The Integrated Information Theory of Consciousness (plus Φ), Rejected; And a Better Way Supplied: The Theory of Cognitive Consciousness (plus Λ)

Naveen Sundar Govindarajulu • Selmer Bringsjord

`ver 1003211130MA`

Can science explain consciousness? In the minds of many, the most exciting attempt now afoot to deliver a theory that entails a "Yes!" to this question is Integrated Information Theory (IIT) and its concomitant, a measure of the amount of consciousness in some thing: Φ. The originator of IIT/Φ is Giulio Tononi, who has now been joined by co-advocate Christof Koch. We summarize IIT/Φ, and then turn to attacks upon it from Searle and Scott Aaronson. Searle argues that any informational theory of consciousness 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, these are not observer-relative at all). A second Searlean argument, specifically in favor of rejecting IIT, is that since IIT entails panpsychism, IIT is false. A more circumspect, and more powerful, version of the second argument has been given by Scott Aaronson. We show that while these two thinkers don't succeed in overthrowing IIT/Φ, they pave the road to our refutation of IIT/Φ. We then present and defend a better way to explain a better way, one that can be pursued in concrete fashion: namely, focus on *cognitive* consciousness, on a system for measuring its level in objects (Lambda/Λ), instead of Φ), and on the engineering of artifacts that have high Λ, and as such are AIs of great promise.

Contents

1 The Scientific Challenge of Consciousness 1

1.1 The Overarching Aim of Science 1

1.2 What Then is the Science of Consciousness? 2

1.3 But What Kind of Consciousness? 2

2 Integrated Information Theory (IIT) *&* Phi (Φ) 4

2.1 Aaronson's Conveniently Compressed Version of IIT 4

3 Searle's Attacks on IIT/Φ 6

3.1 Attack #1: "IIT is observer-relative!" 6

3.2 Attack #2: "But IIT entails panpsychism!" 8

4 High Φ is Neither Necessary nor Sufficient for Consciousness 8

4.1 Necessary Condition: Unfolding Arguments 9

4.2 Sufficient Condition: Aaronson's Attack 10

5 Final Evaluation of Attacks on IIT/Φ 10

6 The Better Way: The Theory of Cognitive Consciousness, Λ, *&* Intelligence 11

6.0.1 Regarding the Axiom System ($\mathcal{CA}$) for Cognitive Consciousness 11

6.1 Cognitive Consciousness, Measured: Λ 12

6.2 Λ and Cognitive Intelligence 13

6.3 TCC/Λ and AI of Today 14

7 Conclusion 15

List of Figures

1 An Evolving-System Perspective 5

2 Effective Information 6

3 Metatheory Damaging to IIT 9

4 Polynomial-based Update Functions 11

5 Λ and Justifications 13

1 The Scientific Challenge of Consciousness

1.1 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[1] is to gradually supply, for a given phenomenon p for which there is observational data,[2] 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 are *declarative*, and hence the content is a collection of propositions (= statements).[3] 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;[4] doing so will pay dividends for us in the present chapter: 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. A pleasing case is classical mechanics in physics, 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*[5] in 1687. In this case, observations of the behavior of macroscopic objects are expressed in our $\Delta(p)$; i.e., this is the explanandum. 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 (McKinsey, Sugar & Suppes 1953). 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 import should be quite clear:

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

Proposition (1) asserts that the axioms of classical mechanics explain the observed phenomena. Of course, in the case of first-rate science, it will also generally be required that declarative content which does the explaining can (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. For ease of reference and some generality, we shall say that some measurement scheme M_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 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 the integrated information theory of consciousness (IIT), the role of its particular measurement scheme, one known as 'Φ,' is 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.

1.2 What Then is the Science of Consciousness?

So then, what about consciousness, our central concern herein? Well, the simple template we have created works just fine to help us make sense of the structure of our chief topic of concern in the present chapter. It would be premature at this point to refer in any detail to what theory of (some kind of) consciousness is in play, and we can hence leave things general via the form of the following schematic proposition:

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

Here, of course, the declarative content E_X does the explaining of observational data $\Delta(\mathit{consciousness})$, 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. The use of this abstract framework may be hard to see at work in some "soft" sciences (e.g., psychology), but it's 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.[6] However, our purposes at hand of course pertain centrally not to physics, but to consciousness, to which we now turn.

