24. Can Consciousness Be Explained by Integrated Information Theory or the Theory of Cognitive Consciousness?[1]

Selmer Bringsjord and Naveen Sundar Govindarajulu

1. Introduction

As readers will doubtless have noted by now, some other chapters in the present volume have expressed the view (rather agreeable to us) that many aspects of human-level mental phenomena are recalcitrant to a mindset that insists upon mathematical and (usually) material explanations. First-person subjectivity, intentionality, mathematical cognition, robust epistemic states, consciousness. . .these phenomena are exceedingly hard to explain in such a manner. It is the final member of that list of challenges that is our focus in the present chapter. Can science operating in the math-and-material manner explain—and perhaps even, courtesy of associated engineering, replicate in artificial agents—consciousness?

This question is now pressed upon at least all technologized societies on Earth, because of the advent of artificial agents able to converse in seemingly flawless English about pretty much anything, including consciousness itself. A famous example is ChatGPT. This class of agents falls into what is now called "generative AI," which includes agents not only able to generate natural language, but also images. In the case of language, these agents are sometimes called "chatbots," but are more precisely known as "Large Language Models." Some of these agents have been declared conscious,[2] and the question of whether they are is really just a special case of the general question taken up in the present chapter. We are very confident that ascriptions of consciousness to artificial agents are only going to grow in frequency, and such ascriptions are going to increasingly be issued by voices that seem balanced and authoritative. This chapter should in our opinion be read and understood by those humans who will find themselves living in the trend we foresee, because it provides at least a starting basis for two fundamental ways of looking not just at consciousness in general, but consciousness in computational artifacts.

It's absolutely crucial for the reader to fully appreciate at the outset that the term "consciousness," even when used in scholarly and scientific contexts, is dizzyingly protean. This state of affairs may be profoundly disturbing to some, but there is no getting around its truth. One of us, Bringsjord, has been actively engaged in discussion and debate about "consciousness" for four decades, within the fields of philosophy, cognitive science, computer science, AI, and computational logic. For Govindarajulu, the timespan is shorter, but the experience is the same. The one indubitable takeaway from both of our cases is this: scholars, scientists, and engineers don't agree about what "consciousness" is.

In the present chapter, we analyze and discuss the two main overarching senses of the term. We present two attempts to explain and measure consciousness in purely mathematical or—to use the better adjective—*formal* terms.[3] Each of these two attempts includes a theory of consciousness, and each theory has its formal framework for measuring the level of consciousness in a given system or agent.

The first type of consciousness we consider is the brand you feel when, for instance, you sample a world-class wine of great complexity; realize that yes, you truly love someone; feel the intense mixture of fear and fun arising from skiing very fast; have a eureka thrill upon finally cracking some puzzle (and so on). Here we are talking about—to use the phrase that tends to dominate at present among philosophers of mind—*phenomenal* consciousness.[4] This is the kind of consciousness that *integrated information theory* (IIT) seeks to explain, in both formal and material terms. IIT's formal measurement scheme is Φ; and the target here is specifically *phenomenal consciousness* ("what-it's-like" consciousness).

Another established sense of consciousness pertains solely to whether internal structure in the relevant agent enables reasoning through time over content encoded in formal languages. This kind of consciousness is known as *cognitive consciousness*. Cognitive consciousness is not, like phenomenal consciousness, "what-it-feels-like" consciousness, at all. Rather, as we shall further explain, it stems directly from the "non-feel" side of a fundamental dichotomy in types of consciousness advanced by John McCarthy, and also from what is called *access* consciousness by Ned Block in a landmark paper of nearly three decades ago.[5] This is the focus of the second theory, the *theory of cognitive consciousness* (TCC); its formal measurement scheme is Λ; and the target is so-called *cognitive consciousness*.[6]

Whereas the IIT/Φ pair emerge out of and are at home in computational neuroscience, the TCC/Λ pair emerge out of and are currently at home in logic-based AI, which—as we shall explain in due course—is currently producing artificial agents having appreciable cognitive consciousness. IIT/Φ originates in work of Tononi and colleagues[7] and is principally defended and moved forward these days by Tononi and Christof Koch.[8] IIT starts from a set of five "axioms," which are rather important to review. For example, whereas some have sought to refute IIT/Φ by pointing out that this pair entails a denial of the unity of consciousness in a given agent, Tononi and Koch specifically take as one of their starting axioms a proposition that insists upon unity of consciousness as fundamental and—for the enterprise of erecting a computational science of phenomenal consciousness—undeniable.[9]

TCC, on the other hand, has its roots unmistakably in the work of John McCarthy,[10] one of the founders of AI, who ascribed consciousness to artificial agents on the strength of

their suitably reasoning in one or more logics over content represented in the formal languages of those logics. TCC has been expanded, refined, axiomatized, and given the Λ measurement scheme, in for example our own work.[11] Bringsjord benefitted from direct conversations with the late McCarthy on these matters.[12]

1.1 Plan for the Remainder

The present chapter unfolds as follows. Next, in §2, we characterize, in general terms, what we take scientific explanation to be, and apply these terms to the specific challenge of scientifically explaining consciousness. Then, in §3, we provide a fuller exposition of IIT/Φ. We first make clear that it conforms neatly to at least the explanatory *structure* needed for an explanation of consciousness, in no small part because IIT is based upon five axioms. We examine these axioms, and then, by making use of two illustrative household robots, High and Low (which allow us to prevent the reader from getting lost in the rather dense mathematics of Φ), convey the core mathematical ideas underlying IIT. (We provide in the appendix a more technical overview of IIT, for interested readers.)

We next (§4) discuss John Searle's sustained and lively case against IIT/Φ. We specifically examine first his argument that *any* informational theory of consciousness (not just IIT) is doomed, because information is "observer-relative" [e.g., the parenthetical you are now reading carries information only insofar as there is a reader of it (= you) with some command of English], whereas consciousness isn't (that there's something it's like to be you, and that you have this thing, are not observer-relative at all). We next look at a second Searlean argument in favor of specifically rejecting IIT, namely that since IIT entails panpsychism, IIT is false.

Next, in §5, we present and defend a better way to explain consciousness, one that can be pursued in concrete fashion: namely, focus on the aforementioned concept of *cognitive* consciousness, and on a system for measuring its level in objects (Λ, instead of Φ), and on the engineering of artifacts that have high Λ, and as such are AIs of much promise. We then wrap up with some concluding remarks (§6), after which follows our appendix (§7) for exceptionally sedulous readers.

2. The Scientific Challenge of Consciousness

As promised, we now summarize the challenge in question, and to do so begin by taking note of the chief aim of science, broadly understood.

2.1 The Overarching Aim of Science

We take it as a given that the chief aim of science is to explain phenomena. Hence, a bit more precisely, the purpose of science for us[13] is to gradually supply, for a given phenomenon *p* for which there is observational data,[14] some third-person content that serves to explain *p*. By "third-person content" we mean that the content is expressed in some medium by which human persons in different cultures and using different natural languages (English, Chinese, Norwegian, Russian, etc.) can nonetheless all understand that content. The medium used since the dawn of science to render relevant content third-person-understandable is specifically formal logic and mathematics; and ultimately use of the formal languages in which content in these disciplines is expressed is *declarative*, and hence the content is a collection of propositions (= statements).[15] To anticipate what is to come below, this is why scientists can refer correctly and helpfully to such things as the "axioms of Newtonian mechanics." An axiom is of course a proposition/statement which usually is regarded

to be of a fundamental nature relative to the domain of inquiry in question.

The situation we have just summarized can be put a bit more precisely, but still informally;[16] doing so will pay dividends for us in the present chapter, in no small part because both theories of consciousness featured herein propose a set of axioms. We shall therefore say that the phenomenon p to be explained is described by some collection of declarative statements about p, and denote this collection by

$$\Delta(p).$$

This description is the *explanandum*, as philosophers of science say. And we can say, following these philosophers further, that the content that explains the explanandum is the *explanans*, E.

The previous paragraph is perhaps a bit abstract. But it has in fact been concretely instantiated before the eyes of humanity time and time again as science has progressed, and, again, the pair of theories of consciousness purported by their proponents to explain consciousness conform to our setup very neatly. A pleasing case in physics is classical mechanics, now standard fare in high school, and something we therefore assume our readers to have studied. Classical mechanics (thunderously) arrived on the scene by way of Newton's *Principia* in 1687.[17] In this case, observations of the behavior of macroscopic objects are expressed in an instantiation of our $\Delta(p)$; i.e., this is the explanandum.[18] And what is the explanans in the Newtonian case? Fortunately, the answer to this question is long completely settled, and hence we can say with total precision (but leaving details aside here), and with no loss of generality, that E in this case is a collection of axioms that can be expressed specifically in formal logic.[19] If we dub this collection E_{Newton}, and again without loss of generality focus on phenomena specifically involving a sub-class of macroscopic objects of much interest to Newton himself, viz., planets, we can encapsulate the happy situation in the case of mechanics by way of the following economical expression, whose summative import should be quite clear:

(1) E_{Newton} **Explains** $\Delta(\textit{planets})$.

Proposition (1) asserts that the axioms of classical mechanics explain the observed phenomena: namely, planetary motion. Of course, in the case of first-rate science, it will also generally be required that declarative content which does the explaining (in some form often suitable for rapid calculation) has predictive power, but this is a requirement that we do not need to have explicitly represented in the present chapter.

One additional thing we do need to explicitly take note of, for reasons that will become clear below, is that the explanans needs to itself enable *measurement* of observational data in some systematic fashion. (Anticipating, in the case of IIT, measurement of the degree of consciousness is via Φ, and in the case of TCC, measurement is achieved by applying Λ.) For ease of reference and some generality, we shall say that some measurement scheme $Measure_E$ associated with a given explanation E can be applied to some description of observational data to produce a value μ. This may strike some readers as needlessly abstract, and perhaps even pedantic, but concrete instantiations of this scheme have been obtained and used for engineering, which is concrete as well.[20]

The current science of consciousness is very much driven by the perceived need to measure descriptions of observational data arising from study of things thought by some to be conscious. As we shall see, in the case of IIT, the role of its particular measurement scheme, Φ, is absolutely central. But before turning to IIT and Φ, we

turn first to what, given the simple framework we have set out, the science of consciousness in general is.

2.2 What Then Is the Science of Consciousness?

So then, what about consciousness, our central concern herein? Well, the template we have created works nicely to help us make sense of the structure of our chief topic of concern in the present chapter. Specifically, but leaving things schematic for the moment, we have this:

(2) E_X **Explains** $\Delta(\textit{consciousness})$ and $M_{E_X}[\Delta(\textit{consciousness})] = \mu$

Here, of course, some as-yet-unseen declarative content

$$E_X$$

explains observational data $\Delta(\textit{consciousness})$ regarding consciousness, and M_{E_X} can be applied to measure such data in certain ways. The "X" here is a variable that can be instantiated to a given theory of consciousness. What about the observational data in this case; that is, what about Δ? These data will vary considerably with the nature of observation used. In the case of IIT/Φ, relevant brains are often carefully examined. For example, if a given human reports having emotional states of various sorts, this is observational data; and if this human lacks some portion of his or her brain (as seen by direct inspection of the brain in question), this is added to what will compose an instance of $\Delta(\textit{consciousness})$. In personal communication with IIT originator Tononi and IIT proponent Koch (both of whom are discussed below), these have in fact sometimes been exactly the kind of observational data they have touted as explained by IIT brought to bear in the case at hand.[21]

The use of the rather abstract framework we have introduced may be somewhat hard to see directly at work in some "soft" sciences (e.g., psychology) that proceed in the absence of formal, declarative content that has been or at least can clearly be expressed in axioms, but the framework is easily seen to be firmly in operation in physics, for instance. Not only is the rigorous science of the kinematics of ordinary macroscopic objects and their behavior captured by the framework, in "Newtonian style," but the kinematics of special relativity is a perfect match as well, since here too there are both axiom systems that explain the relevant (described) observational data, and associated mechanisms enabling measurement of this data.[22] However, our purposes at hand of course pertain centrally not to physics, but to consciousness, to which we now turn.

2.3 But What Kind of Consciousness?

Unfortunately, a crucial fact about scholarship on consciousness is that it has long reflected the fact that the term "consciousness" (as well as, correspondingly, the adjective "conscious") is explosively polysemous. For the present chapter it will fortunately not be necessary to canvass the full, vast landscape of the relevant alternate meanings. It will suffice to have before us, to start, but three meanings, the first two of which are very nicely adumbrated by Ned Block.[23] These two are *access consciousness* and *phenomenal consciousness*; for short, following Block, these are respectively *a-consciousness* and *p-consciousness*.

We of course referred to the latter above, and gave at least a rough-and-ready characterization. P-consciousness is often characterized as "what it's like consciousness." As all (human) readers will agree, there is something it's like to be hot and severely thirsty, and to be able to sit, rest, and drink not just wine, but that first sip (or gulp) of lovely ice water, or iced tea, or lemonade, etc. In fact, we would be willing to

bet you can remember the qualitative aspects of such an experience, even if you are presently in the cold, with no desire whatsoever for such a pleasant beverage. The phenomena to which we refer needn't be so particular as ending thirst: for instance, surely there is something it's like to be you, and to be Naveen, and to be Selmer. In this case we are dealing with p-*self*-consciousness.

