling evidence in support of gained knowledge. It is therefore potentially fallible in a sense.

In general, there is no shortcut to the Halting Problem. The only general way to determine whether a program will halt or not is to wait it out (wait for the computation to reach a point where it halts). This could take a year, a million years, a billion years, or infinite years— one just can't know a priori or predict it. (Moreover, for certain specific programs it is generally believed that no significant bottom-up shortcut (even a tailored one) exists—see Appendix A.) However, the "MIP* = RE" researchers showed that if some provers (Bob) knew that the program does in fact halt (perhaps after one billion years at a fixed computation rate) then Bob could use a relatively short interactive proof to convince the verifier (Alice) that the program does in fact halt.[15] Moreover, the short interactive proof has a small, finite number of computational steps that is independent of how long the program runs before it halts. For example, if the interactive proof takes ten seconds to carry out, it will take the same ten seconds to carry out regardless of whether the program takes one year or one billion years to halt.[16] Alternatively, if the program never halts (runs for infinite steps) then Bob cannot provide compelling evidence of this fact.

By providing objective, compelling evidence (interactive proof) to Alice that a specific program does halt (given a specific Turing machine construction), Bob has gifted Alice a small amount of additional knowledge (information) about her closed system. Prior to her interaction with Bob, Alice could enumerate the possible answers to her specific halting problem: 1) the program halts, or 2) the program doesn't halt. She may also have been able to put rough bounds on the probabilities of each outcome. Now, with Bob's help, if the program does in fact halt, Alice can refine those probabilities (to indicate that the program almost certainly halts). In information theory, this process of refining one's a posteriori credence of a statement, given the outcome of an event, is identical to gaining information. If, on the other hand, the program doesn't ever halt, then Bob will not be able to provide compelling evidence one way or the other, and Alice's initial credence of the statement will remain unchanged (no information gained). For our interests we will focus on the cases where Alice *does* gain information. (Note that if Alice keeps asking these types of questions for different, randomly seeded programs, she is very likely to stumble upon cases where she *does* gain information about her closed system that she could not have derived from the bottom up within her closed system—see Appendix A.)

Let's try to make this hypothetical interplay between the verifier (Alice) and the provers (Bob) a little more concrete. As a first example, in w_2, Alice would exist as a simulated conscious being within the laptop universe whereas Bob could be external, real-world scientists tapping into and interacting with this "laptop universe" through distinct internet connections. Bob could thus project their own distinct avatars into the laptop universe and communicate back and forth with Alice. Note that anything Bob present to Alice must be done in a way that is fully compatible with

15 It may be assumed here that Bob has access to (computational) capabilities well beyond those of Alice. Therefore, Bob can compute the answer in a bottom-up fashion while Alice cannot.

16 Assuming a fixed computation rate.

and containable within the laptop universe (Criterion 1). Nonetheless, Bob may additionally have access to external computational tools that don't or can't exist within the "laptop universe" (such as supercomputers or quantum computers) that enable Bob to derive answers that Alice could not.[17]

In an alternate scenario, we could envision that the closed system is w_1, and Alice is a human scientist within this universe. In this scenario Bob exist as alien beings from beyond the universe (perhaps from a parallel universe). Once again, Bob may manifest themselves within Alice's universe in order to communicate with Alice, but only in ways that are compatible with and containable within Alice's universe.

Alice, as a good scientist, would have to be a priori skeptical of the alien's fantastical claims of having solved the Halting Problem and would put the aliens to the test with an interactive proof. She must assume the laws of physics of w_1 and somehow verify that the aliens are not in direct communication with one another for the duration of her testing. For example, perhaps she would first space the aliens out on either side of her and at great distances. By posing questions to aliens on either side of her simultaneously (by transmitting the questions as light signals) and timing the aliens' responses (also light signals) Alice could verify (based on the timing of the return signals) that the aliens are not secretly communicating with one another within w_1 (colluding during the test), lest they be communicating faster than the speed of light. (Note that faster-than-light communication would break the rules of w_1, irrespective of whether our actual universe is thus constrained.)[18] After completing her full barrage of questions (the interactive proof), Alice will have the data that effectively "proves" something that can't be derived from the bottom up within her universe—that a specific instantiation of the Halting Problem does in fact halt.

Accepting an interactive proof conducted with aliens from another universe may seem like cheating. It may feel as if new information is being smuggled into the "closed system" universe by bringing in beings from *outside* the universe. However, this sentiment is partially misplaced. The aliens are not introducing a new truth into the closed system. They are simply illuminating a truth that already exists within the closed system.[19] The true answer to the specifically posed Halting Problem already existed within (and was entailed by) the scope of the closed system, it was simply unknown (and unknowable from the bottom up within the system). The aliens only helped to make it known (converting the true facts into gained

17 Note that Bob's assertions to Alice do not carry any weight on the basis of Bob's authority, credentials, or assumed methodology. Any weight that these statements carry is on the basis of the results of the interactive test that Alice conducts with Bob.

18 Note that this constraint on faster-than-light communication is only one such example constraint, not an absolute constraint. One may simply posit that there is a means of precluding any cheating facilitated by communication between Bob in the closed system (by alternate mechanisms).

19 Note that this particular example, as presented, still does require that the idealized universe be open to the input of non-random information from outside the idealized universe. Therefore, this particular example would not demonstrate strong emergence. However, in Sections 6 and 7 we describe how a similar effect may be achieved without the input of non-random information from *outside* the idealized universe. We further explore this possibility in relation to the Halting Problem in Appendix D.

information/knowledge that is instantiated in a person or machine within the universe). To further illustrate, we could imagine two metaphysically distinct hypothetical versions of our idealized universe, Universe A in which the specifically posed Halting Problem halts, and Universe B in which it does not. Everything else about the two hypothetical versions of our idealized universe is identical (as far as can be determined from the bottom up within the closed system). Furthermore, we will ensure that everything that the aliens do within Universe A and Universe B is strictly identical. In both hypothetical universes the aliens will act in the exact same way and communicate the exact same signals. However, in Universe A, the aliens' communications will result in conclusive proof that the specifically posed Halting Problem halts, whereas in Universe B, the exact same communications will result in inconclusive proof. (In Universe B the alien's ramblings will likely seem incoherent.) Therefore, even though Universe A and Universe B were fully identical to each other up to this point in time and both receive identical inputs from the aliens, the evolution of Universe A will diverge drastically from Universe B from this point in time forward as a result of different truths already contained within (and entailed by) the scope of each hypothetical universe. (In a sense, the hypothetical universes diverge in behavior due to a difference in their "ethereal information".) To reiterate: it is important to emphasize that the divergent evolution of hypothetical Universe A vs. hypothetical Universe B is independent of (not caused by) the content of the input provided by the aliens.

> Note 4.4.1: The truth (of whether or not the specifically posed Halting Problem halts) is likely to be logically entailed from the bottom up even if it is not derivable from the bottom up (with a bounded computational resource). For example, the answer to the Halting Problem might be entailed by the way in which Universe A and Universe B are physically constituted. Therefore, if Universe A and Universe B are identical in how they are physically constituted, they would both have the same answer to the Halting Problem. This, however, does not change the argument herein. In this thought experiment, Universe A and Universe B are not physically instantiated universes. Rather, they are two hypothetically distinct possibilities of a single universe. Universe A and Universe B are identical in all aspects of their physical constitution (as far as can be known from the bottom up within the system) and are only distinct with respect to their hypothetical answer outcomes (ethereal information), which cannot be known from the bottom up within the system.

> Note 4.4.2: The universes are bounded by their computational capacities. Accordingly, the necessary consequences entailed by the logic of the fundamental rules of the system may not be derivable from the bottom up within the closed system. Furthermore, if the consequences are not derivable, then under a digital physics backdrop they cannot be causally instantiated in a bottom-up fashion.

4.5 Criterion 5

Through the interactive proof conducted between Alice and Bob, Alice *gained knowledge* about her universe.[20] While it remains controversial as to whether pure information can have direct causal consequences on its own (likely not), when information is instantiated into a person or machine as gained knowledge (or gained information), it arguably can have physical, causal consequences. For example, the mere abstract fact that a specific Halting Problem does halt may not have causal power on its own, but if this fact is made evident to Alice, then she can act on this newfound knowledge in physical ways by telling her friends, typing up a paper, or traveling to give a talk on the topic. In fact, we can even remove any direct human or consciousness complications. We could just as easily replace Alice by a mindless robot or computer that will interact with the aliens, conduct the interactive proof, verify the results and store and act on the subsequent conclusion.[21]

The fifth criterion is that the resulting gained knowledge can be induced to have physical, causal consequences within the system. By instantiating the uncovered facts as gained knowledge in a person or machine, this criterion is easily satisfiable. For example, a computer could be programmed to trigger one action in the event that the evidence presented by the aliens conclusively shows that the specific Halting Problem halts (Case A), and do nothing if no evidence is presented, or if the evidence (interactive proof) is not conclusive (Case B). In the morbid extreme, the computer might be programmed to launch and detonate a massive arsenal of nuclear weapons if Case A occurs and do nothing if Case B occurs. Clearly the causal consequences of this gained knowledge can be made very physical indeed.[22]

It is noteworthy that the mechanistic behavior of the computer device described above (that can conduct the interactive proof, verify the results, and store and act on the subsequent conclusion) is effectively determined by three sources of

20 Once again, by "gained knowledge" we only mean to imply that Alice can use the event to update her credence about the answer to the Halting Problem, thereby gaining information (in a Shannon entropy sense).

21 Note that a mechanistic computer could only be induced to interact with, answer, or verify a problem that requires semantic understanding if it is set up to do so by an agent (e.g., a human) who has semantic understanding. Nonetheless, it is reasonable to believe that such a robot could be built with current technology and could mechanistically act on "ethereal information" that it gains and verifies. Furthermore, this robot would not require a mind, free will, or top-down causation in order to function.

22 If one were to argue that this knowledge cannot be made physically causal, then one would have to explain why a computer device or robot that performs both the verification and the actions prescribed could not be built and implemented. Furthermore, it is important to note that "ethereal information" is non-random. Therefore, a robot acting on ethereal information (such as the answer to a specific halting problem) involves the robot physically acting in response to a verifiable regularity/fact of the universe. Accordingly, this verifiable regularity has causal influence in addition to "truth value". A verifiable regularity of the universe that has causal influence but cannot be derived from the bottom up can be considered strongly emergent. This stands in stark contrast to a robot acting on purely random information (such as radioactive decay signals). This latter case would not indicate strong emergence.

information, not just the conventionally expected two sources of information. The two conventional sources of information that directly dictate the computer's behavior are its physical constitution (e.g., the location and connection of its fundamental particles along with the laws dictating the behavior of those fundamental particles), and the digital information fed into its processors (e.g., in bits including its preloaded software and the inputs from Bob). However, in an indirect sense, the computer's behavior is also determined by a third source of information, the actual fact of whether or not the specific Halting Problem does halt. This is a fact that cannot be derived from the bottom up within the bounded system, but it is nonetheless entailed from the bottom up and objectively true. We might therefore call it an *ethereal truth* about the system (i.e., ethereal information). This ethereal truth (whether or not the specific Halting Problem does halt) dictates whether the inputs from Bob pass the verification test in this hypothetical universe, and therefore, whether the nuclear weapons are detonated.