1.3 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 completely suffice to have before us but three meanings, the first two of which are very nicely adumbrated in (Block 1995). The two in question are *access consciousness* and *phenomenal consciousness*; for short, following Block (1995), these are respectively *a-consciousness* and *p-consciousness*. And what are the meanings of these labels, which at the moment will surely be utterly uninformative for those without prior exposure to the literature in question? We briefly answer this question now, starting with the second of the pair of terms just introduced.

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 take 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.

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.[7] 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? What is it? Your two authors would very much like to provide you with a formal definition, but the concept is due to Block [certainly with antecedents he (1995) discusses], and remains murky, perhaps even irremediably so, unless the second author's long-ago issued recommendation to discard the term 'a-consciousness' in favor of using instead terms that refer to the kinds of things this umbrella term is supposed to cover (Bringsjord 1997). Having said this, we convey that Block's (1995) definition (p. 231) is as follows: A state of some agent is a-conscious if and only if it is poised (a) to be used as a premise in reasoning, (b) for rational control of action, and (c) for rational control of speech. Actually, Block tells us (p. 231) 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 the second author has in the past pointed out (1997), 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 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. Two, action is controlled by rational deduction from such sets. 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 application "talks" by producing text, but it could be outfitted with a voice synthesizer. We assume most of our readers will agree, in light of this example, that plenty of systems can be a-conscious (but not necessarily p-conscious).

Alert readers will remember that we promised above to characterize not just two kinds of consciousness, but three. The third happens to be the one near and dear to our hearts (for reasons to be in part shared below), and is *cognitive consciousness*, or just *c-consciousness*. This brand of consciousness is present only when the agent that bears it has a robust ensemble of cognitive attitudes, which correspond directly to a relevant set of verbs bound up with cognition as long investigated in cognitive psychology and cognitive science (e.g. see the cognitive verbs that anchor a number of the chapters in the authoritative Ashcraft & Radvansky 2013). 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) 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.[8]

Let us now take stock. At this point we have available to us three types of consciousness;

accordingly, we can partially instantiate the template (2) to yield three variants:

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

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

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

From this point on in the present chapter, our concern will be fundamentally with p-consciousness and c-consciousness.

2 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 abbreviatory scheme we have allowed ourselves:

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

Of course, this isn't very informative in the absence of a characterization of IIT and Φ. We provide this characterization now.

2.1 Aaronson's Conveniently Compressed Version of IIT

In order to simplify our characterization, we relay heavily below on Aaronson's compressed version of IIT & Φ. 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.

In the version presented here, IIT seeks to measure the p-consciousness of discrete finite systems that evolve over time ts, $t \in \mathbb{N}$. Given a finite alphabet Σ, consider a system θ with its state fully at time t specified by finite strings $\langle s_1, \ldots, s_n\rangle$ of size n drawn from Σ. θ undergoes state changes specified by a function $f : \Sigma^n \to \Sigma^n$. f can be a non-deterministic function. IIT seeks to measure consciousness present in θ using the quantitative scalar measure $\Phi(\sigma)$.

For example, consider a simple binary alphabet $\Sigma = \{0, 1\}$ representing boolean values *False* and *True*, let $n = 4$, and in addition let

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

.

Given an initial state of $\langle 0, 1, 1, 1\rangle$, under f, the system would change to $\langle 0, 1, 1, 0\rangle$. $\Phi(\theta, \langle y_1, \ldots y_n\rangle)$ is defined for the system θ as a concrete state given by $\langle y_1, \ldots y_n\rangle$. See Figure 1 for a pictorial depiction of the kind of evolution through time we have just defined.

For any non-empty partition of the state string into two partitions (A, B), we define the *effective information* $EI(A \to B)$ as the Shannon entropy of B if the characters in A are drawn uniformly at random with B kept fixed at their input values. See further explanation see Figure 2.