At the end of the day, that humans enjoy p-consciousness (i.e., enter into p-conscious states) is why they generally wish to do things, and in fact is in general why they wish to continue living. Why do a host of able-bodied humans spend inordinate amounts of time practicing and competing in the (glorious) game of cricket, rather than simply ending their lives? Because when they do play cricket they enter numerous what-it's-like states of mind that feel quite good, and are hence highly desirable for such agents.[24] P-consciousness is of course not solely the province of people, at least in the minds of most thinkers. Canines, for example, who among nonhuman animals are unique in that they have co-evolved with *H. sapiens sapiens*, can enter p-conscious states of joy, pain, anxiety, and fear, and doubtless other such states as well.

What about a-consciousness? Its definition remains murky—perhaps even irremediably so, unless people can be persuaded to adopt Bringsjord's long-ago-issued recommendation to discard the term "a-consciousness" in favor of terms that refer to the kinds of things this umbrella term is supposed to cover.[25] Having said this, Block's definition is as follows: A state of some agent is a-conscious if and only if it is poised to be used (a) as a premise in reasoning, (b) for rational control of action, and (c) for rational control of speech.[26] Actually, Block tells us that condition (c) isn't necessary, since—as he sees matters—nonlinguistic creatures can be a-conscious in virtue of their states satisfying only (a) and (b).

By this definition, as Bringsjord has in the past pointed out,[27] a run-of-the-mill database application currently running on a laptop is a-conscious, since such an application satisfies Block's three clauses (a)–(c). This is so, one, because such an application can be based directly on standard first-order logic, which ensures that a state of the system is nothing but a set of first-order formulae used as premises in deductive reasoning carried out to answer a data query. Two, action, for instance responding to a query with information, is controlled by rational deduction from such sets of formulae. Three, "speech" can be easily controlled by rational deduction from such sets with help from formal grammars designed to enable simple conversation in English. The simple imagined database system "talks" by producing text, but it could of course be outfitted with a text-to-voice synthesizer.

We assume most of our readers will agree, in light of this example, that many, many computational systems can be a-conscious (but not necessarily p-conscious). After all, if a database system is a-conscious by virtue of satisfying conditions (a)–(c), then many robots are too. So, we thus arrive at a fact that in our experience many readers are unaware of, a fact that we emphasize: viz., *there is in the mainstream cognitive-science literature a longstanding version of consciousness that can be veridically ascribed to computing machines, in the complete absence of these machines having p-consciousness, or anything of the sort*. Bringsjord asserts that robustly a-conscious robots will appear in the future, and will be so human-like that ascriptions of "consciousness" to them by humans will be routine and wholly uncontroversial.[28]

We now have two starkly divergent kinds

of consciousness on the table. But readers will remember that we promised above to characterize not just two kinds of consciousness, but three. The third is philosophically in line with a-consciousness, but more nuanced and robust, and based directly on the core concepts of logic-based AI that were firmly in place after being largely introduced at the 1956 dawn-of-AI Dartmouth conference, by AI founders H. A. Simon, Allen Newell, and John McCarthy.[29] The latter spent a career pushing AI forward, on the strength of the idea that machine intelligence could and should be based on declarative content expressed in logics, and on automated reasoning over this content by the machine in question.

At the conference in question, Simon and Newell introduced the groundbreaking AI system LogicTheorist, which automatically produced theorems in the propositional calculus that had to that point been the province of human minds.[30] While neither of these two broached the subject of whether an AI system of the logicist variety had or soon would have a form of consciousness and other aspects of "deep" human-level mentation, McCarthy did, and he held that such mentation could consist in the reasoning over sufficiently rich declarative information, expressed in formulae in the formal languages of formal logics.[31]

Now, *cognitive consciousness*, with its roots in AI and specifically in the work of Simon, Newell, and especially McCarthy, is much in line with a-consciousness, and is, we admit, the brand of consciousness near and dear to our hearts (for reasons to be in part shared below). We use just *c-consciousness* for short. This brand of consciousness is present only when the agent that bears it has, by virtue of internal declarative formulae and reasoning over them, a robust ensemble of cognitive attitudes, which correspond directly to a relevant set of verbs bound up with human cognition as long investigated in cognitive psychology and cognitive science.[32] The set of these verbs includes: *believing*, *knowing*, *perceiving*, *communicating* (in a natural language, and perhaps also a formal language that might be used in, say, mathematics), *hoping*, *fearing*, *intending*, and so on *ad indefinitum*. As far as we can tell, any agent or system that is cognitively conscious (= i.e. that enters into a series of c-conscious states through an interval of time[33]) is necessarily a-conscious during this stretch. In general, we see no harm in viewing cognitive consciousness to be the most important type of a-consciousness identified by human scientists and engineers thus far.[34]

Let us now take stock. At this point we have available to us three types of consciousness; accordingly, we have three variants of our "explanation template," one for each type:

(2^p) E_X **Explains** Δ(*p-consciousness*) and $M_{E_X}[\Delta(\textit{p-consciousness})] = \mu_p$

(2^a) E_X **Explains** Δ(*a-consciousness*) and $M_{E_X}[\Delta(\textit{a-consciousness})] = \mu_a$

(2^c) E_X **Explains** Δ(*c-consciousness*) and $M_{E_X}[\Delta(\textit{c-consciousness})] = \mu_c$

From this point on in the present chapter, our concern will be mostly with p-consciousness and c-consciousness. Again, the latter is a robust expansion, refinement, and—often, in artificial agents produced by AI—implementation of a-consciousness.

3. Integrated Information Theory (IIT) & Phi (Φ)

Note, then, that the type of consciousness IIT is intended to (and—for e.g. Tononi and Koch—in fact does) explain is none other than p-consciousness. Hence we can encapsulate the

scientific aspiration of IIT by way of the following proposition, using the abbreviation scheme we have allowed ourselves:

(2^p_{IIT}) E_{IIT} **Explains** Δ(*p-consciousness*) and $\Phi_{E_{\text{IIT}}}[\Delta(\textit{p-consciousness})] = \mu_p$

Of course, this isn't very informative in the absence of a characterization of IIT and Φ. We provide this characterization now. As promised, the characterization is done in two stages, and in a way that rescues the reader from having to wade through any of the detailed mathematics of Φ. In the first stage, which immediately follows, we look at the axioms and axiomatic approach of IIT. Following that, we turn to two hypothetical but—given the nature of modern cognitive robotics—very realistic robots in order to convey the mathematical core of Φ in a somewhat metaphorical manner.

3.1 The IIT Progression: Axioms to Postulates to Math

The general structure of what is required of IIT/Φ to meet the scientific challenge of p-consciousness is now before us in the form of (2^p_{IIT}). Does IIT/Φ conform to this required structure? At least at first glance, the pair certainly does. To see this, recall that canonical alignment with the required structure happens when what is doing the explanation of the descriptions of targeted phenomena is an axiom system. We made reference above, remember, to the axioms of Newtonian mechanics, but since those axioms are needlessly complicated for purposes at hand, we can turn to a simpler axiom system to anchor exposition and evaluation of the purported axiomatic basis of IIT: simple arithmetic with addition and multiplication. The particular simple axiom system we bring to bear is commonly known as "Peano Arithmetic," or just **PA**. Here are two simple axioms in **PA**, where '$s(n)$' denotes the successor of n:[35]

(PA1) For all $n : n + 0 = n$

(PA2) For all $n, m :$ if $s(n) = s(m)$ then $n = m$

Please notice immediately, and importantly, both the pure declarative nature and the clarity of these two axioms. They both are direct, simple propositions. As to their internals, both make use of the universal quantifier ("for all"), use variables that range over a set of numbers that even fourth graders begin to understand substantively (the natural numbers), employ only simple addition as the central concept, and ask that only one additional concept be understood by readers: the concept of one natural number being equal to another natural number. That's it. The straightforwardness and clarity are noteworthy and unmistakable. Other axioms in **PA**, and other axioms in other axiom systems (including those specifically designed to directly enable explanation of physical phenomena studied by physicists[36]), follow suit, invariably.

As to the observed phenomena that **PA** explains, this includes such elementary-school observations as that $3 + 0 = 3$, which young children encounter (not just internal to their minds when reflecting, but empirically when, say, moving three cubes from a table into an empty bowl, and then observing that while that bowl had held zero cubes it now contains three); that there exists a (natural) number n greater than prime number 37, which they can also encounter; and so on *ad infinitum*, including of course some rather more surprising things that turn out to be observable in the realm of elementary arithmetic. For example, it is often observed afresh by some humans even today that as you explore the progression of *positive cubic numbers*, that is numbers of the form n^3, start-

ing with 1^3, 2^3, 3^3, 4^3, 5^3, in each case you can find a sum of n consecutive odd numbers that equals the cubic number.[37]

PA does its work of explaining some given arithmetic phenomena when there is a series of deductive inferences from one or more of the axioms in question to the targeted phenomena. This is indeed the mechanism of explanation for any axiom system that succeeds in explaining some phenomena. For a very simple example that nonetheless results in no fundamental loss of generality, consider again the first of our arithmetic phenomena above, that $3 + 0 = 3$. How is this to be proved? The proof is simple. First, note that the natural number 1 is represented as the sucessor of 0, $s(0)$; the natural number 2 then as $s(s(0))$; and finally the natural number 3 as $s(s(s(0)))$. Our next step is to deduce from axiom (3) in one step that

$$s(s(s(0))) + 0 = s(s(s(0))),$$

by substituting for the variable n our encoding of the natural number 3. This inference is trivial, and is made when students are taught simple algebraic equations in (approximately) sixth grade (US). This isn't exactly a robust proof, but it's an ironclad building block for number theory and human mathematical cognition, and besides, the fact is that an enormous body of mathematical phenomena, some of it mind-bendingly complex, is provable in like manner from, and hence explainable from, **PA**.[38]

It is important to note that **PA** is a theory of arithmetic in no way restricted to human cognition. Rather, the axiom system covers, and in principle explains, a formal space that is explored, or at least can be explored, by aliens, intelligent machines, supernatural beings, and so on. This is an important point, because likewise IIT aims to provide an explanation of p-consciousness in not just neurobiologically normal adult humans, but also in human infants, canines, machines, aliens, and so on. And, just like the development of axiomatic theories in logic, mathematics, and physics, the axiomatic approach of IIT is top-down. In this approach, one starts with "essential phenomenal properties of experience, or axioms, and infers postulates about the characteristics that are required of its physical substrate."[39]

The alert reader will have noticed something quite peculiar in the quote just given. You might wish to look at it again. This quote tells us that to explain or ground p-consciousness in various entities, it's not directly a set of axioms that do this, by enabling inference from them to what is targeted for explanation. Rather, the idea expressed here is that the axioms lead inferentially to "postulates," and the postulates then somehow lead to what explains or grounds p-consciousness at the physical level.[40]

Before going further on the nature of the relationship between axioms, then postulates, and then IIT expressed as something that can mathematically dictate, or at least express, the physical substrate that undergirds p-consciousness, we note for convenience and future repeated use in our prose that the progression can be pictured thus:

Axioms re P-Consciousness ⇒

Postulates re P-Consciousness ⇒

Physical Substrate.

Let us refer to this as *the IIT progression*. Before examining this progression, let's determine what the nature of the axioms of IIT is.

Doing so is not difficult, because the prominent proponents of IIT/Φ have been quite clear about this matter. By "axiom" in this context IITers mean a proposition that is self-evident to any p-conscious being who considers that proposition. This characterization of an axiom

in the world of IIT/Φ is easily found, repeatedly, in the relevant publications authored by these thinkers.[41] It seems quite reasonable to hold that the axioms of arithmetic, that is, our **PA**, are self-evident—at least to those who understand the concepts involved. Consider the specific axiom PA2 from above. Is this self-evident to anyone who understands the concept of increment-by-one on the natural numbers, and also understands the concept of identity expressed by "=" in arithmetic? There might be some debate, but it seems more than reasonable to hold the affirmative.

Very well, now to the IIT progression. Let's begin with the progression from axioms to postulates. The first thing to say here is that this part of the progression is strikingly idiosyncratic. Aren't postulates and axioms the same category of things? Aren't the two terms "axiom" and "postulate" interchangeable? Certainly this is true today. Yet most readers, and certainly the two of us, were first exposed to a rigorous example of postulates through the high-school study of "Euclid's Postulates," the first of which is traditionally expressed as the proposition that given any two points, there is a line segment that joins both of them. In this specific context, postulates are sometimes understood to pertain specifically to geometry, while axioms range over branches outside geometry to cover all of mathematics.