Additionally, we note that if the gained knowledge could not be made physically causal (and was perhaps only epiphenomenal), this would not rule out the possibility of strong emergence (see Appendix D for more discussion). In this work we are arguing for a set of sufficient, but not definitively necessary conditions for a particular delineation of strong emergence.

5. Deductive argument for strong emergence

In this section we lay out a deductive argument for strong emergence given the six sufficient criteria.

Thus far we have addressed five of the six sufficient criteria for strong emergence. We have shown that an objectively verifiable truth can be compatible/containable, knowable, and physically causal within our universe, even if it cannot be derived from the bottom up. The sixth and final criterion is the trickiest: to demonstrate that this truth is also *discoverable* within our universe without any outside influences (a la aliens).[23] For example, is it possible that even without external aliens, this truth could become clearly self-evident within the workings of a complex person or machine that is fully bounded by our closed-system universe? Such a discovery would

23 There is a subtle difference between "knowable" and "discoverable" as defined in our framework. "Knowable" indicates that there exists a method by which to verify the truth of the statement or at least to gain knowledge about and therefore update the likelihood of the truth of the statement conditioned on an event (a posteriori). This method must be compatible and containable within the system. For example, in some scenario, we might note that the likelihood that a statement is true will increase if $x + y/z > 1$. Therefore, given the existence of this inequality, the hypothetical statement is "knowable," whether or not the inputs x, y, and z actually exist. However, in order for the statement to be "discoverable", the inputs to the method must exist within the closed system (so that the inputs can be discovered, and the verification method instantiated within the closed system.)

persuasively indicate the existence of strong emergence. We contend that any physically causal phenomenon (i.e., objectively verifiable regularity) that is fully bounded by our universe but cannot be derived from the bottom up within our universe would constitute a clear example of strong emergence (under a digital physics backdrop).

In Section 6 we will argue that answers of the aforementioned variety might in fact be *discoverable* within our universe (Criterion 6). However, first we will summarize our overall argument in a deductive format.

Premise 1. A physically causal phenomenon that is *fully bounded* by the closed system of our universe and impossible to derive from the bottom up within the closed system of our universe is an example of strong emergence.[24]

Premise 2. A phenomenon is *fully bounded* by the closed system of our universe if it is a question/answer pairing wherein both the question and all possible answers are compatible with and containable within our universe, the answer is *knowable* within our universe, and the answer is *discoverable* within our universe.

Premise 3. An answer is *knowable* within our universe if it is objectively true within our universe and the methods and inputs necessary to convincingly argue for this truth are compatible with and containable within our universe.

Premise 4. An answer is *discoverable* within our universe if the inputs necessary to argue for this objective truth do exist within our universe. (*Discoverable* is a subset of *knowable*.)

Premise 5. An answer that is both *knowable* and *discoverable* can have physically causal consequences by the instantiation of the gained knowledge within a physical person or machine. (This assumes that a person or machine can act upon their gained knowledge.[25])

Conclusion 1. A phenomenon that is *fully bounded* by our closed system universe can be made physically causal within our universe.

Premise 6. There exist question/answer pairings that are *fully bounded* by our closed system but not derivable from the bottom up within our universe.

Conclusion 2. Cases of strong emergence exist within our universe.

---

[24] Note that if a new phenomenon is discovered that is not contingent on or supported by lower, more fundamental layers, then that phenomenon itself is a fundamental layer and is therefore derivable from the bottom up.

25 Note, however, that this does not require "free will".

So far, we have argued for the veracity all the key components of this deductive argument, except for the statement that the answer is *discoverable* (Criterion 6) within our universe. Unfortunately, we cannot prove out this last statement in full, but instead will argue for its plausibility by analogy and case studies (see Sections 6-7 and Appendix C).

6. The general case for criterion 6

In this section we describe informally what Criterion 6 might entail and argue for its plausibility by analogy.

How can convincing evidence[26] of a truth[27] be *discoverable*, if it is not *derivable* from the bottom up? One option is for the truth to pop into existence fully formed (self-evident/compelling) or in some such readily verifiable form. Given that this principle is difficult to conceptualize, we will attempt to clarify the idea using a photon-based analogy. According to Einstein's Theory of Special Relativity, nothing can accelerate from speeds below the speed of light up to the speed of light or through the speed of light. For a massive particle to do so would require that an infinite amount of energy be expended during the acceleration process (which is impossible in a universe constrained by finite resources). Once again, we have encountered a fundamental bound (or ontological gap) based on the clash between the finite and the infinite. Nonetheless, massless particles (most notably photons) *do* travel at the speed of light all the time. How do they get to this speed? They don't accelerate into it; they are born at this speed, fully formed. Photons don't take baby steps. They don't learn to walk before they run. As soon as a photon is formed (for example emitted from an electron relaxation process) it is already (and always) traveling at the speed of light. If strongly emergent phenomena are to be discoverable, they are likely to be born fully self-evident/compelling, in much the same way that photons are born at the speed of light.

Some scientists have also theorized the possibility of tachyons, which are hypothetical particles that are born at and always travel at speeds greater than the speed of light (Feinberg, 1967). Compared to normal matter, tachyons would have the opposite problem— they cannot decelerate down to the speed of light or any lower speed. Loosely, if we think back to the 90-foot ladder analogy, normal matter is

26 This evidence does not have to be convincing beyond a shadow of a doubt, but only sufficiently compelling so as to alter the a posteriori probability distribution for the veracity of the answer, conditioned on the evidence.

27 Once again, for our purposes, "truth" or "truth value" simply indicate that the information is non-random and can be verified in an objective sense (e.g., a fact, a law, a pattern, a regularity, etc.).

born and trapped on the lower 30 rungs, at speeds below the speed of light. Likewise, tachyons (if they exist) are born and trapped within the upper 30 rungs at speeds above the speed of light. Neither can cross the gap in the middle.

Bringing this back to the Halting Problem, we might posit an emergent property called "O" that allows for a person or machine to serve as an oracle for the Halting Problem. Moreover, this property might be born fully self-evident/compelling once the person or machine achieves a "latching state" afforded by a requisite type and level of complexity. That is to say that the property of "O" allows them to innately conjure the answer to the Halting Problem. Almost by definition we cannot explain directly how this property "O" works—our general pathway to understand/explain things is to derive them from the bottom up, which is necessarily impossible for any strongly emergent property. Nonetheless, persons or machines with this "O" property could take the place of the aliens (in Section 4) and conduct interactive proofs with other entities within our universe that lack the "O" property. Note that in this way, no appeal to an entity outside of the closed system is necessary. The alien provers from another universe that were used in previous examples could thus be dispensed with. In this way the discovery and verification of an answer that implies strong emergence could be completed fully within the closed system of our universe. The property "O" might involve obtaining several small bits of information by acquaintance (e.g., experience). For example, while asleep I might have a vivid recurring dream that gives me a strong feeling that a specifically posed Halting Problem does in fact halt. The bits of information obtained by acquaintance from the dream could then be used as inputs in a mathematically rigorous interactive proof in order to determine (through verification) whether these dreams are true visions resultant from the emergent property "O" or simply false hallucinations that do not reflect truth.[28] In Section 7 we provide an original case study describing an example verification process for gained knowledge that is discoverable but not derivable from the bottom up. In Appendix C, we provide additional original case study examples.

It is important to note that posing and/or answering the Halting Problem is a process that can only occur at high levels of complexity. Here "high complexity" is taken to mean a high rung on the ladder. Whether done in computers or brains, asking meaningful questions like the Halting Problem requires at minimum a level of complexity that can support semantic understanding. By contrast a few atoms interacting with one another could not meaningfully pose these types of questions. Semantic questions do not make sense to such small clusters of atoms. These questions/answers cannot be posed or made sense of at lower rungs of the ladder (that lack semantic understanding) and cannot be derived from these lower rungs (under finite computational resource assumptions). On the 90-foot ladder (analogy), it may be argued that these truths only begin to have viable existence some 40 feet into the air. Nonetheless they are verifiably true. These questions/answers can only exist at

28 False hallucinations might be considered random information, depending on their source.

higher levels of complexity.[29] Therefore, these questions and their answers supervene on the physical substrate that is contemplating the questions.

What is more, there is a pathway by which these truths could be instantiated as gained knowledge/information and induced to have physically causal effects. Importantly, this pathway (i.e., an interactive proof) only makes use of informational states that are fully compatible with and containable within the closed system of our universe (given rules established from the bottom up). This existing pathway makes these questions/answers *knowable*. What we do not necessarily know is whether the inputs required to go down this pathway exist and are accessible in our specific universe (i.e., are these questions/answers *discoverable*?).

It is likely the case that the objectively true answer to a specific Halting Problem is constrained (i.e., set to a singular value) based on the fundamental truths that exist at the lowest rung (e.g., fundamental axioms of mathematics). This, however, does not change the argument presented herein at all. If the true answer to a Halting Problem cannot be derived from the bottom up (within the closed system), then it cannot propagate causal influence from the bottom up from within the closed system (per a digital physics perspective).[30] If the true answer to a Halting Problem does have causal influence (when instantiated as information/knowledge) then that causal influence must enter at a higher rung (i.e., strong emergence).[31]

So far no one has come forth claiming to be an oracle for the Halting Problem (possess the property of "O"). We don't have any evidence to suggest whether or not this strongly emergent property exists within our universe. Nonetheless, in Section 7 and Appendix C, we present case studies which suggest that some types of truth might be *discoverable* even when they are not *derivable* from the bottom up.

7. Case study in support of criterion 6: testing for Boltzmann brains

In this section we present a case study illustrating how gained knowledge about a truth could be discoverable, even if it is not derivable from the bottom up. This case study is supplemented by additional original case studies in Appendix C. These case

29 Note that a mechanistic computer could only be induced to interact with, answer, or verify a problem that requires semantic understanding if it is set up to do so by an agent (e.g., a human) who has semantic understanding.

30 Note that this assumes a universe whose structure is similar to a computer, in which each subsequent timestep must be computed from the prior timestep(s) and any randomly injected information (e.g., from quantum mechanics).

31 To recap in brief, we contend that the answer to a specific halting problem is an objectively verifiable regularity pertaining to our universe. Furthermore, physical systems within our universe can be induced to physically act on this gained information/knowledge. Therefore, if the question/answer can only be stated at higher levels on the complexity ladder and the answer cannot be derived from lower levels on the complexity ladder, then the answer should be considered an example of strong emergence. The answer has causal influence that cannot (in principle) have propagated in a continuous fashion from the bottom rungs of the ladder up to higher rungs (while staying within the closed system).

studies illustrate the plausibility of discoverable knowledge absent derivability (even if in some cases the discoverable knowledge is only epiphenomenal.