Now let $A = \langle y_{A_1}, \ldots, y_{A_j}\rangle$ and $B = \langle y_{B_1}, \ldots, y_{B_k}\rangle$

$$EI(A \to B) := H\big(B_{t+1} \mid A_t \sim \textit{uniform}, B_t = \langle y_{B_1}, \ldots, y_{B_k}\rangle\big)$$

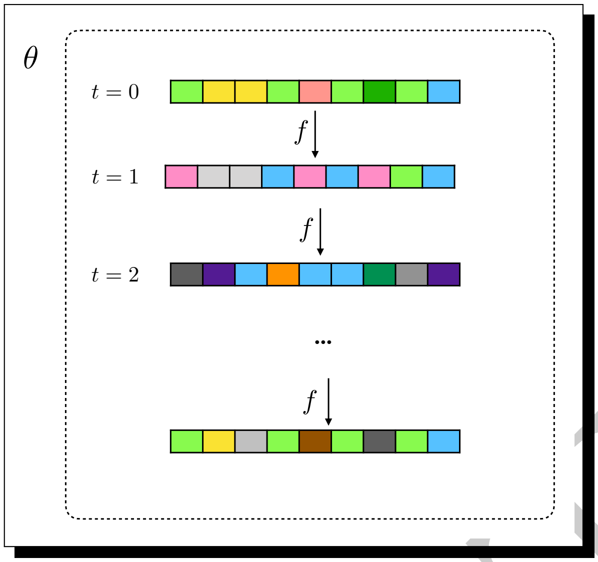

Figure 1: **An Evolving-System Perspective**. Aaronson presents IIT as seeking to measure p-consciousness of systems which evolve under an update function. The state of the system is fully specified by a finite fixed-length string.

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

.

The definition above looks at only one partition (A, B) of the system. To arrive at a quantity for the whole system, which is crucial for IIT, this theory considers all such partitions and takes the quantity $\phi(A, B)$ which minimizes the normalized $\frac{\phi(A,B)}{min(|A|,|B|)}$:

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

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. The *central thesis* of IIT can now be stated as follows:

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 the 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 direction of attack against IIT. We turn now to Searle's attacks on IIT and Φ.

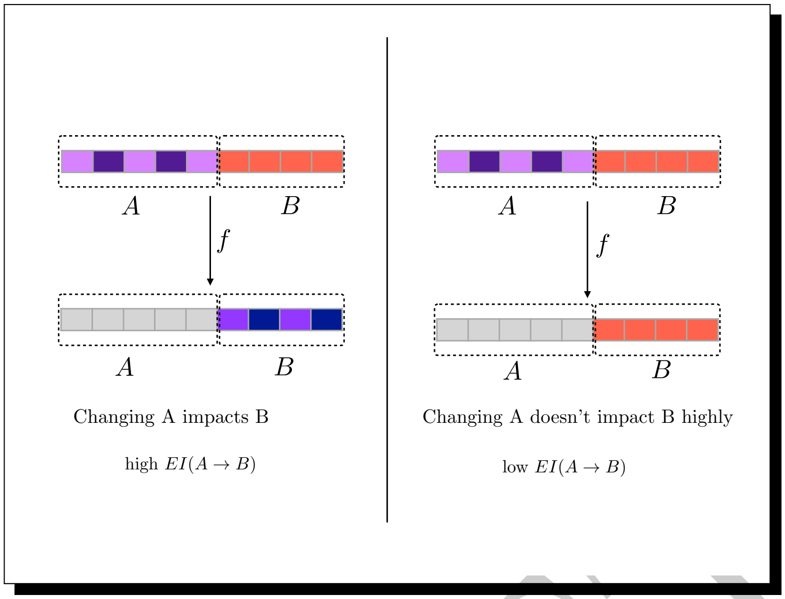

Figure 2: **Effective Information**. Effective information seeks to measure how much the changing of one subsystem (A) effects another subsystem (B). On the left side, we have high effective information, since changing A changes B significantly. On the right side, changing A does not lead to significant changes in B, giving us low effective information.

3 Searle's Attacks on IIT/Φ

John Searle has energetically attacked IIT, in a way he regards to be utterly fatal to the theory.[9] 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 charting this intellectual history would take us too far afield, and would add little to an analysis of the attack itself. 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 pivots around the IIT-and Φ-based treatment of p-consciousness provided in Koch & Tononi's (2012) *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 (Searle 2013), Koch and Tononi replied, and Searle replied to the reply; the exchange is (Koch, Tononi & Searle 2013). As the reader can infer, Searle had the last word in this exchange, and that last word is the best 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 challenge to IIT and Φ. We turn now to the first part of the challenge, Attack #1 from Searle.

3.1 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). (Searle 2013 ¶1)

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, allow us to make at least some sense of Searle's reasoning, and to then 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.