Unfortunately, whether we adopt the modern position, in which "axiom" and "postulate" are coextensive, or adopt the old view that postulates are reserved for geometry, we find nothing helpful with respect to the presentation of IIT in the literature. IIT proponents hold, and the IIT progression indicates and sums this up, that IIT's postulates *follow from* IIT's axioms. But, try as we might, we can make no sense of this whatsoever in the context of the formal sciences, nor in the particular natural sciences. To see why we're non-plussed, let's next take a look at some of the axioms and postulates in question. Here is the first one, the Axiom of Intrinsic Existence (which we abbreviate as "(IntExis)"):

> Consciousness is specific: each experience is the *particular* way it is—it is composed of a specific set of specific phenomenal distinctions—thereby differing from other possible experiences (*differentiation*). Thus, an experience of pure darkness and silence is what it is because, among other things, it is not filled with light and sound, colours and shapes, there are no books, no blue books and so on. And being that way, it necessarily differs from a large number of alternative experiences I could have. Just consider all the frames of all possible movies: the associated visual percepts are but a small subset of all possible experiences.[42]

If this is a scientific axiom, it surely is a most peculiar, unconventional one, since it appears to do anything but assert a clear proposition (relying as it does on a rather vague cinematic metaphor). Contrasted with what we saw above in the case **PA**, specifically in the pair (PA1) and (PA2), IIT's first axiom is profoundly disappointing. We honestly have no idea what to make of it. Are the other axioms clearer and, in the context of other axiom systems in high regard for packing an explanatory punch, more informative? Sadly, it would appear not; but our readers can judge for themselves. Here's a presentation of the axiom from among the quintet that seems to us to be the clearest and sharpest, The Axiom of Integration, or just "(Integ)":

> The axiom of integration states that experience is unitary, meaning that it is composed of a set of phenomenal distinctions, bound

> together in various ways, that is irreducible to noninterdependent subsets.[43]

This may be the clearest among the five, yet we do not know what it means, alas—and this despite much reflection. We can of course guess, and join others in doing so.[44] Our best guess says the following, put informally, but, we think, helpfully:

> Every experience over a given interval of time is had by some agent and by no others, and all the properties possessed by each such experience over that interval are internally perceived by that same agent and by no others.[45]

With this generous rehabilitation, the axiom (Integ) makes considerable sense, and will pay dividends later when we consider objections to IIT—but on the other hand this axiom isn't self-evident. After all, in the history of philosophy of mind, certainly David Hume, who famously said that the "self" is an illusion and a "bundle or collection of different perceptions" and no more,[46] would for example most assuredly reject it (Integ), after saying that confidence in its truth derives from an illusion. But worse for IITers is the brute fact that our rehabilitation is *our* rehabilitation: certainly others will interpret their originai, imprecise version of the axiom differently. (In fact, Bayne sees three additional interpretations, each distinct unto itself and separate from our reworked version.[47]) Needless to say, that's not how axioms in explanatory axiom systems are supposed to work.[48] Of course, as we have already noted, proponents of IIT proceed in idiosyncratic fashion, by following what we have called the "IIT progression."

While we despair of figuring out the inferential relationship between the axioms of IIT and the theory's postulates, each axiom "leads to" or "implies" a postulate, and then in turn, for the last of the third steps in the IIT progression, the postulate leads to formal, computational structure that is expressed mathematically. We read:

> [Postulates are] assumptions, *derived from* axioms, about the physical substrates of consciousness (mechanisms must have causal power, be irreducible, etc.), which can be formalized and form the basis of the mathematical framework of IIT.[49]

This being a chapter that avoids the mathematics in favor of our use of the robots High and Low to express the mathematics in more intuitive fashion (carried out in the next section), we shall rest content with a closer look at the particular postulate that—in some manner that remains entirely mysterious to us, alas—"follows from" the axiom (Integ). Quoting, we have Postulate 4 as this:

> A mechanism can contribute to consciousness only if it specifies a cause-effect repertoire (information) that is irreducible to independent components. Integration/irreducibility [in Φ] is assessed by partitioning the mechanism and measuring what difference this makes to its cause-effect repertoire.

We will momentarily see how our two hypothetical robots, High and Low, bring to life what is said here. The basic idea, in connection with Postulate 4, can be encapsulated thus: High, a robot with high Φ, is such that any attempt to partition its information processing will result in its operational paralysis—whereas Low is expressly designed so that its components have standalone power, and they can be used rather like building blocks for robotics engineering. We turn now to these two robots in order to enable the reader to understand the

mathematics of Φ without having to wrestle with that math directly.

3.2 Conveying the Math of IIT's Φ by Parable

To explain the mathematical essence of Φ we turn to a parable. Our story features two household robots, Low and High. Robot Low is designed by one robotics company, and robot High by another such company. These two companies have long been guided by radically different conceptions of how to best build an effective and useful household robot—the kind of robot that (at least in the marketing literature) cooks succulent meals, makes the laundry sparkle, walks the family dog, vacuums and dusts, declutters, and does yard work as well (e.g., rakes leaves and shovels snow). Low, who—as the reader has likely guessed—eponymously has very low Φ, has been designed and engineered to be as *modular* as possible. For instance, Low's food-preparation module **F** is completely separate from its do-the-laundry module **L**. In fact, module **F** is composed of many sub-modules, and there is no overlap of these sub-modules with any other modules. When for instance Low makes red Italian tomato-based sauce for eggplant parmigiana, the movements needed to do so—lifting and placing pots, pouring finished pasta into the colander, etc.—are all the result of task-specific planning entirely separate from planning out and performing actions with respect to **L**. Low's entertainment module, **E**, which handles the smart TVs in the home, and management of all the content that can be displayed on these TVs, is likewise self-contained and not integrated with any other modules. When Low runs algorithms that result in recommendations to humans regarding what these humans will likely enjoy watching on a given night, none of the computational logic in these algorithms is connected to that which is used for cooking or doing laundry.

Now let's try to make things a bit more rigorous. Let Low_t^k, the global state of robot Low at time t, be the state of each of k propositional variables $p_1, p_2, \ldots, p_k$. Each of these variables represents a piece of declarative content, such as that *the laundry bin is full*, or *the water is boiling*, or *the movie Coda would be good to watch tonight*, and so on. For some added realism, let's assume a trivalent scheme, so that at any given time a propositional variable p_i can be in one and only one of three possible states: 1 (or true), 0 (or false), or U (or unknown).[50]

Through time, Low enters new states that are determined by a function u_{Low} from all the possible permutations of all the relevant propositional variables ($\mathcal{P}$), to this same state. Let's further assume that the global state "tracks" activity and the conditions in the three domains cited above: laundry, food preparation, and what is to be played and modulated on the entertainment system in the house in which Low serves. This means that $\mathcal{P}$ can be partitioned into three sets of propositional variables, one for each of the three domains.

Now, Low has low Φ because as this robot "lives" through time, the function u_{Low} can be defined completely by subsidiary functions that track the propositional variables that pertain *only* to a given module. This is in violation, or at least extreme tension, with the axiom (Integ) and the corresponding postulate for it that we visited above. For additional fixity, suppose that the global state of Low at time t is (where each row is the permutation at t of the module indicated by the leftmost column):

F	1	1	1	0	1
L	1	1	1	0	1
E	1	1	1	0	1

At the next time, t', Low's global state is the result of the application of three entirely separate transition functions, one for each row of this table, working completely unto themselves. If, respectively, they regiment "Leave unchanged," "Invert all numerical values," and "Make everything true unknown," then Low is in this global state at t':

F	1	1	1	0	1
L	0	0	0	1	0
E	U	U	U	0	U

Now let's turn to the robot High, which intuitively is to have high Φ because it conforms to the axioms and postulates of IIT. We suggest that you think about this conformity in connection, specifically, with (Integ) and its postulate. Let's here again focus on the same three areas of service in the same three domains. But now things are very different, in two ways. First, High is capable of carrying out things in parallel, to a significant degree. At the very same time that High is doing the laundry, it can be hearing and speaking through TVs in the house about entertainment options and so on, and indeed at the same time it can also be planning out tonight's dinner.[51]

Put in terms of the matrices from above, we can easily create a depiction that reflects the integrated operation of High. Suppose that High at time t is in this global state:

F	U	1	1	1	0
L	1	1	1	1	0
E	1	1	1	0	1

And suppose in addition that propositional variables in High change in accordance with the "global" function that if the value v of p_i is at a given time 1 (or 0), and its neighbor directly above or below is a match, then its value at the next time will be 0 (1), and is otherwise unchanged. This results in the global state of High at time t':

F	U	0	0	0	1
L	0	0	0	0	1
E	0	0	0	0	1

We encourage the reader to reflect upon the situation when Low and High have their global states determined by enormous matrices, and when the global function u_{High} itself factors in larger and larger "neighborhoods" surrounding a given propositional variable. However large these matrices are, it will be the case that Low continues to operate essentially as a host of "split personalities" bundled together in shallow fashion, whereas in the case of High, its global states through time will hinge on the interconnectivity through time of the values of the variables. High will have high Φ, and Low low Φ.[52]

We hope at this point that the reader has an intuitive grasp of IIT and its measurement scheme Φ.[53]

Of course, the $64,000 question is whether IIT/Φ succeeds. This question distills to whether it's rational to affirm, reject, or. suspend judgement on the proposition (2^p_{IIT}). What can be safely and fairly said at this point is that given how shaky its axioms and postulates are, there is plenty of room for skepticism about whether this proposition holds—but we don't honestly see at this point one or more fatal problems, and opt for suspending judgment. But let's now turn, as planned, to direct attacks on this pair from Searle.

4. Searle's Attacks on IIT/Φ

John Searle has energetically attacked IIT, in a way he regards to be utterly fatal to the theory.[54] For those familiar with Searle's body of work through many years, the general philosophical roots of his attack on IIT are in fact

decades old; but fully charting this intellectual history would take us too far afield and would add little to an analysis of the attack itself. (Nonetheless, below, we shall specifically need to consider the fact that Searle's longstanding complaints about the "observer-relativity" of computation are central to some of his attacks on IIT.) Fortunately, there is a convenient shortcut available to getting clear on what Searle's critique is: namely, we can jump to the latest exchange directly between Searle on the one hand, and IIT/Φ proponents Koch and Tononi on the other. This exchange crystallizes the Searle-versus-IIT-proponents debate, and pivots around the IIT -and Φ-based treatment of p-consciousness provided in Koch and Tononi's *Consciousness: Confessions of a Romantic Reductionist*. This treatment matches exactly the brief orientation we gave to IIT and Φ in the previous section of the present chapter.

Searle reviewed the book in question, Koch and Tononi replied, and Searle replied to the reply.[55] As the reader can infer, Searle had the last word in this exchange, and that last word is the best and most efficient place for us to shine the light of our attention, since as a matter of fact the dialectic between Koch and Tononi, versus Searle, does gradually crystallize into a very efficient two-part presentation of the Searlean challenge to IIT and Φ. We turn now to the first part of the challenge, Attack #1 from Searle.

4.1 Attack #1: "IIT is Observer-Relative!"

Here is Searle's first attack, in his own words:

> [W]e cannot use information theory [= IIT] to explain consciousness because the information in question is only information relative to a consciousness [= agent, for us]. Either the information is carried by a conscious experience of some agent (my thought that Obama is President, for example) or in a non-conscious system the information is observer-relative—a conscious agent attributes information to some non-conscious system (as I attribute information to my computer, for example).

What should we make of this purported refutation? Does it succeed? In our opinion, despite Searle's tone of triumph, not at all. Fortunately, the time we took above to distill the core doctrine of IIT and Φ, viz., (2^{p}_{IIT}), combined with our efficient but accurate setting out of IIT itself,[56] allow us to make at least some sense of Searle's reasoning, and to then justifiably reject it as flatly inadequate for showing that IIT has no explanatory power. This rejection will exploit a simple but illuminating look at the explanatory scheme and power not of a theory of consciousness, but of elementary arithmetic, a domain that we have of course already visited earlier.

To begin our analysis, please examine again the core doctrine of IIT, and you will see afresh that a collection of declarative statements is what is supposed to do the explaining (of p-consciousness); this collection is E_{IIT} notionally, but this is a reference to the axioms and postulates of IIT. So let us consider a case of explanation in the general form of (2) (see §2.2) that's even simpler than the axioms of classical mechanics we alluded to above: viz., the axioms of arithmetic on the natural numbers, $\mathbb{N}$, which we also discussed. Recall that we specifically cited two axioms of **PA**, (PA1) and (PA2). And recall as well that we discussed the observed phenomena that **PA** explains.

This puts us in good position to assess Searle's Attack #1. Put in terms of the key template, Searle is claiming that Δ_{IIT} is observer-relative, and that this is a fatal defect afflicting IIT. This is actually not a new sort of complaint from Searle. In 1992, he claimed that the view that the mind is essentially a computer is

unacceptable, because computation is observer-relative.[57] We shall return below to this claim of Searle's, but at present we don't need to investigate this earlier work, because what we have on the table now regarding **PA** allows us to see that Searle does no harm to IIT at all. How do we see this? Well, first let's have an instantiation of (2) before us for the case of **PA**'s explaining any number of observations about arithmetical propositions, including the ones we cited above:[58]

(2^{a}_{PA}) E_{PA} **Explains** $\Delta(\textit{arithmetic})$ and $M_{E_{\text{PA}}}[\Delta(\textit{arithmetic})] = \mu$.