We present an original case study highlighting a potential truth that might be discoverable even if it is not derivable from the bottom up. The purpose of this case study is to argue for the plausibility of criterion 6 (by example). In this case study, a conscious observer gains information about the likelihood that they are not a Boltzmann brain.[32]

A Boltzmann brain is a type of entity postulated by a famous thought experiment from physics that posits that conscious, physical brains can and will randomly assemble in outer space (over the course of very long timescales) purely as a result of random fluctuations (Cotzen, 2020). For example, it is very improbable for atoms to randomly take on the configuration of a functional brain at any specific time, but over sufficiently long timescales, atoms drifting about in outer space will take on every possible configuration, including configurations that constitute functional brains. Note that this occurs without any teleological impetus. The assembly of a low entropy state (like a functional brain) from a high entropy state is an allowed violation of the second law of thermodynamics if the experiment is conducted over sufficiently long timescales (e.g., the Poincaré recurrence time). These fluctuations could be either thermodynamic or quantum in nature. Cases involving thermodynamic and/or quantum fluctuations have been studied as thought experiments in cosmology (Davenport and Olum, 2010). Surprisingly, under a wide range of universe and/or multiverse conditions it has been predicted that the total quantity of Boltzmann brains that should (statistically) come into existence will vastly outnumber the total quantity of regular brains that will come into existence, i.e., brains that arise through a gradual process like evolution (Carroll, 2020). Furthermore, Boltzmann brains can pop into existence with false memories and false impressions of experiences. Boltzmann brains might come (randomly) preloaded with false memories of having a body, observing the external universe around it, or having lived a long life. To a conscious observer, these false memories of empirical data can be indistinguishable from genuine memories of the external world derived from actual sensory organs. Note that for the purposes of our thought experiment, Boltzmann brains necessarily lack sensory organs. Therefore, such Boltzmann brains cannot interact with or absorb empirical information from outside themselves.

This poses a conundrum of sorts. To illustrate, the predicted total number of Boltzmann brains that randomly happen to contain memories and experiences like my own will vastly outnumber the predicted total number of regular (non-Boltzmann) brains that contain memories and experiences like my own (Carroll, 2020). Therefore, by the principle of indifference, I might conclude that I am vastly more likely to be a conscious observer with a Boltzmann brain than a regular brain. However, if I am a Boltzmann brain, I cannot trust my memories or external experiences or any empirical data that I believe I have observed regarding the external world,

32 Note that gained information does not necessarily imply the discovery of an objective truth with absolute certainty.

because Boltzmann brains don't have sensory organs with which to make any observations about the external world. Therefore, I cannot use empirical data (including the empirical sciences) in order to argue whether or not I am in fact a Boltzmann brain or a regular brain.

It appears that any test that relies on empirical data about the physical universe or empirically derived science is potentially unreliable as it could simply have resulted from false memories or false external experiences. Therefore, it is likely impossible for a specific conscious observer to determine whether he or she is a Boltzmann brain or a regular brain using a bottom-up approach. Any fundamental physics axioms about what the physical world consists of and how it operates (i.e., the axioms that would provide the necessary foundation for the bottom-up derivation) are suspect precisely because they must be empirically founded. Nonetheless, we propose that under a few assumptions there is an interactive proof that can help a conscious observer to statistically constrain the possibility that they are a Boltzmann brain. (These assumptions follow from rationalism, not empiricism.) The statistical constraints that this interactive proof provide constitute gained information/knowledge about the truth of the matter (whether the observer is a Boltzmann brain or a regular brain). This interactive proof provides an example of a truth that is discoverable even when it is not derivable from the bottom up.[33] In a way, we are taking a rationalist stance for this case (although the general notion of information/knowledge we have been considering does not depend crucially on this). In principle, a rational agent (John) could gain knowledge through a non-empirical interactive proof about whether he is likely a Boltzmann brain or a regular brain.

The assumptions for conducting this interactive proof are as follows.[34] First, we assume that an observer (either a Boltzmann brain or a regular brain) is presently aware of their own consciousness and thinking.[35] Second, we assume that mathematics and logic are a priori valid and consistent, independent of the empirical,

33 We assume that no empiricist or rationalist *bottom-up* approach could answer this same question. An empiricist bottom-up approach would require empirically derived axioms that cannot be trusted if obtained in a Boltzmann brain that lacks sensory organs. At the same, a purely rationalist bottom-up approach (without access to empirical axioms) would likely be able to equally arrive at both answers (Boltzmann brain vs. regular brain) without being able to select one answer over the other (therefore precluding a rationalist bottom-up resolution to the question). We leave further exploration of the limitations of such bottom-up approaches to future work.

34 We contend that these eight assumptions are reasonable and defensible metaphysically for a possible idealized universe (regardless of how our universe is actually constituted). Others might posit different fundamental assumptions and create a different test accordingly. Nonetheless, the point is to illustrate a potential example wherein knowledge can be verifiably gained within an idealized universe even if that knowledge cannot be derived from the bottom up.

35 We assume that the observer can distinguish between the sensation of their *current, present* awareness of their own consciousness and memories that might suggest that they were conscious in the past. Note that such memories might be false memories. Therefore, the observer can might have false, past memories of having performed and passed the proposed test, but these memories cannot be relied upon. The observer can only trust thoughts from their current, presently known "continuity of mind".

physical universe and that the observer can consciously and knowingly perform mathematical and logical operations. Third, we assume that larger and more complex Boltzmann brains are less likely to form than smaller or less complex ones (and furthermore that the larger and more complex Boltzmann brains remain functionally intact, by chance, for a shorter duration of time). This follows from statistical (mathematical) arguments concerning the likelihood of low entropy states forming by chance alone (and remaining in that functionally intact, low entropy state for a finite period of time). Fourth, we assume that the computational resource available to a Boltzmann brain (i.e., total number of computational steps that can be undertaken while functionally intact) is constrained by its size, complexity, and the time over which the Boltzmann brain remains functionally intact. Fifth, we assume that an observer can act on random information that is created moment by moment and not fully determined by prior events (e.g., random inputs from quantum mechanics—equivalent to access to truly random coin flips).[36] Sixth, we assume that P≠NP (see Appendix B for elaboration). Seventh, suppose that John can knowingly possess continuity of mind (free from false memories pertaining to his own internal thinking during this period) for at least a small, finite period of time (presumably the actual time over which the Boltzmann brain is functionally intact).[37] Eighth, we assume that there is no retrocausality (i.e., that events in the future cannot causally affect events in the temporal past, taken from John's perspective and within John's static frame of reference). Note that John can become convinced of assumptions 1,2,3,4,6, and 7 by processes of internal thought and reflection (taking a rationalist stance).[38] Even if John cannot become convinced of assumptions 5 and 8, he can still gain knowledge (from the proposed test) that could not otherwise be derived from the bottom up. He could gain the knowledge that either he is not likely to be a Boltzmann brain, or that assumptions 5 and/or 8 do not hold.

Using these assumptions, we can create a test for an individual (e.g., John) to determine if it is statistically unlikely that he is a Boltzmann brain. This test will help to determine the relative likelihood of whether John is actually perceiving and interacting with an external world, or whether he is simply imagining it through false memories/experiences in his brain. We suggest that the blockchain could be an exemplar way to implement this test—a suggestion we will not dig into in detail in this work but will investigate as future work. Nonetheless, we note that a similar

36 Note that this assumption does not require free will. It only requires that truly random information exists within the system and can have causal influence.

37 We assume that while John is in a state of "continuity of mind", he might still have continued false experiences of an external world but will not have false memories of things that he has thought or computations that he has mentally performed within this period of time (and furthermore is aware of the trustworthiness of his internal, mental thinking and computational processes while in this state). This case study assumes a type of rationalist stance.

38 These assumptions fall into three broad buckets (each of which John can assess by internal processes without external empirical data): basic assumptions (e.g., the validity of mathematics and logic), consequences of mathematics and/or logic (e.g., P≠NP, statistical arguments), and assumptions based on direct awareness/intuition (e.g., awareness of one's own present consciousness).

test can be constructed from any NP-complete problem (even if it is not mediated by a blockchain). Note that an "NP-complete problem" is a class of problems in computational complexity theory. Each problem in this class has a solution that can be verified in polynomial time but can only be solved in non-deterministic polynomial time. Moreover, any NP-complete problem can be used to simulate any other NP-complete problem (see Appendix B for additional information).

Test to assess likelihood that John is not a Boltzmann brain.

1) John identifies a system that presumably exists in the physical world independent of and external to his brain (from his perspective) and involves solving computationally expensive NP-complete math problems with regularly published results. For example, a blockchain could be established wherein the proof of work requires solving NP-complete math problems and publicly disseminating their solutions for each block as it is completed. John (from his perspective) would perceive the blockchain to exist and function independently from his brain.

2) On the basis of random information generated moment by moment (e.g., through quantum processes), John undertakes an action to manipulate the input to the NP-complete problem that is next in line to be solved. For example, John measures radioactive decay to obtain a random number and then sends that quantity of cryptocurrency to an address on the blockchain. A cryptographic hash of the next block on the blockchain with this random transaction information will be distinct from a cryptographic hash of the next block without this random transaction information. Assuming that the cryptographic hash sets the seed conditions for the next NP-complete problem to be solved when mining the next block, John has succeeded in his manipulation of the blockchain.

3) John checks (directly using mental math) that the published solution pertaining to the next block is a valid solution to the NP-complete problem that was posed (the specifics of this problem are contingent on John's prior manipulation of the blockchain inputs).

4) If John can complete this entire verification process while maintaining continuity of mind, then he can have increased confidence that he is not a Boltzmann brain (and that he is in fact interacting with a world external to his own brain).

This test functions on the basis of two distinct notions. First, John must argue that it is highly improbable that the correct answer to the NP-complete problem would have been found by dumb luck alone. Second, John must argue that it is highly improbable that John's subconscious would have performed the required computation (unbeknownst to his conscious self) in order to feed the correct answer to his conscious mind.

First off, John can mathematically analyze the probability that the published solution to the NP-complete problem is in fact the correct solution by random happenstance (dumb luck alone). This probability can be determined by considering the solution as a pair mapping between two numbers, the seed number that defines the problem, and the solution number to the problem. By considering the size and

distribution of the set of possible solutions (given a specific seed number), John can compute the likelihood of a seed number and its correct solution number being randomly paired together. In general, this likelihood will be exceedingly low, so as to render the hypothesis that this fortuitous (correct) pairing occurred by random happenstance untenable.

Alternatively, John could posit that while his conscious mind was unaware of how the solution was determined, his subconscious mind was hard at work solving the NP-complete problem in a straightforward and computationally expensive manner. The first question that arises is, "Why would John's subconscious do such a thing?" It is conceivable that his subconscious might be required to perform some interpretation, rectification, error correction, cognitive dissonance dissipation, or fill-in-the-blank work to ensure that John has a reasonably sane, acceptable, or coherent experience for his conscious mind. This might (or might not) be a minimum criterion for John to function as a conscious observer at all.[39] However, it is difficult to fathom that John's conscious mind could not perform as a functional observer if obscure mathematical details in the weeds of his perceived reality did not fully check out. For example, if John found out that the published solution to the blockchain did not mathematically check out, he would likely not be rendered insane or incapacitated to the point of no longer existing as a conscious observer. Rather, he would more likely believe that there was a bug in the blockchain mining code, and that the blockchain as it was set up was flawed. Therefore, if John were a Boltzmann brain, it is difficult to understand why John's subconscious would expend such a computational resource to ensure the validity of mathematical details of such minor importance.