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 $\Delta(\textit{p-consciousness})$. Now consider a case of explanation in the form of (2) that's even simpler than the axioms of classical mechanics we alluded to above: viz., the axioms of arithmetic on the natural numbers, $\mathbb{N}$. This axiom system is commonly known as 'Peano Arithmetic,' or just **PA**. Here are two simple axioms in **PA**, where '$s(n)$' denotes the successor of n:[10]

$$(3) \qquad \text{For all } n : s(n) \neq 0$$

$$(4) \qquad \text{For all } n, m : \text{if } s(n) = s(m), \text{then } n = m$$

As to the observed phenomena that **PA** explains, this includes such elementary-school observations as that $5 \times 5 = 25$, which young children encounter; that there exists a (natural) number n greater than 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 by humans even today that as you explore the progression of *positive cubic numbers*, that is numbers of the form n^3, starting 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.[11]

Now what does all this have to do with Searle's Attack #1? Well, recall that Searle claims 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 (Searle 1992), he claimed that the view that the mind is essentially a computer is unacceptable, because computation is observer-relative. But we don't need to exhume the ins and outs of that earlier work, because what we have on the table now regarding **Peano Arithmetic** 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:

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

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 paper 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 do indeed 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[12] 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** $\Delta(p\textit{-consciousness})$ to some agent A; and $\Phi_{E_{\text{IIT}}}[\Delta(p\textit{-consciousness})] = \mu$

3.2 Attack #2: "But IIT entails panpsychism!"

Searle's second attack is more straightforward than his first. His argument is simply that IIT implies that 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 in his own words, verbatim:

> They [Koch & 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. (Searle 2013 ¶8)

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, but this in no way constitutes a refutation of IIT itself. Why can't Koch and Tononi simply retort that they like 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.[13]

Thus we now leave Searle and come to more serious objections to IIT and its measurement scheme.

4 High Φ is Neither Necessary nor Sufficient for Consciousness

For any phenomenon p, in order to explain observations $\Delta(p)$, we need to explain each *individual* observation $i \in \Delta(p)$. In the case of p-consciousness, we don't have any third-person experimental way of collecting observations. Therefore we must substitute common-sense observations, giving us the following meta-theoretical axiom:

Axiom of Common-Sense Observations (ACO)

$$\forall i\left(\left[\mathit{CommonInstance}(i, \mathit{p\text{-}consciousness})\right] \Leftrightarrow \left[i \in \Delta(\mathit{p\text{-}consciousness})\right]\right)$$

With this quantified biconditional as our anchor, we now proceed to consider first the necessary-condition (left-to-right) direction, and after that discuss the sufficient-condition (right-to-left) direction.