But now look closer at this instantiation of our template. Is it not true that this very proposition is observer-relative? After all, who or what is the explaining that **PA** provides *for*? It's for agents; specifically, for human agents: us. This is thoroughly unsurprising, since the arithmetic that **PA** explains is the arithmetic that human beings have long explored. The upshot is that all along in the present chapter we have been implicitly talking about scientific explanation that is relative to an observer. Which observer? To a human scientist. We can regiment this fact by slightly expanding the templates we have employed above. In the case of elementary arithmetic, the expansion can be as follows:

$(2'^{a}_{\text{PA}})$ E_{PA} **Explains** $\Delta(\textit{arithmetic})$ to agent A; and $M_{E_{\text{PA}}}[\Delta(\textit{arithmetic})] = \mu$

The insertion here is that the explaining is to some agent. In short, the explaining is relative to an observer, the observer who encounters such things as we have cited above regarding cubic and odd numbers.

At this point, we have on hand enough information to see that Searle's Attack #1 is anemic. The reason is pretty obvious. If explanations of phenomena as straightforward as those of an arithmetic[59] nature are relative to an observer, and this fact is benign (indeed, it's outright *desired*), then it can't be objectionable that IIT follows the same pattern. In fact, the pattern can be regimented by a variant isomorphic in structure to what we just presented for arithmetic:

$(2'^{p}_{\text{IIT}})$ E_{IIT} **Explains** $\Delta(\textit{p-consciousness})$ to agent A; and $\Phi_{E_{\text{IIT}}}[\Delta(\textit{p-consciousness})] = \mu$

Searle might retort that what he is fundamentally criticizing in the case of IIT is (and we here use a simplification of explanation as we have laid it out schematically that is in line with Searle's more informal treatment) its providing explanation by X or Y when X is observer-relative, and Y isn't.[60] And, since X in our example is the axiom system **PA**, which is observer-relative, our counter-argument based upon the analogue in which $X =$ **PA** and $Y =$ the arithmetic observations we have canvassed (and the like), our reasoning fails.

We are inclined to believe that, in point of fact, Searle would indeed say something like this—but that in saying it he would attempt intellectual legerdemain. The reason? Well, Searle's reasoning is enthymematic; he snuck in a premise. That he did is revealed by asking: Hey, who says **PA** is observer-relative? It's not; in fact it's *clearly* not, at least as far as we can see. Let all human minds expire ten minutes hence from the face of Earth, yet every iota of declarative information composed by the deductive closure of **PA** will remain firmly true and quite real. And if a non-human race of silicon- and not-carbon-based alien minds springs up on a distant planet, and they reach a stage of cognitive development in which they can count and number things, and size collections, and then add to these collections systematically, well then, this race will discover **PA** (no doubt by an-

other name) and—at least a significant portion of—its deductive consequences.

Searle may insist that **PA** *is* observer-relative, but then he will need to provide an argument in support of this position, since simply insisting that something is true hardly makes for a compelling case. But this means, inevitably, that Searle will need to incorporate a massive new amount of philosophical background, from sub-areas of philosophy that are themselves enormous and quite subtle in their own rights (e.g., philosophy of mathematics)—which is to say that, no matter how we slice it, Searle's attack on IIT, for present purposes, has been neutralized.

Moreover, Searle is in no better shape with respect to IIT and Φ than he is in with respect to **PA**. We say this because Searle's case against IIT/Φ rests upon the claim that the assignment to X in the locution "X explains Y" is observer-relative. Of course, that which instantiates the variable X for Searle is none other than Φ. But Φ is no more observer-relative than **PA** is (the axioms and postulates alleged to "lead" to Φ *are*, we have noted, horribly imprecise, esp. compared to **PA**). Φ is after all an abstract characterization, and a formal one to boot, that matches in general the concepts and structures at the heart of theoretical computer science. So, if Φ is observer-relative, then for example the established formal hierarchies of ever more powerful computation are as well! One of these hierarchies, the so-called *Arithmetic Hierarchy*, lays out a spectrum of increasingly powerful computation over the natural numbers. Is this spectrum observer-relative? It has standalone mathematical structure that in no way disappears when humans aren't thinking about it and its constituents. So it's not observer-relative. And, just as in the case of **PA**, which of course is nothing more than description in formal logic of $\mathbb{N}$ and the simplest of arithmetic functions on it, if Searle sticks to his guns, then he is obliged to show us *why* the Arithmetic Hierarchy is observer-relative. Such a demonstration, which would be in and of itself a major contribution to philosophy of mathematics, is obviously not forthcoming.[61]

4.2 Attack #2: "But IIT Entails Panpsychism!"

Searle's second attack is more straightforward than his first. It's simply that IIT implies everything is conscious, that is, that panpsychism holds. Since—so the argument goes—panpsychism is absurd, and thus false, IIT falls. Here is Searle again verbatim:

> They [Koch and Tononi] claim not to be endorsing any version of panpsychism. But Koch is explicit in his endorsement and I will quote the passage over again: "By postulating that consciousness is a fundamental feature of the universe, rather than emerging out of simpler elements, integrated information theory is an elaborate version of *panpsychism*." (p. 132, emphasis in the original) [. . . IIT] has panpsychism as a consequence.[62]

Is Attack #2 successful? Searle in our opinion has managed to reveal that Koch and Tononi aren't exactly clear about their attitude toward panpsychism over the course of their book (Do they *thoroughly like* the fact that IIT entails panpsychism? What import does "elaborate" have in his quote of them just above?), but this in no way constitutes a refutation of IIT itself. Why can't Koch and Tononi simply retort to Searle that, as a matter of fact, they do very much like a version of panpsychism's being a consequence of IIT? Unless Searle has an independent refutation of the proposition that consciousness is everywhere in the universe, his second attack does no damage to IIT and Φ at all.

Well, as a matter of fact, Searle *does* have an argument for the proposition that panpsychism is absurd; he gives it in his original review (it's not discussed in the second-round exchange we have drawn from so far). Here's the argument, in full, in Searle's own words:

> Consciousness comes in units. The qualitative state of drinking beer is different from finding the money in your wallet to pay for it. But a consequence of its subjectivity is its unity. So for example, I am conscious and you are conscious but each consciousness is separate from the other; they do not smear into each other like adjoining puddles of mud. Consciousness cannot be spread over the universe like a thin veneer of jam; there has to be a point where my consciousness ends and yours begins. For people who accept panpsychism, who attribute consciousness, as Koch does, to the iPhone, the question is: Why the iPhone? Why not each part of it? Each microprocessor? Why not each molecule? Why not the whole communication system of which the iPhone is a part? The problem with panpsychism is not that it is false; it does not get up to the level of being false. It is strictly speaking meaningless because no clear notion has been given to the claim. Consciousness comes in units and panpsychism cannot specify the units.[63]

Certainly the metaphors relied upon here will be intuitively attractive to many who read them; but even a slight grasp of the mathematical essence of Φ, combined with basic knowledge of how information-processing artifacts in our world work, it seems to us, negates this attractiveness. We both, and perhaps you the reader as well, have spread "a thin veneer of jam" over many a piece of toast, and we accede to the claim that no such maneuver can be pulled off for consciousness as it can be for apricot jam. Likewise, we suspect that our life experience with puddles of mud, and yours too, suggest that the mixing of one or more of them, if each figuratively denotes a conscious creature, is hard to make sense of. But why would such visceral feelings, however much they are in agreement with Searle's position metaphorically expressed, carry any weight in a dialectic that must be discursive, at least ultimately? IIT is nothing if not a theory about information-processing systems, and are such systems really like jam and mud?

In point of fact, the answer to this question, given our coverage above of the axioms of IIT, and specifically the axiom (Integ), which insists on a unity of consciousness for conscious experiences, should be a firm negative. In fact, our formal version of the axiom (Integ) could be used to *prove* that Searle's complaint is unfounded. The fact of the matter is that when IIT says that consciousness is everywhere, its stated basis excludes a dissolution of the unity of consciousness. If IIT is right, then the universe is flooded in every corner with consciousness, yes—but taking place in equally ubiquitous agents that are the bearers of conscious states.

In addition, we can return profitably to the robots High and Low. Each robot is a determinate, separate-unto-itself information-processing system. If we needed to, we could describe both robots in technical glory, in accord with the practice of cognitive robotics.[64] This description would enable the manufacture and deployment of High and Low, eventually; it can be considered a blueprint, at the level of information flow and control, and it would have in it algorithms and dataflow symbols in flowcharts and so on. Let us focus our thoughts specifically upon the description specifically for Low; let's dub it "Des_{Low}." In this description, does anything jump out at us as a prime candidate for, or at least a prime candidate for a correlate of, the "I" in consciousness? No. We routinely

use the first-person pronoun to say such things as

> (†) "I am capable of doing the laundry, doing the dishes, and of selecting a movie for tonight,"

and this is the kind of phenomenon, at least at a linguistic level, that Searle is seeking to exploit in his argument. In the case of Des_{Low}, there is nothing at all to be seen that suggests an "I" in any way. In particular, Low, by definition, can't declare truthfully that (†), that is that it has the capacity to clean laundry and dishes and select entertainment. The reason, fundamentally, is that Des_{Low} is just a collection of sub-systems, each one for a separate activity in a separate domain, with zero integration between, let alone executive management of, these sub-systems. (We learned this during the articulation of the parable of High and Low.) But note well: the absence of anything information-theoretic that suggests an "I" is, in the case of Low, exactly what Tononi and Koch would expect and want, and indeed their expectation aligns with the mathematics of Φ. Φ essentially explains that Low can't correctly utter (†). We find that noteworthy, if not downright impressive.

But, when we turn to the case of High, things are very different, and the situation further neutralizes Searle's argument. To see this, imagine the corresponding blueprint for High: Des_{High}. And now imagine what it is in this blueprint that might suggest some correlate for the "I." Does anything come to your mind? Well, surely the blueprint is going to show some sort of executive control mechanism that can coordinate the operation of **F**, **L**, and **E**—and this "executive" is surely suggestive of a unifying element in High. Moreover, there is in fact every reason to think that this unity of processing in High, with mathematical details fleshed out, falls naturally out of IIT itself; which is to say, Searle is wrong in asserting that IIT has nothing to say about "units" of consciousness.

The upshot for us, and, we recommend, for the reader, is that while IIT/Φ has some serious deficiencies, it's far from dead. Regardless, we believe that there is a better game in town anyway. That game is to focus not on p-consciousness, but rather on c-consciousness, and its own measurement scheme, Λ. According to plan, we turn to this pair now.

5. The Theory of Cognitive Consciousness, Λ, and Intelligence

We characterized *cognitive consciousness* (= *c-consciousness*) above. As a brief reminder, and now put starkly, an agent a is c-conscious when it is in one or more c-conscious states (through time), and these states consist in a's X-ing such and such a proposition, where X is a cognitive verb and the "-ing" employs a gerundive nominal to denote the state in question. Examples quickly illuminate the core idea. For instance, *Robbie's believing that he knows the combination to a lock the location of which is unknown to Sally* is a cognitive state. If Robbie is a robot and this state holds at some time t_1, perhaps Robbie immediately thereafter searches his knowledge-base for the combination of the lock in question and sees it there; in which case at t_2 he enters the c-conscious state *Robbie's knowing that he knows the combination to a lock the location of which is unknown to Sally*. As we said above, the collection of cognitive verbs include: *believing*, *knowing*, *perceiving*, *communicating*, *hoping*, *fearing*, *intending*, and so on *ad indefinitum*. Now, what is the Theory of Cognitive Consciousness (TCC)? TCC consists of its own axiom system, $\mathcal{CA}$, combined with meta-

propositions regarding this system, which together characterize c-consciousness.[65] We do not in the present chapter spend time explaining various meta-propositions regarding $\mathcal{CA}$, and turn now directly to the axioms in question.[66] We suspect that the reader will agree that the axioms of TCC are a bit sharper than those of IIT.

5.1 Regarding the Axiom System ($\mathcal{CA}$) for Cognitive Consciousness

It would exceed the scope of the present chapter to even slightly approach here a recapitulation of all the axioms of cognitive consciousness (= c-consciousness). For the full axiomatic treatment, the reader is directed elsewhere.[67]

It will suffice in the current context if we show the reader but two of the simpler axioms of $\mathcal{CA}$, the first of which is:

Perception to Belief	
P2B	Human persons perceive internally[68] and externally,[69] and in both cases the percepts in question are believed [at varying degrees of strength, with external perception at the strength of *evident* (which here can be understood as "overwhelmingly likely"), but never *certain*] by these agents, whereas most of what is internally perceived is indeed certain.