However, this proposition (that John's subconscious solved the NP-complete problem unbeknownst to his conscious mind) becomes even more untenable under the assumptions that computational resources are limited by the size and complexity of a Boltzmann brain and that there is an (exponential) inverse correlation between the size/complexity of an object and the likelihood of that object assembling from random fluctuations. (This exponential inverse correlation comes from statistics, and we assumed the trustworthiness of mathematics from the outset.) John may test his conscious mind and its computational power using phenomenological tests (e.g., one such test might check how quickly John can consciously solve math problems of a certain type). He may also presume that the computational resource needed by his subconscious mind is a finite quantity related to the computational power (scope) of his conscious mind. The necessary computational resource of John's subconscious mind would be based on the subconscious computational work necessary to support John's conscious mind (e.g., maintain sanity or conscious observer status). Arbitrarily adding the need to solve computationally expensive NP-complete problems could significantly increase the necessary computational resource of John's subconscious mind. Furthermore, the difficulty of the posed NP-complete

39 Here we do not attempt to adjudicate on the issue of what requirements are necessary for an entity to be considered a conscious observer and instead leave this as an open question for future work.

problems can be scaled somewhat arbitrarily. Therefore, if John is in fact a Boltzmann brain that can solve arbitrarily large and computationally expensive NP-complete problems, then John's subconscious mind must have an arbitrarily large computational resource.

In the absence of trustworthy empirical data about the physical universe, we cannot put definitive limits on John's subconscious computational capacity (if he is a Boltzmann brain). However, we can readily show that Boltzmann brains that have expanded subconscious computational resources in order to solve computationally expensive NP-complete problems should be vastly outnumbered by Boltzmann brains that do not have these expanded subconscious computational resources.[40] This follows from the assumption that computational resource is constrained by size/complexity of the Boltzmann brain and increasing size/complexity of assembled objects exponentially decreases their likelihood of assembly. Therefore, by the principle of indifference, John is much more likely to be a Boltzmann brain without expanded subconscious computational resources (if he is a Boltzmann brain). This argument, therefore, makes it highly unlikely that John's subconscious could or would solve the NP-complete problem.

Additionally, there might exist a statistically constrained maximal computational capacity for any Boltzmann brain. This follows from the fact that the total number of computations that a Boltzmann brain can carry out (before it falls apart) is a product of the rate at which it can compute (correlated with its size) and the time over which it can compute before it falls apart or becomes non-functional (inversely correlated with its size). As the size of the Boltzmann brain increases and its computational speed increases, its functional lifespan decreases, suggesting that a maximal computational power might exist.[41]

In recap, John can test whether he is a Boltzmann brain by checking the solutions to asymmetrical math problems that are difficult to solve, but easy to check the solution. If John believes he lives in an external world where a blockchain regularly publishes these solutions, he can test the validity of one such solution, following each step in his mind throughout a finite time period wherein he has continuity of mind. If this blockchain truly exists in the external world, outside of John's own mind, then it is making use of significant computational resources pooled together by blockchain miners from around the world. If John is truly interacting with an external world, this explanation makes sense. However, if John is a Boltzmann brain, his subconscious would have had to race ahead of his conscious mind by

40 Note that the ratio between the expected number Boltzmann brains with vs. without the capacity to solve large NP-complete problems has an exponential dependence on the ratio between the computational resource needed for each type of Boltzmann brain. Therefore, small increases in the computational expense of solving the posed NP-complete problem lead to significant increases in the credence that John in not a Boltzmann brain. It is therefore advantageous (for this thought experiment) to take advantage of the most computationally expensive solved and publicly published NP-complete problems available, such as those that can be found in blockchain mining.

41 Note that this case study does not crucially depend on the existence of a maximal computational capacity for Boltzmann brains.

using vast computational resources in order to solve the NP-complete problem correctly on its own. (Note that John's conscious mind has a sufficient computational resource to verify a published answer rapidly, but not to find the answer.) The possibility of John's subconscious successfully racing ahead is highly unlikely for the aforementioned reasons. Therefore, if John finds that (upon conscious verification) the published solution to the NP-complete problem is correct, then he can have increased confidence that he is truly interacting with an external world and is not a Boltzmann brain. He has gained information about the truth of his circumstances based on an event (the results of his verification efforts).[42]

> Note: It is important that John verify the solution to the math problem himself and during a period of continuity of mind. He cannot outsource this step to anyone or anything else as these could yield false memories/results. Furthermore, it behooves John to verify the solutions of the most computationally expensive NP-complete problem that is available to him as his confidence (that he is not a Boltzmann brain) can grow exponentially with the computational expense required to solve the math problem. In this work we only outline this case study. We leave a more rigorous formalization of this case study argument as future work.

This case study showcases an example of a truth (gained information/knowledge) that could be learned through a type of interactive test, but not derived from the bottom up. Here John's direct mental verification of the solution to the math problem is like an interactive proof. However, this truth (that John is likely not a Boltzmann brain) could not be derived from the bottom up (i.e., from some sort of first principles) given that empirical evidence about the physical universe could not be trusted. This truth is therefore potentially discoverable even though it is likely not derivable.

> Note: The test described above makes use of statistics and random sampling and could potentially be applied in other skeptical scenarios (e.g., brain in a vat) as long as random sampling criteria are met. Other applications of this argument are beyond the scope of this work.

In Sections 6-7 and Appendix C we have argued that truths may be discoverable within a closed system even if they cannot be derived from the bottom up (Criterion 6). First, we gave an analogy to help conceptualize how strong emergence might be

---

42 Of course, if John is too skeptical to believe in the 8 assumptions we have laid out, then this test will not work for him. For example, if John does not trust the logic or mathematics performed by his mind, then this test will not work. However, if John does trust the 8 assumptions we have laid out, then by performing the test on increasingly difficult math problems, John can gain increased confidence that he is not a Boltzmann brain. In theory, if arbitrarily large math problems can be used (and John trusts the 8 assumptions), then John's increased confidence that he is not a Boltzmann brain can overcome any prior prejudice to the contrary (any prior greater than zero).

born fully formed (self-evident/compelling). Second, we presented a case study of a truth that might be discoverable within our universe even if it is not derivable from the bottom up within our universe. This case study was presented as a brief sketch of an argument without going into significant detail, which is left for future work. Additional original case studies are presented in Appendix C. It is notable that the case studies we have presented in this work all rely on a conscious observer and private information (that is accessible to the conscious observer themselves but is seemingly not equally accessible to an outside party). It is unclear whether all discoverable but non-derivable phenomena must share in these traits or whether these case studies are limited by the authors' abilities to imagine examples of a totally different nature. We leave these further questions as future work. Furthermore, we note that in each case study, knowledge by acquaintance is accumulated and used as evidence in a process that loosely resembles an interactive proof wherein informational knowledge can be gained.

In our overall work, the use of the Halting Problem and its verification by the interactive proof of "MIP*=RE" is critically important. Within the backdrop of digital physics, the verification of answers to the Halting Problem makes formal and explicit use of an interactive proof between multiple external parties that demonstrates verification in an external fashion that is credible to external observers (unlike case studies #1-#4). This external verification increases the reasonableness of assuming that the gained knowledge can be made physically causal and affect the system at large. Case studies #1-#4 are only used to illustrate the plausibility of truths that are discoverable but are not derivable. The deductive argument (Section 5) that makes use of the six sufficient criteria is based upon the Halting Problem and "MIP*=RE". We also note that it is possible that facts of the Boltzmann brain case study type are discoverable, while facts of the "O" type are not. However, if true, this would lead to questions for future work, namely determining the bounds on what types of truths are discoverable but not derivable and why those bounds exist (see Appendix D for further discussion).

Taken together with the previously discussed criteria, Criterion 6 completes the deductive argument for strong emergence (Section 5). We have analyzed known cases of question/answer pairings that are 1) compatible/containable, 2) non-random but objectively true, 3) non-derivable, 4) knowable, and 5) have the potential to be made physically causal within our universe. Moreover, we have argued for the plausible existence of truths that are 6) discoverable, even if they are not derivable from the bottom up.

8. The bigger picture

In this section we briefly summarize one possible reason why the existence or non-existence of strong emergence remains an open question.

If strong emergence really does exist, why haven't more cases been observed and reported? We believe this may result from a selection bias that has more to do with where and how scientists and philosophers look for strong emergence, than the actual abundance or scarcity of strongly emergent phenomena. To illustrate, when Kurt Gödel set off to determine whether truths exist that cannot be derived from the bottom up within their own formal systems, the answer was not obvious. In order to prove that at least one truth exists that cannot be derived from the bottom up, Gödel had to devise subtle, ingenious, and incredibly roundabout proofs. Even today we have no idea whether underivable truths of the type that Gödel identified are rare or commonplace. By definition, there is no direct path for finding them. (Moreover, it is likely that if strong emergence were encountered it might be difficult or impossible to identify it as such in certain cases.) If a phenomenon cannot be predicted from the bottom up (strong emergence) and is not known from direct exposure, then it tends to evade our scientific tools as we know them. Reductionism has proven an incredibly powerful tool in the scientific toolchest thus far. It is therefore easy to extrapolate and assume that what has worked so well in some domains will continue to work through every domain. However, this is an unsupported extrapolation—an overextrapolation.

In general, we test our assumptions by looking for agreement between prediction and experiment. We might test the reductionist assumption that there are no strongly emergent causal properties at any level except for the bottom rung by predicting the results of complex phenomena from the bottom upwards and comparing predicted results to real-world experiments. However, to date, scientists cannot predict complex phenomena (or their downstream effects) at any reasonable scale. The world's fastest supercomputers still struggle to simulate more than about a billion atoms interacting together at a time. This size is approximately 15 orders of magnitude smaller than the scale at which humans live. Since we can't yet make predictions regarding complex phenomena at an appropriate scale, we cannot and have not begun to test the assumptions of reductionism. We cannot reasonably begin to rule out the possibility that strongly emergent properties have causal effects—chaos theory has thus far prevented us from conducting the necessary simulations and experiments. At present, there is simply no sufficient experimental data for or against the hypothesis of reductionism. Reductionism is an interesting, though untested hypothesis.

Conclusion

In this work we have developed a conceptual framework for strong emergence and argued for the plausibility that strongly emergent phenomena exist in our universe. An important facet of the conceptualization of strong emergence presented in this work is that while these strongly emergent phenomena have causal power to influence physical events, that causal power cannot have propagated from the bottom up. We have identified six sufficient criteria for strong emergence: a question/answer pairing that is 1) compatible and containable, 2) non-random but objectively true, 3) non-derivable from the bottom up, 4) knowable, 5) physically causal, and 6) discoverable within our universe. Furthermore, we have presented a deductive argument based on six premises and two conclusions in support of strong emergence (based on the six sufficient criteria). Of the six sufficient criteria, we make a case for the real existence of five of the criteria and argue for the plausibility of the sixth criterion by analogy and case studies. If discovered, the existence of strongly emergent phenomena would shatter the perspective that all of reality is reductionistic and would have implications for questions regarding free will and other philosophical questions. The computer science-based framework presented in this paper (used for the analysis of strong emergence) will be applied to these alternate philosophical questions in future work.

Appendices

Appendix A
Information about the Halting Problem, contiued

At several points throughout this work, we refer to a "specific", "specifically posed", or "specific instantiations" of the Halting Problem. This language is shorthand used throughout the main text for the purpose of brevity. Herein we will address more explicitly what types of questions qualify under this category for the interests of this work. To summarize up front, we are interested in programs for which there is reason to believe that it is *impossible* to determine whether or not the program will halt (from the bottom up, within the computational limits of our universe or closed system). We refer to such programs as "specific", "specifically posed", or "specific instantiations" of the Halting Problem.