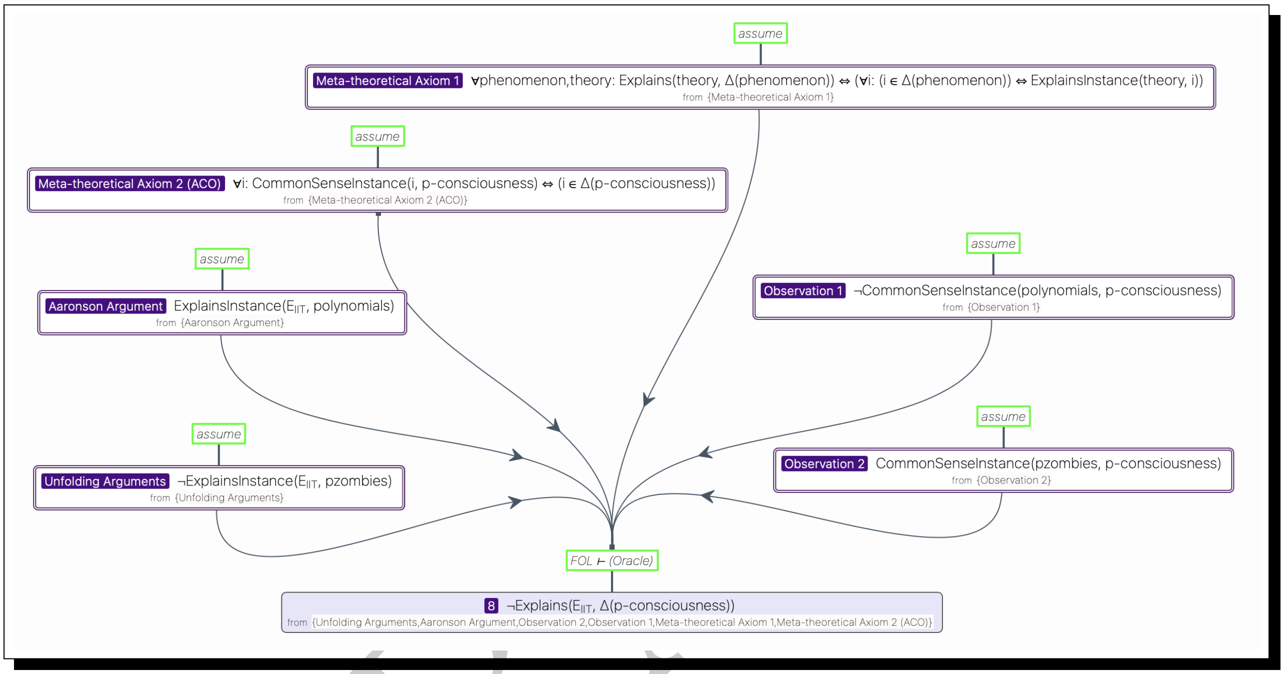
assume
Meta-theoretical Axiom 1 ∀phenomenon,theory: Explains(theory, Δ(phenomenon)) ⇔ (∀i: (i ∈ Δ(phenomenon)) ⇔ ExplainsInstance(theory, i))
from {Meta-theoretical Axiom 1}
assume
Meta-theoretical Axiom 2 (ACO) ∀i: CommonSenseInstance(i, p-consciousness) ⇔ (i ∈ Δ(p-consciousness))
from {Meta-theoretical Axiom 2 (ACO)}
assume
Observation 1 ¬CommonSenseInstance(polynomials, p-consciousness)
from {Observation 1}
assume
Aaronson Argument ExplainsInstance(E_IIT, polynomials)
from {Aaronson Argument}
assume
Observation 2 CommonSenseInstance(pzombies, p-consciousness)
from {Observation 2}
assume
Unfolding Arguments ¬ExplainsInstance(E_IIT, pzombies)
from {Unfolding Arguments}
FOL ⊢ (Oracle)
8 ¬Explains(E_IIT, Δ(p-consciousness))
from {Unfolding Arguments,Aaronson Argument,Observation 2,Observation 1,Meta-theoretical Axiom 1,Meta-theoretical Axiom 2 (ACO)}

Figure 3: **Metatheory Damaging to IIT**. This is a proof in first-order logic using HyperSlate® (Bringsjord et al. 2008) that shows how the unfolding argument and Aaronson's argument lead us to the conclusion that IIT does not explain p-consciousness. (Please note that the notation used in HyperSlate® is slightly different from the notation we use in the text, due to HyperSlate® displaying formulae in a uniform, graphical way.)

4.1 Necessary Condition: Unfolding Arguments

For the necessary-condition direction, we begin by noting that several authors have shown that IIT and other similar theories are vulnerable to what is called the *unfolding argument* (Doerig, Schurger, Hess & Herzog 2019). Formulations of IIT & Φ require the presence of feedback loops for a system to have non-zero Φ values. It is well-known in computability theory that any system with feedback loops can be simulated by a system without feedback loops in a "feed-forward" fashion. This entails as a matter of mathematics that IIT will assign high Φ values to a system σ_1 with intricate feedback loops; IIT will also then assign zero or very low values to another system σ_2 that has the same outward behavior as σ_1 but is implemented in a feed-forward fashion without any loops.

For a concrete example, Hanson and Walker (2021) construct two versions of an electronic tollbooth; yes, a tollbooth. While both versions have the same functionality and implement the same function, the IIT-"conscious" version, made with feedback loops, has non-zero Φ values in all states; the IIT-"unconscious" version, constructed in a feedforward fashion without any loops, has in stark contrast zero Φ values in all states. This argument establishes that IIT assigns different Φ values to systems that have the same behavior. The upshot of this is that we can in principle show that there can be a new class of "feedforward" IIT-zombies,[14] organisms that look and behave like humans, but are assigned zero Φ values due to their brains being wired in a feedforward fashion. These arguments show that high Φ values are not necessary for systems that we might consider to be p-conscious under common-sense but naïve observations. By using **ACO**, we have that Φ values are not necessary for p-consciousness. Summarizing economically, the unfolding argument gives us the following proposition, rather a problem for IIT:

$$\neg\big(E_{\text{IIT}}\ \mathbf{Explains}\ \textit{pzombies}\big)$$

4.2 Sufficient Condition: Aaronson's Attack

Aaronson defines a system whose update function f reduces to evaluating a polynomial on a set of points. Given this, we present an intuitive argument based on Aaronson's attack that gets us to the gist of the issue without delving into the technical minutiae. Delving deeply into Aaronson's technical details would greatly exceed scope in the present chapter.