P2B is pretty easy to understand. When we perceive such things as that seven is a prime number or that we seem to be sad, we believe these propositions, and they are *certain* for us. But when we perceive in a garden a pink rose, *ceteris paribus* we believe that there is a pink rose before us, but it could be an illusion. (We may have forgotten that we are wearing rose-tinted sunglasses, and we are in fact looking at a white rose.) In c-consciousness as we rigorize it, belief is stratified, in that a belief is accompanied by a *strength factor*. So for example Jones, if having ingested a powerful pain reliever in a hospital, and knowing that such drugs can have serious side effects, may believe only at the level of *more probable than not* that there is a walrus before him. With stratification in place, belief becomes graded from *certain* to *certainly false*, and so will knowledge. This means that our framework for $\mathcal{CA}$, in contrast with elementary standard logics, which have binary values TRUE and FALSE, or sometimes those two plus INDETERMINATE (recall the trivalent setup used in our parable of robots Low and High), has thirteen possible values. In large measure due to the research and engineering of Govindarajulu, and to significant contributions from Mike Giancola, we have some fairly robust implementations of artificial agents that embody axiom **P2B**, and bring this framework to concrete life.[70]

Now the second axiom we share is here:

Introspection (positive)	
Intro	Humans persons know that they know what they know.

This axiom is in fact well-known in formal logic because it corresponds to a much-discussed axiom from so-called *alethic* modal logic—an axiom customarily written

$$\Box \phi \rightarrow \Box\Box\phi,$$

when symbolized as the characteristic axiom of the modal logic **S4**, first introduced, along with four other modal logics, by Clarence Lewis and Cooper Langford.[71] In **S4**, the boxes here are read as "it's necessary that." In epistemic logic, we instead read $\Box$ as "knows that," often denoted by simply "**K**." A bound $k \in \mathbb{N}$ can be placed on the amount of iteration of **K**, but it

would we think need to be at least 5 for human-level cognition.[72] The axiom here can also be expanded to include provision for negative introspection (i.e., $\neg\mathbf{K}\phi \rightarrow \mathbf{K}\neg\mathbf{K}\phi$), and once again a bound can be placed on the iteration, if desired.

5.2 Cognitive Consciousness, Measured: Λ

Λ differs from IIT's measurement scheme Φ in two significant ways. First, Λ measures consciousness, specifically, as the reader now knows, c-consciousness, based on how the system observably behaves and how its internal operation observably works, instead of on the peculiarities of vague, unseen internal structures in the system. This is exactly as John McCarthy would have it, as discussed above. He wanted to see observable, external behavior of AIs/robots match the behavior in the human case that impels us to ascribe consciousness to humans—and he also wanted to see, "under the hood," that information processing conforms to reasoning that we know from the human case is directly associated with the requisite external behavior.

In short, unlike today's chatbots, McCarthy insisted that spitting out coherent prose would be insufficient: you would also need to find, internal to the AI, the structures and content and—most importantly—justificatory reasoning, in arguments and proofs, behind such prose. Finding only numbers swimming dynamically in sea of nodes as in the case of today's so-called "deep learning" artificial neural networks like ChatGPT[73] wouldn't qualify. Overall, then, rather than striving to measure phenomenal consciousness (p-consciousness), Λ explicitly aims to explain and account only for cognitive consciousness (c-consciousness). We have thus moved a considerable distance away from IIT/Φ and are hence shielded from the attacks of Searle, and from skepticism based upon lingering imprecision.

As to conveying precision, we already gave a sense of the clarity and crispness of the axioms of TCC, and we now present a condensed version of Λ. For the setting we use for exposition here, we assume we have an agent a that acts at discrete time points. For some of the agent's actions $\alpha(t)$, the agent possesses internally or outputs externally a justification/rationale *justification*(a, α, t). Λ is based on the richness of structures found in the justifications produced by the agent. The justification can be a semi-formal structure and can include a mix of different modalities (non-verbal actions, gestures, written content, etc.). If the structures include references to cognitive states of other agents or the agent itself, we in general assign a high Λ score to the agent at those points in time. Unlike Φ, we don't provide a single Λ value for an agent or system or creature which is to be measured with respect to c-consciousness; rather, what is provided is a sequence or vector of values corresponding to different cognitive components such as knowledge **K**, belief **B**, desire **D**, intention **I**, temporal structures $\vec{t}$, quantifiers $\forall, \exists, \ldots$, etc. Semi-formally, if we have justification *justification*(a, α, t) produced by an agent a for action α at time t, then:

$$\Lambda\big[\textit{justification}(a, \alpha, t)\big] = \big\langle \lambda_{\mathbf{B}}, \lambda_{\mathbf{D}}, \lambda_{\mathbf{I}}, \lambda_{\mathbf{K}}, \lambda_{\vec{t}}, \lambda_{\forall}, \lambda_{\exists} \ldots \big\rangle$$

In the above equation, Λ maps a justification to a vector of values λ, where each λ is a natural number $\{0, 1, 2, \ldots\}$. For example, if a justification j is constructed using multiple nested beliefs of other agents (*"John believed that Jack believed that Mary believed that the apple is red"*), but is deficient in other cognitive structures,

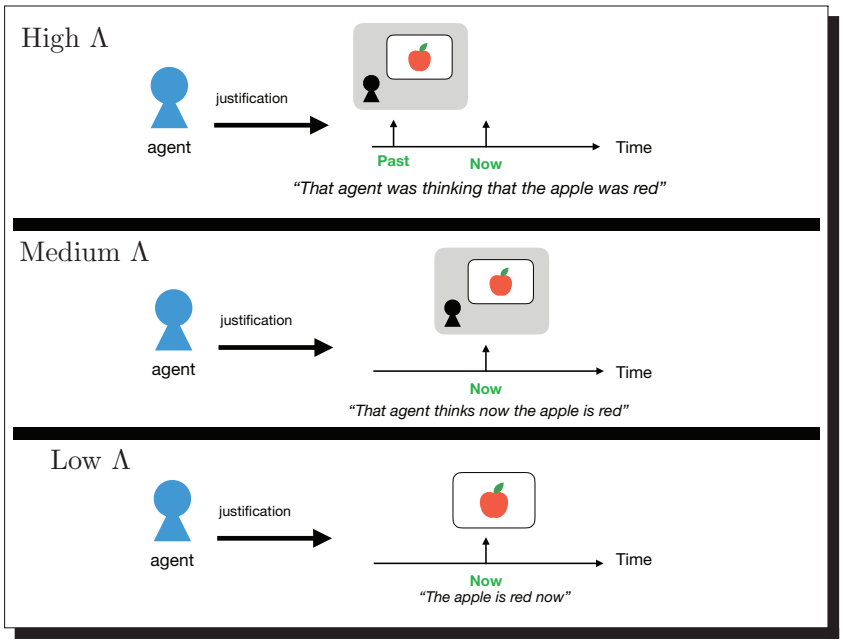


Figure 24.1: Λ and Justifications. We receive higher Λ values when the agent measured must consider other agents and handle richer temporal structures, as depicted here.

we could have the following mapping:

$$\Lambda[j] = \langle \lambda_{\mathbf{B}} = 3, \lambda_{\mathbf{D}} = 0, \lambda_{\mathbf{I}} = 0, \lambda_{\mathbf{K}} = 0, \lambda_{\overrightarrow{t}} = 0, \lambda_{\forall} = 0, \lambda_{\exists} = 0 \ldots \rangle$$

5.3 Λ and Cognitive Intelligence

One prominent drawback of IIT is that the theory has no linkage with any formal theory of computational intelligence[74] or for that matter any computational hierarchy used to consider intelligence in an abstract way, such as the Arithmetic Hierarchy.[75] Λ rectifies this situation immediately and decisively, since, by construction, Λ's measures reflect the depth of different components of cognitive intelligence, which bijectively correspond to cognitive consciousness. The basic intuition is simple: it's that any system that lacks rich cognitive structures is one that we wouldn't think of as being intelligent.

Consider for instance a web-page ranking algorithm such as the familiar PageRank. PageRank computes search results by considering how pages link to each other. It is possible to build two versions of a search engine using PageRank; the first version could have high Φ, while the second version could have low or zero Φ; the building of this pair would be in keeping with what we showed above. While such an algorithm is useful, one wouldn't seriously consider any implementation of PageRank to possess any amount of intelligence. In a match with our basic intuition, Λ will assign zero values to all the individual λ components in this case. On other hand, Λ will assign high scores to artificial characters, and even simulations of fictional characters that have rich cognitive lives; this outcome mirrors our intuition that such creatures ought to be regarded to possess

substantive cognitive consciousness, and, correspondingly, high cognitive intelligence.

5.4 TCC/Λ and AI of Today

Before concluding the present chapter in our next and final section, we make a few brief remarks about the relationship between TCC/Λ and the field of AI as it stands today. Because we have built out from TCC/Λ into systematizing cognitive intelligence, the most important part of the TCC/Λ–AI relationship, as we explain, centers around artificial *general* intelligence, or as it's commonly known, AGI.[76]

To begin, we simply report that we are under no illusion that quite soon the majority of AI researchers and engineers will begin to use Λ in order to assess the levels of c-consciousness and cognitive intelligence in the artificial agents they design and build. (We certainly recommend and hope that this happens.) But nonetheless, as a matter of fact, the current intellectual landscape is at least tacitly fertile ground for Λ, it seems to us. Why? The reason for our optimism in this regard pertains to the fact that a crucial distinction has been explicitly (albeit informally)[77] made by researchers between AI *simpliciter* versus AGI.

The start of any reasonable characterization of AGI, which may be new to some of our readers, probably consists in simply taking note of the way AI of the standard sort is defined in the dominant textbooks for the field of AI. By far the most influential such volume in the world today is the recently released fourth edition of Stuart J. Russell and Peter Norvig's *Artificial Intelligence: A Modern Approach*, what is by any reasonable metric a massive tome.[78] All four editions have been clear as can be in holding that AI is the field devoted to designing, implementing, and analyzing *artificial agents*.

And what is such an agent in this framework? It is a thing that maps percepts of its environment to actions performed in that environment, where the mapping is carried out by computation. This account has the immediate consequence that a simple, efficient computer program π which computes, say, the factorial function on the natural numbers qualifies as an artificial agent. But surely any sense in which such an agent is intelligent must be subjected to scrutiny. The reason is that printing out 6,227,020,800 after having perceived 13 (and so on for many other pairs in the graph of the factorial function) doesn't exactly seem sufficient to warrant ascriptions of "intelligent" to the program in question. At the very least, it would seem that, relatively speaking, π isn't all that intelligent. As a matter of fact, Λ applied to π yields zero. The reason, as the reader will have already grasped (and in fact saw in our example of the algorithm PageRank, given above), is that π doesn't have any c-consciousness at all. And the reason for this, in turn, is that π, for example, has absolutely no epistemic attitudes (in fact, no cognitive attitudes of any sort) that target any declarative formulae whatsoever. And when there is no c-consciousness there is no cognitive intelligence either. Likewise, Λ applied to chatbots like ChatGPT that are all the rage as we write will return a big fat zero, and thus such artificial agents have no cognitive consciousness at all.

But AGI leads directly to a different situation. To see this, let us ask: What sort of artificial-and-intelligent agents do people in AGI aim at? There is no consensus answer to this question. In addition, given our space constraints, we certainly cannot present and adjudicate competing characterizations of AGI. Our solution in the present context is to simply rely upon a nice characterization of AGI that among competitors seems to be the most cogent and ecumenical available: viz., Pei Wang's "On Defin-

ing Artificial Intelligence."[79]

For our purposes, we can focus on a key *sine qua non* for AGI of any level in an artificial agent, according to Wang: viz., that such an agent have general-purpose problem-solving capability, where the problems are at least as difficult as those we issue to human agents as a matter of routine course in our technologized world.[80] Given this requirement, it follows that an agent with AGI must have and exploit many cognitive attitudes. This can be immediately seen by considering tests of general problem-solving given to human agents in order to determine that they are maturing intellectually at an adequate pace. A very nice example is the so-called "false belief task" (FBT) which children over the age of five can solve, but which younger children generally can't.[81]

In FBT, we ask the agent $a^\star$ to be assigned to watch the following activity unfold among three other agents, a_1, a_2, and a_3 in a room: Agent a_3 places an object o into the first of two cardboard boxes b_1 and b_2 upon a table in plain view of all three agents, and then puts a top on this box b_1. Next, agent a_2 leaves and goes to a remote location from which no activity in the room can be seen. With a_2 gone, a_3 moves o into the other box b_2. Then agent a_2 returns to the room. Now $a^\star$ is asked this question (by the experimenter/tester): "If a_3 asks a_2 to retrieve o, which box will a_2 open first to do that?" Children with enough cognitive intelligence reply with "b_1," but younger children with insufficient cognitive intelligence say "b_2," which is of course incorrect.