The Halting Problem (which Alan Turing proved to be undecidable over Turing machines) asks the question as to whether there exists a general algorithm that can accurately determine whether any particular program/input pair (*x*) will halt or not halt, for all possible *x* on a Turing machine (*M*). Here a program/input pair refers to a program that has been uniquely modified by a specific seed number (i.e., an input). It has been proven that this general algorithm cannot exist.

However, this does not imply that for any given program/input pair it is impossible to find a shortcut to determine whether or not that *x* on that *M* will halt. Indeed, there are many trivial such *x* for which it is readily apparent whether or not each will halt. Two examples are provided below:

The program (pseudocode)

```
print "Hello, world!"
```

will obviously halt. Whereas the program (pseudocode)

```
while (1=1)
     print "Hello, world!"
```

will obviously continue forever and never halt.

As the programs become more complex, the question as to whether or not each will halt becomes less and less tractable, and shortcuts to determine this truth become less and less obvious. In fact, we contend that for some special classes of programs it is reasonable to believe that no significant shortcut can be used to determine

whether or not *x* will halt when run on *M*. We defend this assertion with the following argument:

> 1) Does each and every unique computer program that halts have a unique (though unknown) shortcut that accurately indicates that the program in question will halt? 2) Can every shortcut of the aforementioned type complete its execution within the same fixed, finite number of steps, *N*? If both questions were answered in the affirmative, then given any arbitrary computer program one could simply search the entire finite space of shortcuts of less than *N*+1 steps.

If the shortcut were found during this finite search, then it would be known that the program halts. If no shortcut were found during this finite search, then it would be known that the program doesn't halt. However, this process, if possible, would provide a general solution to the Halting Problem (and a way to calculate Chaitin's constant, which is known to be impossible (Chaitin, 1975)). Thus, by contradiction, this process is not possible. Therefore, we can conclude that not every program that halts will have a shortcut that takes less than *N*+1 steps. If we choose *N* to be greater than the total number of computational steps allowed within our computationally bounded system, then we can show that computer programs exist that will halt, but for which no shortcut can be successfully run to completion within our closed (computationally bounded) system.

Below we explain an example of one such special class of programs for which it is reasonable to believe that no shortcut can be run to completion within our closed system.

One of the most frustrating features of Gödel's incompleteness theorems is that they prove not only that there exist questions whose truths cannot be derived from the bottom up within their own formal systems, but also that we cannot know definitively which questions fall into this category. Accordingly, we cannot know for certain which question/answer pairings cannot be derived from the bottom up, but we can be assured that they do exist, and can even make guesses as to which questions are most likely to fall into this category. One strong contender (as a question/answer pairing that cannot be solved from the bottom up) is Goldbach's conjecture (Wang, 2002).

Goldbach's conjecture (in number theory) asserts that every even whole number greater than two can be expressed as the sum of two prime numbers. This conjecture has been successfully tested for every even whole number up to 4×10^{18} and has resisted every attempt at a proof since it was first proposed in 1742 (Oliveira e Silva et al., 2014). Many mathematicians suspect that Goldbach's conjecture is in fact true but unprovable. No entity that is constrained by finite resources could possibly prove the truth of Goldbach's conjecture by directly testing every even number—there are infinite even numbers and infinite prime numbers. If no roundabout, finite-length proof of Goldbach's conjecture exists, then its truth (if true) cannot be verified from the bottom up. Alternatively, Goldbach's conjecture may be false. There may exist one or more counterexamples—very large even numbers that haven't yet been tested that cannot be written as the sum of two primes.

Interestingly, Goldbach's conjecture can be posed as a Halting Problem (on a Turing machine with infinite memory). First, we can write a simple subroutine that will take an even whole number, u, as an input and check whether or not u can be expressed as the sum of two prime numbers. (If we are not concerned about efficiency, we can simply have the subroutine check the sum value for every combination of two prime numbers where each prime number is less than u.) Next, we can write a program that will take this subroutine and progressively run it on every even whole number starting with the number four (incrementing the input by two in each step). If the subroutine finds an even number that cannot be expressed as the sum of two primes, then it will trigger the program to halt. Otherwise, the program will continue running indefinitely. Therefore, if we could determine whether or not this program (which can be written as a finite set of instructions) halts, then we could solve Goldbach's conjecture (one way or the other). The answer to this Halting Problem would give us the answer to Goldbach's conjecture.

Let us assume for the moment that Goldbach's conjecture is false and there exists at least one counterexample—i.e. an even whole number that cannot be expressed as the sum of the two primes. (Here we will assume that no finite length proof exists that can identify all such counterexamples.) The smallest counterexample may be such a large number that running the program described above will not reach the counterexample within the finite number of computational steps allowed by our finite universe. Nonetheless, it may be possible (though improbable) to find the counterexample by simply guessing and checking random even numbers. Therefore, even though the program may not find the counterexample through its process of incrementation, the ultimate fate of the program (that it would eventually halt) might (improbably) be determined by a guess and check process (shortcut). This would make the answer to this Halting Problem improbable to find but not impossible to determine within the confines of our universe.

For the purposes of this paper, we are only interested in programs for which it is believed to be i*mpossible*, not improbable (from within the confines of our finite universe) to determine in a bottom-up fashion whether or not the program should eventually halt. Therefore, we will modify the program described above to instead pose a modified version of Goldbach's conjecture as a Halting Problem. In lieu of Goldbach's conjecture, which asks whether or not ALL even whole numbers greater than two can be composed as the sum of two primes, our Modified Goldbach's conjecture only applies to a subset of all even whole numbers.

Our Modified Goldbach's conjecture problem, posed as a Halting Problem, is composed as follows. First an arbitrary seed number, y_n, is randomly selected (where y_n must be an even whole number greater than two). Second, the aforementioned subroutine is used to determine whether y_n is the sum of two prime numbers. If so, then y_n and its two prime addends are fed into an inherently sequential algorithm that undergoes a predetermined computation. An inherently sequential algorithm cannot be sped up by parallelization (assuming NC ≠ P). (Note that NC comprises the set of decision problems that are decidable in polylogarithmic time when using a parallel computer with a polynomial quantity of processors. P comprises the set of all decision problems that can be solved in polynomial time on a deterministic

Turing machine. It is unknown whether or not NC = P, but it is generally believed that NC ≠ P.) The results of this computation are then fed into a cryptographic hash function. In turn, the output of the hash function is fed into a second function as its argument. This second function will select for a deterministic but arbitrary even number, z. This whole process is then repeated for a new value of y_n, y_{n+1}, where $y_{n+1} = y_n + z$. This process will continue to iterate as described unless a value of y_n is found which cannot be composed as the sum of two primes, in which case the program will halt.

As can be seen, this modified program cannot be shortcut by a guess and check process. One cannot know a priori which even whole numbers will be a part of the subset considered. This can be ensured by the appropriate selection of an inherently sequential algorithm (assuming NC ≠ P). Therefore, a guess and check strategy might uncover counterexamples to Goldbach's conjecture, but these counterexamples on their own will not indicate whether or not the modified program will halt. This is because the counterexamples might not be included in the subset of even whole numbers considered by this program.

If counterexamples to Goldbach's conjecture do exist, then for certain seed values that are input into the modified program, it will be impossible to determine (from the bottom up, within the computational constraints of our universe) whether or not the program will halt. This is because the quantity of computations necessary to uncover counterexamples (that fall within the subset of even whole numbers considered by the program) will exceed the finite computational limits of our finite universe. Note that this can be arranged even for a Turing machine with a large, but finite memory. Furthermore, we note that there is likely no way to determine (a priori) which inputs will result in specific instantiations of the Halting Problem that cannot be decided from the bottom up within the constraints of our finite universe. Nonetheless, we argue that if counterexamples to Goldbach's conjecture do exist, then by posing a multitude of different instantiations of the Halting Problem (as described), each with a different arbitrary seed, there is a possibility and a probability of stumbling upon a program with the desired criterion (i.e., cannot be derived from the bottom up within our universe.)[43]

In this appendix we have elaborated on the key criterion indicated when we refer to a "specific", "specifically posed", or "specific instantiations" of the Halting Problem. To recap, this criterion is that it is impossible to determine in a bottom-up fashion from within the constraints of our computationally bounded universe whether or not that program will halt. Furthermore, we have identified a class of programs that may meet this criterion (based on a modified version of Goldbach's conjecture). Different instantiations of this program can be produced by changing the initial seed number. We note that this is likely just one of many classes of programs that might meet the desired criterion. Further exploration is likely to uncover

43 Note that just as we cannot definitively know whether a specific truth is unprovable (by Gödel's incompleteness theorems), we cannot know specifically which seeded programs cannot be computed from the bottom up given a finite computational resource. Nonetheless, the mere likelihood of encountering such a program is sufficient to test for strong emergence.

other such classes of programs (possibly based on other open questions in mathematics such as the Collatz Conjecture, etc.).

A key idea in this work is that there exist programs for which it is impossible to determine whether or not they will halt eventually, from a bottom-up perspective and given the finite computational constraints of our universe. At the same time, based on the recent results from "MIP* = RE", answers to the Halting Problem are verifiable in cases where the true result is "halt" (Ji et al., 2021). Therefore, while an answer to a specifically posed Halting Problem cannot be derived from the bottom up, it can be verified. For an informal discussion of this result, see the excerpt below written by Henry Yuen (2020), an author of "MIP*=RE".:[44]

> In the Halting problem, you want to decide if whether a Turing machine M, if you started running it, would eventually terminate with a well-defined answer, or would it get stuck in an infinite loop. Alan Turing showed that this problem is *undecidable*: there is no algorithm that can solve this problem in general. Loosely speaking, the best thing you can do is to just flick on the power switch to M, and wait to see if it eventually stops. If M gets stuck in an infinite loop — well, you're going to be waiting forever.
>
> *MIP* = RE* shows with the help of all-powerful Alice and Bob, a time-limited verifier can run an interactive proof to "shortcut" the waiting. Given the Turing machine M's description (its "source code"), the verifier can efficiently compute a description of a nonlocal game GM whose behavior reflects that of M. If M does eventually halt (which could happen after a million years), then there is a strategy for Alice and Bob that causes the verifier to accept with probability 1. In other words, $\omega^*(GM)=1$. If M gets stuck in an infinite loop, then no matter what strategy Alice and Bob use, the verifier always rejects with high probability, so $\omega^*(GM)$ is close to 0.
>
> By playing this nonlocal game, the verifier can obtain *statistical evidence* that M is a Turing machine that eventually terminates. If the verifier plays GM and the provers win, then the verifier should believe that it is likely that M halts. If they lose, then the verifier concludes there isn't enough evidence that M halts. The verifier never actually runs M in this game; she has offloaded the task to Alice and Bob, who we can assume are computational gods capable of performing million-year-long computations instantly. For them, the challenge is instead to *convince* the verifier that if she *were* to wait millions of years, she would witness the termination of M. Incredibly, the amount of work put in by the verifier in the interactive proof is *independent* of the time it takes for M to halt!

44 Note that in this excerpt, both Alice and Bob are provers. In the main text, Bob is the prover and Alice is the verifier. This is only a difference in labeling.