The internal state of the Aaronson's system corresponds to the coefficients $\langle c_0, c_1, \ldots, c_n\rangle$ of a polynomial $s(x)$. The update function evaluates the polynomial over a finite set of points $\langle p_0, p_1, \ldots, p_n\rangle$, giving us the new state. As can be easily seen, the value of the polynomial evaluated at any point depends on all the coefficients. A setup such as this creates a highly interconnected system. The important point to note is that Aaronson's construction allows us to create a family of systems with unbounded ϕ values. Increasing ϕ is essentially accomplished by simply increasing the degree of the polynomials used. An example is shown in Figure 4.2, which you should now consult. In this example, changing one component in the input state, $3 \rightarrow 2$, results in all the components in the next time step, except for one, being different. Summarizing, this argument gives us the following proposition:

$$\big(E_{\text{IIT}}\ \mathbf{Explains}\ \textit{polynomials}\big)$$

So, Aaronson's argument shows that we can get high Φ values for systems that would be irrational to classify as p-conscious, at least under general common-sense constraints. Using **ACO**, this line of argument shows that high Φ values are not sufficient for p-consciousness.

5 Final Evaluation of Attacks on IIT/Φ

What, then, is the final assessment of IIT and Φ, in light of the foregoing? Should this pair be rejected? Our conclusion is that IIT, given the objections we have now summarized, simplified, and advanced, fails to explain p-consciousness, given the axiom of common-sense observations (**ACO**). This argument is formally shown in Figure 3; it uses only standard first-order logic in the HyperSlate® proof environment. The logical validity of the proof is formally verified by the checker built into this environment.

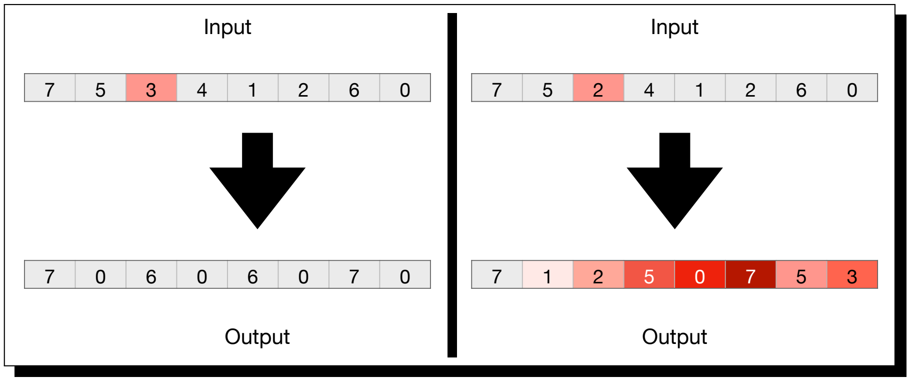

Figure 4: **Polynomial-based Update Functions**. In the example shown here, the update function f is a polynomial evaluation over the Galois field $GF(8)$. The state has 8 components. The polynomial is evaluated over the points $\{0, 1, 2, 3, 4, 5, 6, 7\}$. The coefficients of the polynomial are given by the values in the current state. The polynomial for the starting state on the left is $6x^6 + 2x^5 + x^4 + 4x^3 + 3x^2 + 5x + 7$ and the polynomial corresponding to the state on the right is $6x^6 + 2x^5 + x^4 + 4x^3 + 2x^2 + 5x + 7$. As the figure shows, a small change in one component $3 \rightarrow 2$ impacts all other components of the state; this yields a system with many feedback loops. To easily visualize the changes, the difference in the output state values are represented by color intensity.

However, a defender of IIT retains the option of "biting the bullet" if they reject **ACO**. 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. We turn to this pair now.

6 The Better Way: The Theory of Cognitive Consciousness, Λ, & Intelligence

We already characterized TCC above, and so now move to a discussion of its axiomatization.