The cognitive intelligence possessed by the older FBT-passing humans is directly relevant to TCC/Λ. The reason should be clear: it is that such humans have beliefs about the beliefs of other agents. This directly entails that such humans are in c-conscious states. And it certainly seems to be the case that the capacity to enter into such states is what enables intelligent responses to questions issued by the experimenter in FBT. Moreover, as the complexity of FBT is allowed to increase, the level of Λ required to give intelligent responses grows. For instance, in the *second-order* FBT, the agent a_2, in his/her remote location, can watch a video monitor secretly fed by a camera back in the room holding the boxes. This allows a_2 to covertly see that a_3 moves o into the other box b_2. All of this is taken in by the subject, that is by $a^\star$. If $a^\star$ is sufficiently cognitively intelligent, and is asked the same question, the correct answer is now no longer b_1, but b_2. Of course, $a^\star$ can readily provide a justification for the correct answer—and at least AGI researchers would be likely to request such a justification.

The upshot is simply that on the assumption that the current distinction between AI vs. AGI is real and sensible, TCC/Λ are quite relevant to this distinction, and in particular quite suitable as a formal explanation of AGI. As the AI-vs-AGI distinction grows in importance, and as artificial agents with only narrow and non-cognitive intelligence continue to fail when faced with the nuances of the real world, we believe that TCC/Λ will correspondingly grow in importance.

Finally, we claim that the relevant instantiation of the explanation template is at this point extremely plausible, which is to say we claim that

($2'^c$) E_{TCC} **Explains** $\Delta(c\text{-}consciousness)$ and $M_{E_{\text{TCC}}}[\Delta(c\text{-}consciousness)] = \Lambda_c$

and expect our rational readers to affirm this proposition.

6. Conclusion

To SUM up, at least in our view (and, needless to say, in the view too of Searle and

other IIT skeptics—and hopefully also now in your view), minimally the proposition (2^{p}_{IIT}), which expresses that IIT scientifically explains p-consciousness, should meet with agnosticism in the minds of those systematically and objectively seeking a scientific explanation of p-consciousness.[82]

Realistically, we are inclined to believe that IIT and Φ, at least in the form of variants, will continue to arrive on the scene, and on the strength of the unfolding of such a future, IIT/Φ will survive, in the sense of being considered credible by at least a remnant of cognitive scientists and AI engineers. At the same time, we are equally confident that many such scientists and engineers, including (in the second of these groups) roboticists, will pursue construction of artificial agents (including robots) that are c-conscious, and have high levels of Λ, and thereby high levels of cognitive intelligence—and this focus will be firm, undying, and energetic, with not the slightest concern for whether or not these artificial agents are such that it's something to be them. In the other words we introduced above, these engineers will concern themselves not a bit with whether their creations are p-conscious, but only with whether and to what degree these creations are c-conscious. As to whether what we envision will materialize, only time, of course, will tell.

7. Appendix: A Deeper Dive into IIT

In order to simplify our characterization, we rely on Aaronson's compressed version of IIT and Φ.[83] While there is an updated version of IIT, we note that attacks upon this theory considered in the present chapter are not thwarted by the newer version of IIT.

The informational content of any computing device θ (e.g., our robots High and Low) can be described in terms of a text of ones and zeroes (a finite string of bits). Such devices belong to a class of systems that computer scientists term *discrete finite systems*. A discrete finite system is made up of a finite number of components $\{c_1, c_2, \ldots, c_n\}$. Almost all modern computational systems can be viewed as discrete finite systems. To some extent, biological systems can also be approximated by such systems. For these reasons, discrete finite systems are a reasonably general starting point for many kinds of analysis. In the version presented here, IIT seeks to measure the p-consciousness of discrete finite systems that evolve over discrete time steps.

Consider a discrete finite system θ. At time t, its state is represented by $\theta(t)$. The state $\theta(t)$ is fully described by a finite string of bits $\sigma_{\theta(t)}$ of size n. Such strings are represented by the notation $\langle s_1, s_2, \ldots, s_n \rangle$ where each $s_i \in \{0, 1\}$. The system undergoes state changes specified by a function f operating on binary strings:

$$\theta(t+1) = f(\sigma_{\theta(t)})$$

f can be a non-deterministic function. IIT seeks to measure consciousness present in θ using the quantitative scalar measure $\Phi(\theta)$. A starting example is shown here:

Example

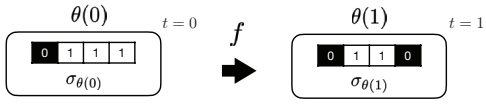
Take a system with four bits ($n = 4$), and let

$$f(\langle s_1, s_2, s_3, s_4 \rangle) = \langle s_1 \wedge s_2, s_2 \wedge s_3, s_3 \wedge s_4, s_4 \wedge s_1 \rangle$$

The symbol $\wedge$ represents the "and" operation:

$x \wedge y = 1$ if and only if both $x = 1$ and $y = 1$

Given an initial state of $\langle 0, 1, 1, 1 \rangle$, under f, the system would change to $\langle 0, 1, 1, 0 \rangle$.



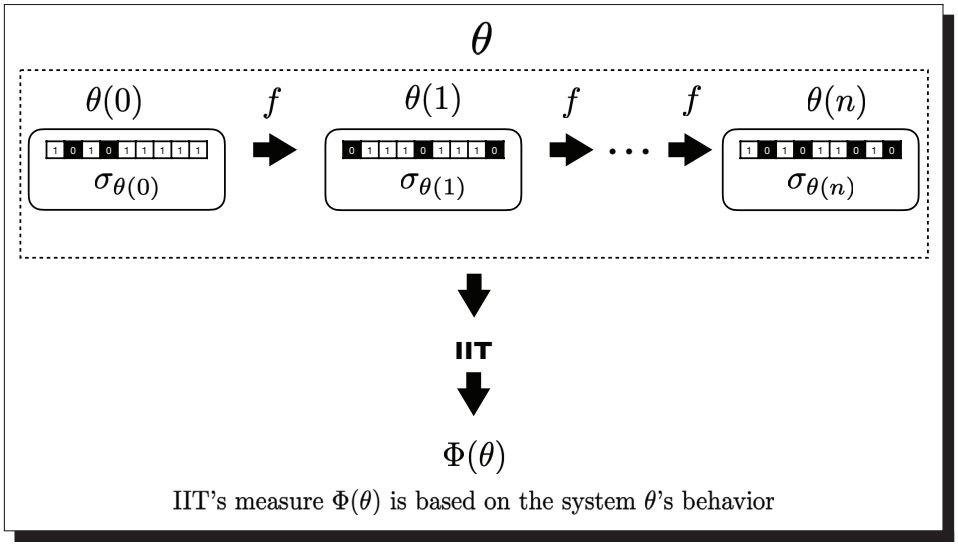


Figure 24.2: An Evolving-System Perspective. Aaronson presents IIT as seeking to measure p-consciousness of systems which evolve under an update function. The state of the system is fully specified by a finite fixed-length string.

IIT seeks to measure the level of connectedness of different components in a given system. Since a system is fully described by a bit string, different subsystems correspond to different parts of the string. A *non-empty partition* of a bit string σ divides the string into two parts A and B. These partitions correspond to two different components. E.g., if σ is $\langle s_1, s_2, s_3, s_4, s_5 \rangle$, one possible non-empty partition is $A = \langle s_1, s_3 \rangle$ and $B = \langle s_2, s_4, s_5 \rangle$.

Given a state string σ, its *Shannon entropy*, $E(\sigma)$, measures uncertainty or lack of knowledge in the possible values the string can have. If the string is always fixed at a certain sequence of bits, then its Shannon entropy is $E(\sigma) = 0$. Shannon entropy is maximized when the string can take on all possible values with equal probability; in this case, we would have no prior knowledge of what value the string might have. We can also measure the Shannon entropy of non-empty partitions A and B of σ.

For any non-empty partition of the state string σ into two partitions (A, B), we define the *effective information* $EI(A \rightarrow B)$ as the Shannon entropy of B if the bits in A are drawn uniformly at random with bits in B kept fixed at their input values. This quantity measures how much impact current values of A can have on future values of B, and measures the connectedness of the two partitions. See Figure 24.3. We then define $\phi(A, B)$ as:

$$\phi(A, B) := EI(A \rightarrow B) + EI(B \rightarrow A).$$

The definition above looks at only one partition (A, B) of the system. To arrive at a connectedness measure for the whole system, which is crucial for IIT, we look at all partitions and take the partition which has the minimum value for $\phi(A, B)$. Any system which is highly interconnected will have high $\phi(A, B)$ between all

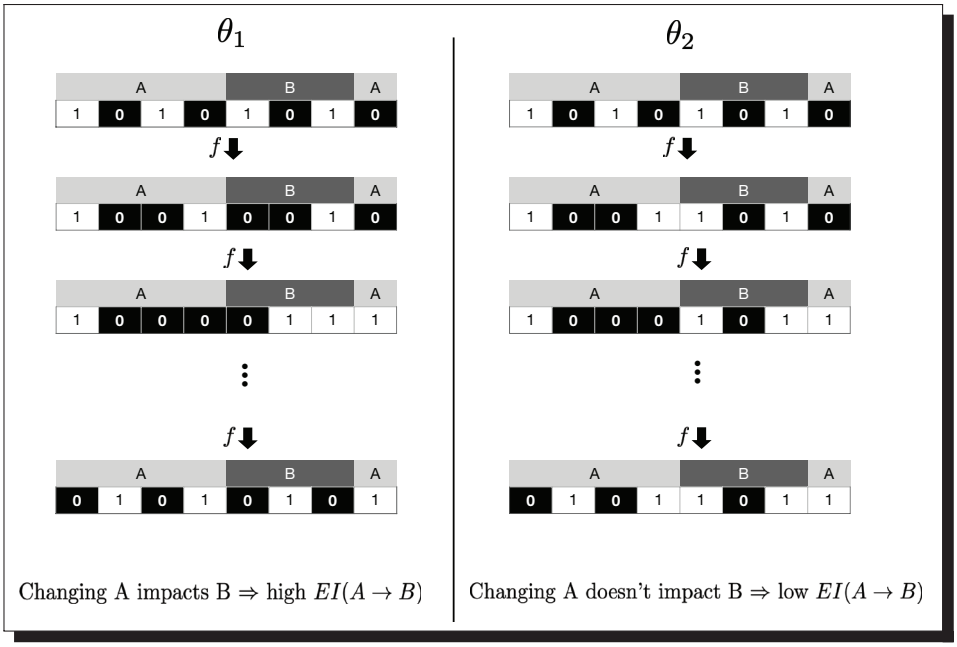


Figure 24.3: Effective Information. Effective information seeks to measure how much the changing of one subsystem (A) affects another subsystem (B). On the left side, in system θ_1, we have high effective information, since changing A changes B significantly. On the right side, in system θ_2, changing A does not lead to any changes in B, giving us low effective information.

partitions. Finally, we need to account for the size of partitions. The maximum entropy of a string is limited by the size of the string. When one partition is very small, the maximum possible $\phi(A, B)$ will also be limited. To handle this, we divide $\phi(A, B)$ by the size of the smaller partition ($|A|$ gives us the size of A):

$$\Phi(\sigma) := \underset{\phi(A,B)}{\text{argmax}} \frac{\phi(A, B)}{min(|A|, |B|)}$$

Intuitively put, effective information measures how changes in one part of a system impact a different part of the system; in short, it measures information on a "global" scale. (At this juncture, the parable of robots Low and High hopefully makes further sense to the reader.) The *central thesis* of IIT can now be stated:

Central Thesis of IIT

The quantity $\Phi(\sigma)$, labeled "**integrated information,**" measures the p-consciousness of σ.

IIT does not present a formal proof for its central thesis. The various arguments in IIT literature in support of the central thesis are markedly informal, with a mixture of empirical results sprinkled in. The central thesis implies that high values of Φ are necessary and sufficient for high p-consciousness. While arguments against IIT generally attack the central thesis, there are some arguments, such as Searle's attack, that fall outside of this but present another dimension of attack against IIT.