Appendix B

An aside on why NP-complete problems cannot be substituted in place of the Halting Problem for the purposes of this work

To readers who are familiar with cryptography, the notion of a true statement that cannot be derived (within a closed system) but can be verified might sound vaguely familiar. Cryptographic protocols make use of mathematical problems for which finding the correct answers is ludicrously difficult, and yet verifying correct answers once they have been discovered is relatively easy. This is almost analogous to a "Where's Waldo" puzzle book. It could take someone hours to locate Waldo's hidden location, but once he or she points Waldo's location out to a friend, the friend can verify immediately that Waldo has indeed been found. As a more mathematical example, given a large composite number, it is computationally very expensive to find its unique prime factors (using conventional computing). However, if already provided with the prime factors, a simple multiplication would verify that the given answer is correct. There are many other mathematical problems that have similar structures: very difficult to find the answer, but easy to verify the answer if it is provided. These "asymmetrical" problems can be found in modular exponentiation, cryptographic hashes, and the general class of NP-complete problems.

In fact, it is relatively straightforward to use an asymmetrical problem to pose a question with a true and "knowable" answer that is *unlikely* to ever be found within the computational constraints of our universe. In this case, the odds of stumbling upon the correct answer are stacked heavily against. However, small though the odds may be, they are not zero. This is a subtle, but very important distinction (from the Halting Problem). For appropriately selected specific Halting Problems, the probability of deriving (and knowing) the correct answer from the bottom up within our universe is truly zero. There is no guess and check process by which a fortuitous guess might give knowledge of the correct answer. Bottom-up derivation is truly impossible. This condition can be made even stronger by adding inherently sequential computations into the Halting Problem program, assuming $NC \neq P$ (see Appendix A.)

An appropriately selected specific Halting Problem cannot be solved from the bottom up within our finite universe; there is an uncrossable chasm. An NP-complete problem, on the other hand, *can* be solved by some bottom-up pathways within our universe; it is just unlikely that these pathways will be fortuitously stumbled upon (on account of their rarity). There is no uncrossable chasm here. (Note that once the answer to an NP-complete problem is found, the bottom-up path to get there becomes obvious.) Furthermore, whereas a guess and check strategy can work for NP-complete problems, it cannot work for the Halting Problem or for interactive

proofs in general. To illustrate, NP-complete problems can in principle be solved using a shotgun approach—keep guessing random but possible answers and checking until an answer is found that checks out correctly. However, the shotgun approach would not result in a valid interactive proof, whose "proof" is based on the statistically derived improbability of getting a series of correct answers by chance alone.

Appendix C

Additional case studies of information that might be discoverable even if it is not derivable in a bottom-up fashion

Note that here we present brief sketches of several potential truths that might be discoverable even if they cannot be derived using a bottom-up approach. We leave more in-depth discussion of these case studies as future work.

Case study #2: The Hard Problem of Consciousness

The second case study we will discuss (briefly) is a famous historical example that may point to strong emergence—the existence of experiential (or phenomenological), qualia-based aspects of consciousness (what it is like to be/feel a certain way). David Chalmers (1995) proposed dividing the questions of consciousness and of mind into two categories: the "easy problems" and "The Hard Problem". The Hard Problem can be paraphrased as, "Why do some organisms have a subjective, experiential sensation (or awareness) that accompanies the information they are processing?"

Using lines of reasoning such as the "Philosophical Zombies" argument (Chalmers, 1997), the "Inverted Qualia" argument (Block, 1990), or "Mary's Room" argument (Jackson, 1982), many have argued that to directly test the existence of another person's experience of qualia is intrinsically impossible. Chalmers has argued that even if we could measure and map every single fundamental physical particle and all of their associated interactions in a person's brain, we would still not be any closer to answering the Hard Problem of Consciousness. We would still be unable to determine whether another individual experiences qualia and, if so, the character of that experience. Measuring a person's brainwaves, electrical impulses, chemical and neurological structures, and behavior won't help because the problem is fundamentally intractable to this type of approach. A vast body of literature has explored the Hard Problem of Consciousness and its implications for strong emergence (cf. Block, 2002; Levine, 2009; McGinn, 2012). Accordingly, we will forego an in-depth discussion here.

Nonetheless, we see parallels between the Hard Problem of Consciousness and the Halting Problem. If Chalmers is correct, the question of whether or not a specific individual has qualia-derived experiences, and whether a specific program will halt on a specific computer both have definitive answers within our closed system, and yet those answers cannot be derived from the bottom up. Moreover, in each case the physical circumstances can be fully specified without providing an answer or a route to an answer. In the case of the Halting Problem, we can fully specify the designs and working of both the computer program and the computer (without solving the

problem). In the Hard Problem of Consciousness, we might fully specify all of the fundamental particles and their locations in the individual's brain (without solving the problem). Additionally, in each case we can fully enumerate the possible answer space (the answer must be either "yes" or "no" when asking whether an individual does experience qualia and when asking whether a specific program will halt.)

Moreover, in both cases the answer is hypothetically knowable within the closed system. In the case of the Halting Problem, knowledge of the answer can be obtained through an interactive proof between the provers and the verifier. For the Hard Problem of Consciousness, at least one individual (the subject in question) can know whether or not they experience qualia. In fact, we might even choose to view the process by which an individual gains awareness of their own experiences of qualia as a subtle interactive proof wherein that individual tests themselves by effectively playing the roles of both the prover and the verifier. Like a photon, experience of qualia seems to be born fully self-evident/compelling if the appropriate introspective self-tests are conducted. Regardless of whether or not consciousness is epiphenomenal, the Hard Problem of Consciousness suggests that a property can be discoverable, even if it is not derivable from the bottom up.

Case study #3: Many-Worlds Interpretation of quantum mechanics

We present a third case study highlighting a potential truth that might be reasonably discoverable even if it is not derivable from the bottom up. While quantum mechanics is generally considered to be the most rigorously tested theory in all of science, the interpretation of what quantum mechanics implies about reality remains contentious. Over a dozen distinct interpretations of quantum mechanics have been put forward. However, among these interpretations, the two most popular (the Copenhagen Interpretation (Stapp, 1972) and the Many Worlds Interpretation (Everett, 2015)) are generally believed to be indistinguishable to observers within our closed system using bottom-up experimentation or derivation. It remains controversial as to whether the differences between these interpretations is indistinguishable in practice or in principle (using an empirical bottom-up approach).[45] Nonetheless, it is therefore possible that if either the Copenhagen Interpretation or the Many Worlds Interpretation is correct, the truth of the matter could not be derived from the bottom up (from within the system).

However, Max Tegmark (1998) has formalized a possible test by which to falsify interpretations of quantum mechanics that do not incorporate many worlds. (Note that the viability of this test remains controversial (Gao, 2022)). This test, known as

45 Note that for the purposes of this thought experiment it is not necessary to delve into the physics controversy of whether or not the Copenhagen Interpretation and the Many Worlds Interpretation are distinguishable by experimental methods from the bottom up within *our* universe. Instead, what *is* important to notice is that it is metaphysically plausible for a universe to exist in which the Copenhagen Interpretation and the Many Worlds Interpretation are indistinguishable from a bottom-up approach but *can* be distinguished by the quantum suicide experiment.

"quantum suicide," necessitates that the experimenter take the place of Schrödinger's cat. By climbing into a box that will rapidly kill the experimenter with a high probability (mediated by a quantum event), the experimenter is all but assured death under the Copenhagen Interpretation of quantum mechanics. However, if the Many Worlds Interpretation of quantum mechanics is in fact correct, then there will exist universes (rare though they may be) in which the experimenter survives. The Many Worlds Interpretation alleges that all universes (branches of the wavefunction) are equally real. However, the experimenter cannot be aware of branches of the wavefunction in which he is dead. Therefore, by the anthropic principle the experimenter will find himself (by necessity) in one of the branches of the wavefunction where he fortuitously (and improbably) survived. At this point, the experimenter will be reasonably convinced (to within an arbitrarily high certainty) that the Many Worlds Interpretation of quantum mechanics is correct. (This credence follows from the improbability of the experimenter having survived under the singular outcome of the Copenhagen Interpretation, in contrast to the alleged certainty of survival under the Many Worlds Interpretation.) Note that this credence can be further strengthened by reducing the experimenter's chance of survival in any randomly selected branch of the wavefunction. Unfortunately, like in the case of experiencing qualia, no one apart from the experimenter will be convinced by the evidence presented by the experimenter who risked his or her life. The evidence is only compelling from the first-person perspective (to someone who has gone inside the box and survived). From a third-person perspective, there is no mechanism for a self-selection bias that would downselect specifically for universes in which the experimenter survived (i.e., the anthropic principle). Therefore, from a third-person perspective, the experimenter, if he or she survived, should be considered extremely lucky, but nothing more can be gleaned.

In a sense, the experimenter can be thought of as verifying the truth of the Many Worlds Interpretation by an interactive proof (iterated with his or her own life in various branches of the wavefunction).[46] Implicitly, the multitude of branches under the Many Worlds Interpretation serves as the prover and the surviving experimentalist acts as a verifier. Moreover, this interactive proof can only compellingly convey knowledge (as to the truth of the Many Worlds Interpretation) to an entity with a sufficient minimum level of complexity (e.g., consciousness, sematic understanding). This scenario therefore closely parallels the verification of the answer to the Halting Problem (by an interactive proof). While there is possibly no bottom-up path by which to decide between the Copenhagen Interpretation and the Many Worlds Interpretation (though this remains controversial), there may exist an emergent pathway to gain this information/knowledge (via the quantum suicide experiment). If the Many Worlds Interpretation is correct, this truth might be discoverable, even if it is not derivable from the bottom up.

46 Note that as with other "interactive proofs" these verification methods do not provide absolute, infallible proof of a conclusion. Rather they provide compelling evidence (given assumptions) in support of gained knowledge.

Case study #4: Multiverse

We present a fourth (and original) case study highlighting a potential truth that might be discoverable even if it is not derivable from the bottom up. This case study provides a means for falsifying the hypothesis of the existence of a vast level II multiverse—composed of near infinite parallel universes, each with different values for fundamental constants and initial conditions apart from our own (Tegmark, 2003). This case study makes use of anthropic reasoning and makes use of the following assumptions. (Note that this is a thought experiment and not intended to reflect the most recent advancements in cosmology and theoretical physics.)

First, we assume that our universe is fine-tuned for complex, chemistry-based intelligence, as is a popular opinion among cosmologists on the basis of observations and modeling (Rees, 2008). This is to say that the constants and initial conditions of our universe take on very precise values (from within the presumed range of all possible values) that happen to be permitting to the existence of complex, chemistry-based intelligence. If these constants and initial conditions were even a hair's breadth different, then the universe would either explode or implode (or otherwise be altered) so fast that the formation of stars and the assembly of complex, chemistry-based intelligence would be impossible anywhere in the universe. Changing these constants or initial conditions even slightly would preclude all complex, chemistry-based intelligence in our universe. Let us call complex, chemistry-based intelligence, "Type A Intelligence". Second, we assume that a physical theory called "Theory S" predicts the existence of near infinite[47] parallel universes (in a vast multiverse that contains our own universe). Furthermore, Theory S predicts a very large (but finite) range (and probability distribution) of possible values for constants and initial conditions with which these near infinite parallel universes could be instantiated.

Additionally, we assume that we can roughly simulate (on a supercomputer) each possible universe in order to get the gist of what it could or could not contain, given its set of constant and initial condition values (e.g., could that universe contain complex, chemistry-based intelligence). Granted, completing these numerous simulations would require immense computational power, but this is only a thought experiment. Finally, let us define a form of omni-survivable intelligence that would come into existence, survive, and thrive (i.e., multiply) in almost any universe allowed for by Theory S (with high probability). Let us call this intelligence form "Type B Intelligence". (Note that the values of the constants and initial conditions that would support "Type A Intelligence" are a small subset of the full range of constant and initial condition values allowed for by Theory S.)[48] It is presumed that under Theory

47 A very large number that is nonetheless finite.

48 Note that we can posit other types of intelligence, C, D, E…, each of which is fine-tuned for its own universe in different ways (not necessarily relying on the existence of complex chemistry). However, as long as the vast majority of universes predicted to exist in the multiverse (by Theory S) are barren of any form of intelligence, save for Type B Intelligence, then the argument remains unaffected.

S, distinct universes within the multiverse are instantiated randomly, with randomly selected values for their constants and initial conditions (given the ranges and probability distributions entailed by Theory S). It is further presumed that under Theory S almost all universes in the multiverse are not capable of supporting any form of intelligence, save for "Type B Intelligence."

Under these assumptions, if a vast level II multiverse exists (owing to Theory S), then throughout the multiverse as a whole, cases of Type B Intelligence will vastly outnumber cases of Type A Intelligence (or any other intelligence that can only exist in highly fine-tuned universes). This follows from the fact that Type B Intelligence can and is expected to exist in abundance in almost all of the universes in the multiverse. In contrast, Type A Intelligence and other types of intelligence that can only flourish in highly fine-tuned universes can only exist in a small fraction of the total universes in the multiverse.[49] Therefore, if one were to randomly sample from the total set of intelligent observers in the multiverse, one would expect to draw an observer of Type B Intelligence with much higher probability that any other Type of Intelligence. One way in which to sample from the total pool of intelligent observers in the multiverse (in a presumably unbiased and random fashion) is to consider oneself.

An individual, e.g., Jane, could awaken one day and realize that she is an intelligent observer, and presumably a randomly sampled observer in the absence of any pressure to have awoken as "Jane" as opposed to any other intelligent observer. As an observer she can make measurements on the values of the constants and initial conditions of her universe and furthermore determine that her type of intelligence could only exist within a very small subset of the total parameter space predicted by Theory S. Jane can thus discover that she is of Type A Intelligence. She can use these realizations to put to the test the hypothesis that there exist a vast number of universes (in a level II multiverse) beyond her own universe (randomly instantiated within the predicted parameter-space distribution of Theory S).

Multiverse Test. Imagine that Jane would like to rule out the existence of a near infinite level II multiverse (randomly instantiated by Theory S). She could simply run computer simulations to search for Type B Intelligence, given the ensemble of universes allowed for by Theory S. If she discovers Type B Intelligence (and can through simulation validate the other aforementioned assumptions), then she has statistically precluded the existence of a near infinite level II multiverse, randomly instantiated by Theory S. (Here we assume that Jane is aware that she is of Type A Intelligence.)

This conclusion follows from anthropic reasoning, codified in the following deductive argument:

---

49 It is also important to consider the quantity of individual intelligent observers that would be expected to exist in each universe (sorted by their Intelligence Type). This could be also addressed by computer simulations.

(1) If there are near infinite parallel universes randomly instantiated by Theory S, and Type B Intelligence exists (given Theory S), then (by statistics) Jane will not find that her universe is fine-tuned for her type of intelligence (Type A Intelligence).
(2) Jane finds that her universe is fine-tuned for Type A Intelligence.
(3) Therefore, (by statistics) there do not exist both near infinite parallel universes randomly instantiated by Theory S, and Type B Intelligence (given Theory S).
(4) Type B Intelligence (given Theory S) does exist (per Jane's computer simulations).
(5) Therefore, near infinite parallel universes (randomly instantiated by Theory S) do not exist.

In this interactive test, Jane discovers that near infinite parallel universes (randomly instantiated by Theory S) are very unlikely to exist. Put another way, Jane's universe is very likely not surrounded by near infinite parallel universes (randomly instantiated by Theory S). This is a truth that Jane can potentially learn about her own universe (gained information/knowledge), despite the fact that there likely does not exist any way for Jane to derive this truth from the bottom up (from within her own universe). Therefore, this provides another case study of a truth that might be discoverable even if it is not derivable from the bottom up (within the closed system). (Once again, the truth does not need to be known with a credence of 1, but only an increased credence conditioned on the interactive test—i.e., gained knowledge.)

Note that two important pieces of this interactive test are that Jane knows herself to be an observer of Type A Intelligence and that Jane knows herself to be a randomly selected sample of an intelligent observer. Again, this argument would not be compelling to others as it requires private (first person) information (pertaining to Jane's status as a randomly selected observer among the set of all possible observers throughout the multiverse). The argument only works because Jane can assume *herself* to be a randomly selected element from the whole set of observers. If Jane were to select any other individual as the representative intelligent observer, the selection would presumably be based on limited information (accessible to Jane) that would create a selection bias and invalidate any result.

Appendix D

Thought experiment in the search for strong emergence

Historically, a lack of rigorous specificity in the definition of strong emergence and the criteria that are sufficient for it have muddled the concept and made it difficult to seek out or study. In this work we provide a sufficient set of criteria for strong emergence and a deductive argument based on these six criteria. This new conceptual framework for understanding strong emergence suggests new ways to search for strong emergence in our universe. An example thought experiment is provided below.

Let us imagine that a hypothetical scientist, Sally, is determined to design and build a fleet of robots that will search for strong emergence. (Sally plays an important role as she possesses semantic understanding of the experimental setup and results.) The robots function as follows. Each robot has a computer, an interface for communicating with humans, an internal atomic clock, encrypted radio signal-based communication channels, a quantum communication channel, and an internal quantum random number generator. The robots are mobile, spread themselves out across the planet, and randomly approach different individuals living across the planet. The robots have the purpose of searching for strong emergence by seeking out groups of individuals with the property "O". As defined in the main text, this property entails the hypothetical, emergent ability of an individual to serve as an oracle for the Halting Problem (i.e., the individual intuitively knows the answer to the Halting Problem without having to derive it from the bottom up). It is likely that the property "O" does not actually exist within our universe, but that does not pose a hindrance to this thought experiment. The deeper question at hand is whether it is *impossible* for the property "O" to exist within our universe, and if so, why?

The robots act as verifiers to the Halting Problem. The randomly selected humans ("contestants" from around the world) who communicate with the robots act as Bob, providing oracle-like information regarding the Halting Problem. In order to have multiple independent provers, as required by "MIP* = RE" (Ji et al., 2021), the robots will randomly select and assemble simultaneous groups of individuals (contestants) at distant geographic locations around the globe. For example, the robots might approach and engage with random individuals found in public spaces around the world (e.g., Times Square in New York, Shibuya in Tokyo, and Piccadilly Circus in London). Separating the simultaneous contestants out geographically helps to ensure that they are acting independently. In theory this independence can be further ensured by timing the contestants' responses using the robots' internal atomic clocks and only accepting responses within specified timeframes (that preclude collusion amongst the contestants, assuming no faster-than-light communication between the contestants). In practice, this anti-cheating measure would be difficult to implement, and alternative anti-cheating measures could be conceived of.

The robots communicate with each other via radio signals at the speed of light. Also, the humans are allowed to access and make use of the quantum communication channels embedded in each robot. In this way the multiple independent humans serving as Bob cannot communicate directly with each other but can make use of shared quantum information (as required by "MIP* = RE"; Ji et al., 2021).

First the robots will collectively generate a random number using their quantum random number generators. This random number will serve as a seed number (y_n) in a Modified Goldbach's conjecture Halting Problem (see Appendix A). The contestants will be asked to take part in an interactive proof to verify whether this particular Halting Problem halts by playing a nonlocal game (using the shared quantum information). If the contestants have the property "O", they will be able to win the game and, in the process, provide input data to the robots verifying that the particular Halting Problem does halt (when true). Otherwise, the inputs provided by the contestants will be inconclusive. If the contestants win a particular game (and thereby "prove"[50] that a particular Halting Problem does halt), then the robots will unfurl robotic butterfly wings and flap them three times as a congratulatory celebration. This also has the effect of converting a largely informational game into a physical disturbance in the world with chaotic downstream physical consequences (e.g., by the butterfly effect).

The robots will continue to approach new contestants and play the game over and over (each time starting with a different, randomly selected seed number). If the contestants win the game for a particular Halting Problem (and thereby prove that the particular Halting Problem does halt), two possibilities follow: 1) the program can be shown to halt within the computational bounds of our universe, or 2) the computational resource necessary to derive the solution to this Halting Problem is greater than the computational bounds of our universe. Only possibility #2 unambiguously demonstrates strong emergence as delineated in this work. Moreover, in general it will not be clear if a particular Halting Problem falls within possibility #1 or possibility #2. However, as the robots continuously play this game with different, randomly selected seed numbers, they are likely to stumble upon Halting Problems that fall within possibility #2 (even if we are unable to determine which cases do).[51]

In this thought experiment we have laid out a hypothetical method by which strong emergence might be sought out (and plausibly stumbled upon). Indeed, we find it unlikely that strong emergence of this particular form (humans possessing the property "O") does in fact exist within our universe, but that is not the point.

---

50 As previously discussed, in "interactive proofs" the term "proof" is used loosely. These verification methods do not provide absolute, infallible proof of a conclusion. Rather they provide compelling evidence (given assumptions) in support of gained knowledge.

51 Note that even if the particular halting problem falls within possibility #1, it might be possible to determine constraints on the computational power available to the contestants (e.g., based on the total mass and volume over which they can exert control). These constraints will presumably limit the contestants to a smaller computational resource than that of the entire universe. Therefore, it might still be possible to show that computing the answer to the particular halting problem from the bottom up is impossible for the contestants (within the allotted time and computational resource constraints.)

The point is that in order to argue for strict reductionism (i.e., that strong emergence does not exist), one must specifically argue why it would be impossible (or at least unreasonable) for the aforementioned thought experiment to be able to stumble upon strong emergence.[52] This thought experiment therefore helps to clarify where such strict reductionist arguments might focus and whether they are likely to succeed. Additionally, this thought experiment demonstrates that one might have an encounter with strong emergence without having the ability to discern that it is strong emergence, in support of Section 8. (Note that this thought experiment assumes the same digital physics backdrop as the main text.)

In order to argue for strict reductionism, one must argue that the aforementioned thought experiment must necessarily fail to encounter strong emergence. Below we go through different possible hindrances to the aforementioned thought experiment ("arguments" as to why it should fail) and reasons why each is lacking as an argument for strict reductionism ("counterarguments"). The conclusion is that the arguments for strict reductionism are weak, leaving open the plausibility of strong emergence.

Possible Arguments for Strict Reductionism

(Possible hindrances wherein the aforementioned thought experiment is alleged to be destined to fail)

Argument 1. Robots of this functionality could not be designed and built.

Counterargument: The sophistication of the robots described in this thought experiment is not significantly beyond the capabilities of current technology. Importantly, there is no presumption that these robots are conscious or have computational abilities exceeding that of today's computers. It would be difficult to argue that robots with this functionality could not be designed and built.

Argument 2. The robots could not convert informational states into a causal, physical disturbances.