Notes

1. The authors are deeply grateful to numerous anonymous reviewers for their time and trenchant feedback; to Brian Krouse for his wise guidance (and preternatural patience); and, in Bringsjord's case, to the late AI founder John McCarthy for expressing in person, first in 1991, that the internal structure (and content therein) of artificial agents is sufficient for the type of consciousness that will increasingly take root on earth as the field of AI progresses. The authors thank AFOSR for funding to explore the nature of computational intelligence and consciousness; this funding was crucial to the development of TCC and Λ. Support over a summer from SRI was also invaluable, and we are thankful.
2. See Kevin Roose, "Help, Bing Won't Stop Declaring Its Love for Me," *New York Times*, February 16, 2023.
3. Why is "formal" better? It's better for two reasons: (1) Mathematics is technically smaller than formal logic, as it can be obtained from specific axioms expressed in some relatively straightforward logics that are a proper subset of the space of all logics. For the axioms in question, expressed in (so-called) third-, second-, and first-order logic, see Stephen G. Simpson, *Subsystems of Second Order Arithmetic*, 2nd ed. (Cambridge, UK: Cambridge University Press, 2010.) By "formal" we denote *both* mathematics and logic. Some readers might like to see us use "formal logic" rather than just "logic," but we are of the opinion that so-called "informal" logic is just based on formal logic anyway. An overview of informal logic is provided by Leo Groarke, "Informal Logic," *Stanford Encyclopedia of Philosophy* (Winter 2022), https://plato.stanford.edu/entries/logic-informal. Prior to that essay Groarke provides a way to understand any informal logic as an instantiation of a certain quintuple of elements: Leo Groarke, "How to Define an Informal Logic," in J. Anthony Blair and Christopher W. Tindale, eds., *Rigour and Reason: Essays in Honour of Hans Vilhelm Hansen* (Ontario, Canada: University of Windsor, 2020), 231–251. Each of these elements can be directly constituted in all their variations from the established resources of formal logic. The second reason why "formal" is wiser than "mathematical": (2) There are a number of disciplines not classified as being within mathematics that are most assuredly "in on the game" of trying to explain the sort of mental phenomena targeted in the present book, and targeted specifically by us—yet these disciplines can be accurately said to be in the "formal sciences," while they cannot be uncontroversially said to be part of mathematics. One example is game theory; another is decision theory. We count all such disciplines as falling under our umbrella term of "formal."
4. Philosophers since at least 1996 have also often called the attempt to make sense of phenomenal consciousness *the hard problem*. See David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford, UK: Oxford University Press, 1996).
5. Ned Block, "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences* 18, no. 2 (1995): 227–247.
6. Just as Φ is the majuscule (uppercase) version of ϕ in Greek (pronounced "fie" or "fee"), Λ is the uppercase Greek letter for λ, and is pronounced "lamduh."
7. Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi, "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," *PLoS Computational Biology* 10, no. 5 (May 8, 2014): 1–25, https://pubmed.ncbi.nlm.nih.gov/24811198/.
8. Christof Koch and Giulio Tononi, Consciousness: Confessions of a Romantic Reductionist (Cambridge, MA: MIT Press, 2012).
9. As we shall see, this will be the axiom (Integ).
10. See, for example, John McCarthy, "Making Robots Conscious of Their Mental States," *AAAI Technical Report* SS-95-05 (1995), https://cdn.aaai.org/Symposia/Spring/1995/SS-95-05/SS95-05-013.pdf. McCarthy continued to refine this paper through the years, from at least 1995 to 2002. A later version of the paper can be found at http://jmc.stanford.edu/articles/consciousness/consciousness.
11. Selmer Bringsjord, Paul Bello, and Naveen Sundar Govindarajulu, "Toward Axiomatizing Consciousness," in Dale Jacquette, ed., *The Bloomsbury Companion to the Philosophy of Consciousness* (London: Bloomsbury Academic, 2018), 289–324. Selmer Bringsjord and Naveen Sundar Govindarajulu, "The Theory of Cognitive

Consciousness, and Λ (Lambda)," *Journal of Artificial Intelligence and Consciousness* 7, no. 2 (2020): 155–181, https://www.worldscientific.com/doi/abs/10.1142/S2705078520500095.

12. For good measure, we mention here that McCarthy in no way stopped at ascribing consciousness to AIs/robots on the strength of their having the requisite internal logic-based processes and structures. E.g., he ascribed "free will" to suitably logic-based artificial agents; and this ascription was avowedly and painstakingly made only on the strength of whether the relevant internal structures in these agents have a certain logical nature. See John McCarthy, "Free Will—Even For Robots," *Journal of Experimental and Theoretical Artificial Intelligence* 12, no. 3 (2000): 341–352.
13. While we write "for us," this view of science is of course perfectly standard. Interested readers unfamiliar with philosophy of science and wishing a deeper presentation of our view can consult a classic presentation of philosophy of science given in an explanation-centric manner. See, for example, Ernest Nagel, *The Structure of Science: Problems in the Logic of Scientific Explanation* [1961], 2nd ed. (Indianapolis, IN: Hackett, 1979).
14. Notice we refer to "observational data." We do *not* simply say that things are "observed." We refer in a moment to macroscopic objects/phenomena, which by any familiar sense of "observe" can be observed. However, when phenomena to be scientifically explained involve things that are either very small or very big (think of quantum mechanics and general relativity), direct observation is unattainable. But there is still, if science is being brought to bear in search of explanation, observational data.
15. For needed economy, we are not concerned herein with the history of science, and note only that appreciable robust mathematics, replete with rigorous proof, is found in Euclid, and a *bona fide* (albeit limited, by modern metrics) formal logic is specified by Aristotle in his *Organon*. See Robin Smith, "Aristotle's Logic," *Stanford Encyclopedia of Philosophy* (Winter 2022), https://plato.stanford.edu/entries/aristotle-logic. And Aristotle firmly held that scientific knowledge had to be expressed in and obtained by way of his syllogistic logic. See 71b in his *Analytica Posteriora* in Richard McKeon, ed., *The Basic Works of Aristotle* (New York: Random House, 1941). For a lucid history of systematic thought from Euclid to Frege that revolves around formal logic, see Clark Glymour, *Thinking Things Through* (Cambridge, MA: MIT Press, 1992).
16. Readers who wish a more formal framework for what scientists do through time to hypothesize regarding, and gradually understand, "hidden" phenomena, are directed to an excellent text we have used to teach such matters: Sanjay Jain, Daniel Osherson, James Royer, and Arun Sharma, *Systems That Learn: An Introduction to Learning Theory*, 2nd ed. (Cambridge, MA: MIT Press, 1999). This book presents a paradigm rooted in formal logic that constitutes a science of science, and in particular provides a rigorous way to understand a significant part of what a scientist does when, as inquiry unfolds through time, she offers candidate explanations for what she observes.
17. The full title, rarely used, is *Philosophiæ Naturalis Principia Mathematica* (Mathematical Principles of Natural Philosophy). Even the very first edition, with Newton's annotations, is available online upon a brief search.
18. Newton himself, as is well known, was a rather maniacal observer of phenomena. The classic (but now somewhat dated) biography of Newton makes this clear. See Richard S. Westfall, *Never at Rest: A Biography of Isaac Newton* (1980), 5th ed. (Cambridge, UK: Cambridge University Press, 1983). Note that without question Newton passionately and indefatigably searched for formal explanations, happy to leave aside "ultimate causes" at a more informal level, a crystal-clear case in point being gravity.
19. J. C. C. McKinsey, A. C. Sugar, and Patrick Suppes, "Axiomatic Foundations of Classical Particle Mechanics," *Journal of Rational Mechanics and Analysis* 2 (1953): 253–272.
20. Details are beyond our scope here, but as readers will likely recall, in the Newtonian case of (1), an axiomatization of gravity (E_{Newton} in (1)) explained not only the motion of a falling apple, but the planets too, since according to the axioms, the gravitational force between any two objects is proportional to the product of their masses, and inversely proportional to the square of the distance between them, which in turn deductively entails Kepler's description of planetary motion (e.g. that planets move around the sun in elliptical orbits). Measurement that yields that location of a planet at time t, expressed in keeping with the language of E_{Newton}, combined with the

axiomatization itself, yields μ, viz., the relevant subsequent location of the planet at later time t'. Of course, what we describe here not only also works for falling apples, but works for predicting the subsequent location of rockets and spaceships, and is fundamentally why sending humans to the moon in 1969 was a Newtonian affair.

21. Note that Δ, as we have said, is *observational* data. This means it's third-person data. What, then, about first-person information? E.g., what about your feeling fear? Or, to harken back to §1, what about your finding in that complex Brunello a certain cherry-sugared sweetness? Unless this is expressed in third-person terms (when you for instance express to a neurologist that you are afraid while your brain is in some manner being scanned), the information isn't part of Δ.
22. Interested readers can start by consulting Hajnal Andréka, Judit X. Madarász, István Németi, and Gergely Székely, "A Logic Road from Special Relativity to General Relativity," *Synthese* (2011): 1–17, http://dx.doi.org/10.1007/s11229-011-9914-8. Note that everyday life in the twenty-first century provides a straightforward example of how measurement is tied to explanation by axiom systems for science, the reason being that clocks aboard GPS satellites need to be reset every single day, as entailed by the axioms of relativity theory. Here, the time of such a clock is the measurement μ, and that measurement is predicted by relativity theory, formalized axiomatically. A wonderful exposition of this situation is provided by Gergely Székely, "New Challenges in the Axiomatization of Relativity Theory," in Á. Poroszlai, G. Poroszlai, Z. Petrák, eds., *Proceedings of the New Challenges in the Field of Military Sciences* (Budapest: Bolyai János Military Foundation, 2010), https://users.renyi.hu/ turms/NewChallenges2010.pdf.
23. Block, "On a Confusion."
24. This is not to say that there isn't a highly cognitive side to cricket. There is; see for example Mike Brearley, *The Art of Captaincy: The Principles of Leadership in Sport and Business* (London: Pan Macmillan, 2015). Hence, as will shortly be seen, our theory of consciousness, the cognitive theory of consciousness (TCC), or *c-consciousness*, is also embodied in the playing of serious cricket.
25. Selmer Bringsjord, "Consciousness by the Lights of Logic and Common Sense," *Behavioral and Brain Sciences* 20, no. 1 (1997): 144–146.
26. Block, "On a Confusion," 231.
27. Bringsjord, "Consciousness by the Lights of Logic and Common Sense."
28. Selmer Bringsjord, "A Refutation of Searle on Bostrom (re: Malicious Machines) and Floridi (re: Information)," *Newsletter on Philosophy and Computers* 15, no. 1 (2015): 7–9. *Newsletter on Philosophy and Computers* is published by the American Philosophical Association.
29. The original proposal for the conference, a profound and profoundly illuminating read even today, can be found here: John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence" [August 31, 1955], *AI Magazine* 27, no. 4 (Winter 2006), https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904.
30. Allen Newell and Herbert A. Simon, "The Logic Theory Machine: A Complex Information Processing System," *P-868 The RAND Corporation*, 25–63. An almost exact version of this paper can be found in *IRE Transactions on Information Theory* 2, no. 3 (September 1956): 61–79, https://ieeexplore.ieee.org/document/1056797.
31. To again cite a key paper in this regard, see McCarthy, "Making Robots Conscious of Their Mental States." An overview and defense of logic-based AI can be found in Selmer Bringsjord, "The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself," *Journal of Applied Logic* 6, no. 4 (2008): 502–525, http://kryten.mm.rpi.edu/SB LAI Manifesto 091808.pdf.
32. See the cognitive verbs that anchor a number of the chapters in the authoritative M. Ashcraft and G. Radvansky, *Cognition*, 6th ed. (London: Pearson, 2013).
33. These states can be defined ostensively, given what we have said earlier in the present paragraph, by way of a few examples of them given in connection with some arbitrary (neurologically normal, mature, educated) human: e.g., *Alvin's believing that Beatrice believes that Alvin is an opera lover*; *Charlie's fearing that bears are hibernating in his basement*, *Doreen's knowing that she knows that the combination is known to Frank*, and so on. We here follow the tradition of referring to states (or, as they are sometimes called, *states-of-affairs*) using gerundive nominals.