Counterargument: In cases of the mind controlling the body there exist controversies about the plausibility of downward causation. However, in this thought experiment there is no presumption that the robots are conscious, have minds, or free wills. The robots are programmed by Sally (a conscious scientist with sematic understanding) to mechanistically take in inputs and produce outputs. Direct inputs

52 Note that here we only use the property "O" and its associated interactive proof as an arbitrarily selected example of strong emergence. However, the prescribed search process can be generalized to other cases of strong emergence.

include the information provided by the humans (through the communication interfaces) and communications received from other robots in the fleet. Indirect inputs of information include the programming of the robot and the robot's physical constitution (both designed and built by Sally). Given these various inputs, each robot will act mechanistically to either flap its wings or not (in accordance with its design). At an abstract level, if a specific instantiation of the Halting Problem is verified to halt, then the robot's subsequent actions will also be codetermined by "ethereal information" (see Section 1 of the main text). This thought experiment sidesteps the controversial concept of downward causation (i.e., a mind controlling a body) in the robots. Therefore, it would be difficult to argue that the robots could not be built to function as described and would not actually function as described (processing on informational inputs to produce distinct physical actions or physical disturbances in their environments).

Argument 3. Specific halting problems that halt, but whose answer cannot be derived in a bottom-up fashion within our computationally bounded universe cannot exist or cannot be stumbled upon.

Counterargument: We have detailed the likely existence of such Specific Halting Problems in Appendix A.

Argument 4. Answers to the specific halting problems of Argument 3 cannot be verified (knowable) within our computationally bounded universe.

Counterargument: This argument is strongly refuted by the cited work, "MIP* = RE" (Ji et al., 2021).

Argument 5. The oracle-like information necessary to verify answers to the specific Halting Problems of Argument 3 cannot be discovered and communicated to the robots.

Counterargument: Argument 5 is likely the most interesting of the (strict reductionism) arguments proposed here. One might initially claim that information cannot be discoverable if it is not derivable from the bottom up. However, in case studies #1-#4 we have presented several examples to refute this claim. Admittedly, all of these case studies involve a conscious being and private information discovered within the mind of the conscious being that relates to the direct experiences of the conscious being. We say that the information is private because the conscious mind within which that information is discovered has (seemingly) privileged access to the information above and beyond what an outside observer would easily be able to access.

The private nature of the information discovered in case studies #1-#4 raises the question as to whether certain types of information are discoverable but not derivable, while other types of information are not. This raises the possibility that perhaps information that is discoverable but not derivable cannot be publicly and persuasively communicated to others. Perhaps, for example the information necessary to verify answers to the Halting Problem is discoverable but cannot be communicated to the robots. For example, the information might be epiphenomenal[53].

However, if this constraint were true, it would raise other thorny questions for strict reductionism within a digital physics backdrop. We might ask whether strict reductionism within a monist digital physics backdrop has room for epiphenomena. Would the information of this epiphenomena take a digital form (a seeming requirement of a monist digital physics backdrop)? If so, where would this digital information be stored or computed? Would this epiphenomenal data not be a part of the digital physics universe, and if so, would adding or altering the epiphenomenal data not also constitute a physical change within some part of the digital physics universe? Perhaps more conspicuously, does strict reductionism within a monist digital physics backdrop allow for private information? In this system, wouldn't everything be equally privileged and equally a part of the same digital/physical substrate? In this system, what would delineate the boundary between public and private, between a specific mind and the rest of everything?

53 I.e. caused by physical states but without causal power on the physical

References

Anjum, R. L., & Mumford, S. (2017). Emergence and demergence. In *Philosophical and Scientific Perspectives on Downward Causation*. Routledge.

Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives*, 11, 375-399.

Bekenstein, J. D. (1981). Universal upper bound on the entropy-to-energy ratio for bounded systems. *Physical Review D, 23*(2), 287. https://doi.org/10.1103/PhysRevD.23.287

Beraldo-de-Araújo, A., & Baravalle, L. (2017). The ontology of digital physics. *Erkenntnis*, 82, 1211-1231. https://doi.org/10.1007/s10670-016-9866-y

Block, N. (1990). Inverted earth. *Philosophical Perspectives*, *4*, 53-79. https://doi.org/10.2307/2214187

Block, N. (2002). The harder problem of consciousness. *The Journal of Philosophy*, *99*(8), 391-425. https://doi.org/10.2307/3655621

Bostrom, N. (2003). Are we living in a computer simulation?. *The Philosophical Quarterly*, *53*(211), 243-255. https://doi.org/10.1111/1467-9213.00309

Bremermann, H. J. (1962). Optimization through evolution and recombination. *Self-Organizing Systems*, *93*, 106.

Carroll, S. M. (2020). Why Boltzmann brains are bad. In *Current Controversies in Philosophy of Science* (pp. 7-20). Routledge.

Chaitin, G. J. (1975). A theory of program size formally identical to information theory. *Journal of the ACM (JACM)*, *22*(3), 329-340.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200-219.

Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.

Chalmers, D. J. (2006). Strong and weak emergence. In *Clayton, P. & Davies, P. The Re-emergence of Emergence*. Oxford University Press.

Collier, J. (2011). Holism and emergence: Dynamical complexity defeats Laplace's Demon. *South African Journal of Philosophy*, *30*(2), 229-243. https://doi.org/10.4314/sajpem.v30i2.67786

Davenport, M., & Olum, K. D. (2010). Are there Boltzmann brains in the vacuum. *arXiv preprint arXiv:1008.0808*. https://doi.org/10.48550/arXiv.1008.0808

Davies, P. C. (2004). Emergent biological principles and the computational properties of the universe. *arXiv preprint astro-ph/0408014*. https://doi.org/10.48550/arXiv.astro-ph/0408014

El-Hani, C. N., & Pereira, A. M. (2000). Higher-level descriptions: why should we preserve them. *Downward Causation: Minds, Bodies and Matter*, *133*, 118-42.

Everett, H. (2015). The theory of the universal wave function. In *The Many-Worlds Interpretation of Quantum Mechanics* (pp. 1-140). Princeton University Press.

Feinberg, G. (1967). Possibility of faster-than-light particles. *Physical Review*, *159*(5), 1089. https://doi.org/10.1103/PhysRev.159.1089

Fredkin, E. (2003). An introduction to digital philosophy. *International Journal of Theoretical Physics*, *42*(2), 189-247. https://doi.org/10.1023/A:1024443232206

Gao, S. (2022). Quantum suicide and many worlds. http://philsci-archive.pitt.edu/20926/1/qs%202022.pdf

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, *38*(1), 173-198. https://doi.org/10.1007/BF01700692

Gompert, Z., Flaxman, S. M., Feder, J. L., Chevin, L. M., & Nosil, P. (2022). Laplace's demon in biology: Models of evolutionary prediction. *Evolution*. https://doi.org/10.1111/evo.14628

Hao, Z., Liu, J., Wu, B., Yu, M., & Wegner, L. H. (2021). Strong Emergence in Biological Systems: Is It Open to Mathematical Reasoning?. *Acta Biotheoretica*, 69(4), 841-856. https://doi.org/10.1007/s10441-021-09423-1

Hartnett, K. (2020). *Landmark Computer Science Proof Cascades Through Physics and Math*. Quantamagazine. https://www.quantamagazine.org/landmark-computer-science-proof-cascades-through-physics-and-math-20200304/

Holland, J. 2014. Complexity: A Very Short Introduction. Oxford University Press.

Humphreys, P. (2016). Emergence: A philosophical account. Oxford University Press.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly*, *32*(127), 127-36.

Ji, Z., Natarajan, A., Vidick, T., Wright, J., & Yuen, H. (2021). MIP*= RE. *Communications of the ACM*, *64*(11), 131-138. https://doi.org/10.1145/3485628

Kim, J. (2006). Emergence: Core ideas and issues. *Synthese*, 151(3), 547-559. https://doi.org/10.1007/s11229-006-9025-0

Kotzen, M. (2020). What Follows from the Possibility of Boltzmann Brains?. In *Current Controversies in Philosophy of Science* (pp. 21-34). Routledge.

Levine, J. (2009). The explanatory gap. *The Oxford Handbook of Philosophy of Mind*, 281-291.

Lloyd, S. (2007). *Programming the universe: a quantum computer scientist takes on the cosmos*. Vintage.

McGinn, C. (2012). All machine and no ghost. *New Statesman*, *20*(02).

Meinong, A. (1910). *Über Annahmen* (Vol. 2). Lipport.

Mill, J. S. 1843. A System of Logic. In: J. Robson (ed.) 1973, The Collected Works of John
Stuart Mill, vol. 7-8. Indianapolis: Liberty Fund.

Oliveira e Silva, T., Herzog, S., & Pardi, S. (2014). Empirical verification of the even Goldbach conjecture and computation of prime gaps up to $4 \cdot 10^{18}$. *Mathematics of Computation*, *83*(288), 2033-2060.

Planck, M. (1899). Uber irreversible Strahlungsvorgänge. *Sitz. König. Preuss. Akad. Wissen*, *26*, 440. https://doi.org/10.1007/978-3-663-13885-3_13

Rees, M. (2008). *Just six numbers: The deep forces that shape the universe*. Hachette UK.

Stapp, H. P. (1972). The Copenhagen Interpretation. *American Journal of Physics*, *40*(8), 1098-1116. https://doi.org/10.1119/1.1986768

Tegmark, M. (1998). The interpretation of quantum mechanics: Many worlds or many words?. *Fortschritte der Physik: Progress of Physics*, *46*(6-8), 855-862.

https://doi.org/10.1002/(SICI)1521-3978(199811)46:6/8<855::AID-PROP855>3.0.CO;2-Q

Tegmark, M. (2003). Parallel universes. *Scientific American*, *288*(5), 40-51.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Journal of Math*, *58*(345-363), 5.

Wang, Y. (2002). *The Goldbach Conjecture* (Vol. 4). World scientific.

Whitworth, B. (2008). The physical world as a virtual reality. *arXiv preprint arXiv:0801.0337*. https://doi.org/10.48550/arXiv.0801.0337

Wolfram, S. (1985). Undecidability and Intractability. *Physical Review Letters*, 54(8), 735. https://doi.org/10.1103/PhysRevLett.54.735

Yuen, H. (2020). *The Shape of MIP* = RE*. Quantum Frontiers. https://quantumfrontiers.com/2020/03/01/the-shape-of-mip-re/

Zuse, K. (1969). Rechnender Raum (calculating space). *Schriften Zur Dataver-arbeitung*, *1*.

Acknowledgements

We are indebted to Henry Yuen and Aldo Filomeno for insightful comments on this work. Also, Esteban Céspedes would like to thank Miguel Ángel Fuentes for fruitful discussions on issues related to the main topic of the paper, as well as to the Chilean Agency for Research and Development (FONDECYT projects #1211323 and #1241630) and to the Spanish State Research Agency (PID2023-150396OA-I00).

Ethical statement

Funding: This work was supported by the Chilean Agency for Research and Development (FONDECYT projects # 1211323 and # 1241630).

Conflict of Interest: The authors have no relevant financial or non-financial competing interests to disclose.

Ethical approval: Ethical approval is not needed.

Informed consent: All authors read and approved the final manuscript.

Author contribution: All authors contributed to the conception and writing of this manuscript.

Data availability: The data that support the findings of this study are openly available.

Affiliation: This work was not funded or commissioned by Lawrence Livermore National Laboratory or the U.S. Department of Energy. The views expressed herein do not necessarily represent the views of Lawrence Livermore National Laboratory or the U.S. Department of Energy.