34. Leaving aside for simplification the mind of God, which is purportedly, at least to a degree, known to a number of scientists.
35. We shall not spend the considerable time that would be needed to list all the axioms and explain them. For nice coverage of **PA** (and illuminating commentary on this axiom system) readers can consult the elegant Heinz-Dieter Ebbinghaus, Jörg Flum, and Wolfgang Thomas, *Mathematical Logic*, 2nd ed. (New York: SpringerVerlag, 1994). There are theories of arithmetic even simpler than **PA**, because **PA** includes an axiom relating to mathematical induction, and the simpler systems leave this axiom aside. For example, readers unfamiliar with mathematical induction can, if motivated, consult the induction-free theory of arithmetic known as "Robinson Arithmetic," or sometimes just as "Q." For elegant coverage, see George S. Boolos, John P. Burgess, and Richard C. Jeffrey, *Computability and Logic* (Cambridge, UK: Cambridge University Press, 2003).
36. A wonderful example of explanation of kinematics in special relativity (and we indirectly alluded to this above) can be obtained by the axiom system **SpecRel**. For a truly wonderful overview, see Hajnal Andréka, Judit X. Madarász, and István Németi, "Logical Axiomatizations of Space-Time. Samples From the Literature," in András Prékopa and Emil Molnár, eds., *Non-Euclidean Geometries: Janos Bolyai Memorial Volume, Mathematics and Its Applications* 581 (New York: Springer, 2006): 151–185.
37. It's (in our opinion) fun to find the answers yourself. Here's one answer for you: $13 + 15 + 17 + 19 = 64 = 4^3$. Can you find the other four for the list we just gave? From **PA** it can be proved that, in fact, *every* positive cubic number n^3 is a sum of some n consecutive odd numbers! (This was first proved, apparently, by Nicomachus.) Hence, **PA** explains the phenomenon in question.
38. For the definitive account of how essentially all of mathematics can be obtained from a group of axiom systems (one group being **PA**), see Simpson, *Subsystems of Second Order Arithmetic*. The author here is the leading authority on what is known as *reverse mathematics*.
39. Giulio Tononi, Melanie Boly, Marcello Massamini,and Christof Koch, "Integrated Information Theory: From Consciousness to its Physical Substrate," *Nature Reviews of Neuroscience* 17 (2016): 450.
40. Thus, more precisely, we would have this emendation:

    $(2^{p'}_{\text{IIT}})$ *Postulates*$_{\text{IIT}}$ **Explains** Δ(*p-consciousness*) and $\Phi_{\textit{Postulates}_{\text{IIT}}}[\Delta(\textit{p-consciousness})] = \mu_p$

41. See for example Tononi, Boly, Massamini and Koch, "Integrated Information Theory"; and Giulio Tononi and Christof Koch, "Consciousness: Here, There and Everywhere?," *Philosophical Transactions of the Royal Society* B: Biological Sciences 370 (2015): 1–18.
42. Tononi and Koch, "Consciousness: Here, There and Everywhere?," 6. Italics in original.
43. Tononi et al., "Integrated Information Theory," 452.
44. See, for example, the balanced and patient guesses given by Tim Bayne, "On the Axiomatic Foundations of the Integrated Information Theory of Consciousness," *Neuroscience of Consciousness* 2018, no. 1 (2018), https://doi.org/10.1093/nc/niy007.
45. A hyper-charitable formal version of the axiom, given here for cognoscenti and for Tononi, Koch, and other IIT adherents, is possible using the kind of machinery that e.g. mathematical physicists have used (recall our comments about **SpecRel** above). This machinery is simply a background, primitive vocabulary for the overall axiom system we dub **AITT**. We charitably stipulate that this vocabulary has—in each case with or without subscripts, as needed—(i) the symbol $\mathfrak{a}$ for agents that are the bearers of experiences; (ii) the symbol e for "experiences;" (iii) a supply of symbols R to indicate properties experiences can have, and when we write the atomic formula-form "$R(e)$" that means that e has property R; (iv) the symbol t for timepoints, with an interval of time from t_1 to t_2 denoted by $(t_1 - t_2)$; and (v) the atomic-formula form $Had[\mathfrak{a}, e, (t, t')]$ for saying that the agent here has the experience over the interval shown. And now the logicization of (Integ), in second-order logic, is this:

    $$\forall e\, \exists t_1, t_2\, \exists^{=1}\mathfrak{a}\, \{Had[\mathfrak{a}, e, (t, t')] \wedge \forall R\, (R(e, t_1, t_2) \rightarrow \mathbf{P}(\mathfrak{a}, (t_1, t_2), R))\}$$

46. David Hume, *A Treatise of Human Nature* [1738] (Amazon Digital Services: Kindle, 2011).
47. Bayne, "On the Axiomatic Foundations."
48. It's nonetheless true that plenty of axioms in mathematics are objectionable to some skeptics. The classic example is the Axiom of Choice in Zermelo-Fraenkel axiomatic set theory. For coverage, see e.g. the classic and still perfectly accurate Patrick Suppes, *Axiomatic Set Theory* (New York: Dover, 1972).

49. Oizumi, Albantakis, and Tononi, "From the Phenomenology to the Mechanisms of Consciousness," 4. Emphasis ours.
50. We thus have in our parable not a bivalent framework restricted to only TRUE and FALSE for the value of propositional variables, but a *tri*valent framework of a type long ago (circa 1938) introduced by Kleene (and others independently), a subsequent overview of which can be found here: Stephen Cole Kleene, *Mathematical Logic* (New York: Wiley & Sons, 1967). We recommend a 2002 Dover unabridged republication of the original 1967 book from Wiley, if you cannot obtain the original book.
51. The details of how such engineering can be achieved is outside of scope for the parable, but such a thing is quite feasible.
52. We find it very tempting, in the light thrown by our parable, to assert that IIT/Φ have in many ways been anticipated by the theory of consciousness advanced near the middle of the twentieth century by Julian Jaynes, according to whom consciousness in the human case (to put matters barbarically) emerged only when integration between the two halves of our minds was achieved by a certain kind of socialization. Robot Low is composed of systems that are not in any way unified by a "self." See Julian Jaynes, *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (New York: Houghton Mifflin, 1976).
53. Tononi has attempted to provide an intuitive, artistic overview of Φ, in a fascinating, aesthetically pleasing little book: *Phi: A Voyage from the Brain to the Soul*. It is in fact much more intuitive than our parable about Low and High; indeed, overall, it's an attempt at a sort of historico-philosophical poetry. Each chapter is a burst of philosophical fiction featuring characters loosely based on famous scientists of the past—Galileo, Darwin, Turing, and so on; and then comes for each chapter a wrap-up in which Tononi explains things in a more professorial mode. See Giulio Tononi, *Phi: A Voyage from the Brain to the Soul* (New York: Pantheon, 2012).
54. And if he's right, by extension, Φ would of course fall as well.
55. John Searle, "Can Information Theory Explain Consciousness?," *New York Review of Books* (January 10, 2013): 54–58, http://www.nybooks.com/articles/2013/01/10/can-information-theory-explain-consciousness. This is the online version of Searle's original review, composed of seven sections. Next came Christof Koch and Giulio Tononi, reply by John Searle, "Can a Photodiode be Conscious?," *The New York Review of Books*, March 7, 2013, https://www.nybooks.com/articles/2013/03/07/can-photodiode-be-conscious/. This is an exchange between the three, in which Searle is given the last word.
56. We of course haven't presented IIT in *full* technical detail, but certainly, it can be said, have supplied at least a "Scientific-American"-level presentation of the theory.
57. John Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992).
58. Some readers may be curious as to what specific things correspond to μ, the measurement, appearing in the instantiation of the template given here. We ask these readers to consider the number-theoretic observation we cited above: viz., that every cubic number c is equal to a sum of (finite) consecutive odd numbers $o_1, o_2, \ldots, o_m$. In this case, μ is the measurement confirming that

$$o_1 + o_2 + \ldots + o_m = c.$$

59. Here used an an adjective.
60. We are indebted to a rather insightful (and charitable-to-Searle?) reviewer for anticipating the reply here given on Searle's behalf.
61. If Searle persists, he would need to take on the body of work explaining that amazing correspondence between math/logic and the physical world (Eugene P. Wigner, "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," *Communications in Pure and Applied Mathematics* 13, no. 1 (February 1960): 1–14), and the parallel stunning correspondence between formal logic specifically and computation (Joseph Halpern, Robert Harper, Neil Immerman, Phokion Kolaitis, Moshe Vardi, and Victor Vianu, "On the Unusual Effectiveness of Logic in Computer Science," *The Bulletin of Symbolic Logic* 7, no. 2 (2001): 213–236).
62. Koch and Tononi, reply by Searle, "Can a Photodiode be Conscious?"
63. Searle, "Can Information Theory Explain Consciousness?"
64. An overview of which is presented in Hector Levesque and Gerhard Lakemeyer, "Chapter 24: Cognitive Robotics," in *Handbook of Knowledge Representation* (Amsterdam, The Netherlands: Elsevier, 2007), 869–882.

65. This is directly analogous to such things as "a theory of arithmetic," or "a theory of special relativity," etc. from our point of view.
66. In case of the axiom system **PA** we relied upon above, which gives the heart of a theory of (elementary) arithmetic, a sample meta-proposition would include "As far as we know from our storehouse of relevant theorems, **PA** is consistent."
67. The reader is directed first to the introduction of this axiomatization (and cognitive consciousness in general), provided in Bringsjord, Bello, and Govindarajulu, "Toward Axiomatizing Consciousness." Then, for a more detailed (and more technical) presentation of the axioms, which presents the axiom system $\mathcal{CA}$ in its expanded and more rigorous form: Bringsjord and Govindarajulu, "The Theory of Cognitive Consciousness, and Λ (Lambda)."
68. E.g., we perceive that we are in pain when we are.
69. E.g., we perceive creatures whose behavior indicates to us that they are in pain.
70. See Selmer Bringsjord, Naveen Sundar Govindarajulu, and Michael Giancola, "Automated Argument Adjudication to Solve Ethical Problems in Multi-Agent Environments," *Paladyn, Journal of Behavioral Robotics* 12 (2021): 310–335, https://www.degruyter.com/document/doi/10.1515/pjbr-2021-0009/html. The URL here goes to a rough, uncorrected, truncated preprint as of July 14, 2021.
71. Clarence Irving Lewis and Cooper Harold Langford, *Symbolic Logic* (New York: Century Company, 1932).
72. For a rationale, see Selmer Bringsjord and David Ferrucci, *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine* (Mahwah, NJ: Lawrence Erlbaum, 2000).
73. Stephen Wolfram, "What Is ChatGPT Doing... and Why Does It Work?," *Stephen Wolfram: Writings*, February 14, 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.
74. Such as Hutter's AIXI, which is such a theory. See Marcus Hutter, *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability* (New York: Springer, 2005).
75. Mentioned above, and explained and used in Selmer Bringsjord and Michael Zenzen, *Superminds: People Harness Hypercomputation, and More* (Dordrecht, The Netherlands: Kluwer Academic Publishers, 2003).
76. For recommendations regarding characterizations of AGI in the literature, and crisp summaries of these characterizations, we are greatly indebted to James Oswald.
77. Though see Naveen Sundar Govindarajulu, John Licato, and Selmer Bringsjord, "Toward a Formalization of QA Problem Classes," in B. Goertzel, L. Orseau, and J. Snaider, eds., *Artificial General Intelligence* (Basel, Switzerland: Springer, 2014), 228–233.
78. Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (New York: Pearson, 2020).
79. Pei Wang, "On Defining Artificial Intelligence," *Journal of Artificial General Intelligence* 10, no. 2 (2019): 1–37, https://doi.org/10.2478/jagi-2019-0002. We recognize that our readers may wish to study other accounts of AGI, and hence provide some pointers: Goertzel, long a pioneer in AGI research, places considerable emphasis upon the need for generalization capability in any AGI agent, which is compatible with our own emphasis in the present section on general-purpose problem-solving capability. See Ben Goertzel, "Artificial General Intelligence: Concept, State of the Art, and Future Prospects," *Journal of Artificial General Intelligence* 5, no. 1 (2014), 1–46. Hutter (recall that we referred above to this work) offers a rigorous account of "universal" AI, which might be thought of as an account of AGI—but unfortunately the account leaves out any notion of cognitive intelligence, including knowledge. See Hutter, *Universal Artificial Intelligence*. Finally, in a very nice paper that is generally in line with the paper by Wang we here rely upon, Voss provides an account of AGI that for specific technical reasons is beyond the scope of the present chapter but which we find very attractive (in a word, Voss insists that AGI requires an ability to reason through time in a way that allows its conclusions and hypotheses about the world at time t to change into different conclusions and hypotheses at a subsequent time t'). See Peter Voss, "Essentials of General Intelligence: The Direct Path to Artificial General Intelligence," in Ben Goertzel and Cassio Pennachin, eds., *Artificial General Intelligence* (Berlin, Germany: Springer, 2007): 131–157.
80. AGI cognoscenti (such as James Oswald, consultation with whom has greatly helped us in the case of the present chapter), will not be unjustified in

pointing out that while Wang does emphasize solving "general" problems, he doesn't emphasize *human-level* problems of this type.

81. FBT is sometimes referred to as the "Sally-Anne" task. For a definition and discussion of FBT in psychology/cognitive science, see D. Premack and G. Woodruff, "Does the Chimpanzee have a Theory of Mind?," *Behavioral and Brain Sciences* 4 (1978), 515–526. For a general-purpose formal-and-computational model of the task, one that makes it crystal clear that cognitive intelligence is needed to solve it, see Konstantine Arkoudas and Selmer Bringsjord, "Propositional Attitudes and Causation," *International Journal of Software and Informatics* 3, no. 1 (2009): 47–65, http://ijsi.cnjournals.com/ch/reader/create_pdf.aspx?file_no=32&flag=1&journal_id=ijsi&year_id=2009. For a more recent, elegant analysis and formal model of the task using hybrid logic, see Torben Braüner, "Hybrid-Logical Reasoning in the Smarties and Sally-Anne Tasks," *Journal of Logic, Language and Information* 23 (2014): 415–439.
82. Of course, again, we don't think that such an explanation is even conceptually possible, since it would by definition consist of a collection of third-person declarative assertions, expressed as formulae in a formal language, and this is something no scientist has any reason to think is possible for p-consciousness. In particular, as Bringsjord has explained, certainly no *AI* scientist has reason to think such a third-person scheme is both possible and implementable in a computing machine; see Selmer Bringsjord, "Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline," *Journal of Consciousness Studies* 14, no. 7 (2007): 28–43, http://kryten.mm.rpi.edu/jcsonebillion2.pdf.
83. Scott J. Aaronson, "Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)," *Shtetl-Optimized: The Blog of Scott Aaronson*, 2014, https://scottaaronson.blog/?p=1799.