6.0.1 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 the axioms of cognitive consciousness (= c-consciousness). For the full axiomatic treatment, the reader is directed first to the introduction of this axiomatization (and cognitive consciousness in general), provided in (Bringsjord, Bello & Govindarajulu 2018), and then, for a more detailed (and more technical) presentation of the axioms, to (Bringsjord & Govindarajulu 2020), which presents the axiom system $\mathcal{CA}$ in its expanded and more rigorous form. It will suffice in the current context if we show the reader but two of the simpler axioms of $\mathcal{CA}$, to wit:

Perception to Belief

P2B Human persons perceive internally[15] and externally,[16] 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, has 13 possible values. In large measure due to the research and engineering of the first author, and 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; see (Bringsjord, Govindarajulu & Giancola 2021).

Now the second axiom we share here:

Introspection (positive)

Intro Humans persons know that they know what they know, etc.

This axiom is in fact well-known in formal logic because it corresponds to a much-discussed axiom from so-called *alethic* modal logic ($\Box\phi \rightarrow \Box\Box\phi$, when symbolized as the characteristic axiom of modal logic **S4**), which in epistemic logic reads $\Box$/necessarily as 'Knows.' A bound $k \in \mathbb{N}$ can be placed on the iteration of **K**, but it would we think need to be at least 5 for human-level cognition (for a rationale, see Bringsjord & Ferrucci 2000). 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.

6.1 Cognitive Consciousness, Measured: Λ

Λ differs from IIT/Φ in two significant ways. First, Λ measures consciousness, specifically, as the reader now knows, c-consciousness, based on how the system observably behaves, instead of the peculiarities of vague internal structures in the system. Importantly, rather than striving to measure phenomenal consciousness (p-consciousness), Λ explicitly aims to explain and account only for cognitive consciousness (c-consciousness). These two differences take us a considerable distance away from IIT/Φ and shield us from the attacks presented earlier.

We 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 outputs 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

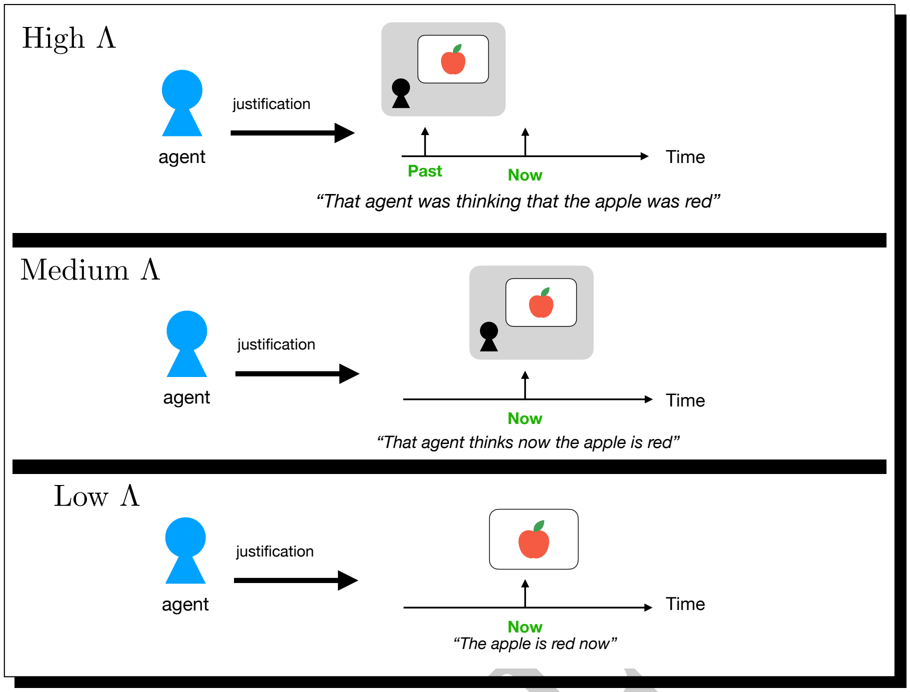

Figure 5: Λ **and Justifications**. We have higher Λ values when the agent has to consider other agents and handle richer temporal structures.

a single Λ value for an agent or system or creature which is to be measured with respect to c-consciousness; rather, a sequence or vector 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 $\mathit{justification}(a, \alpha, t)$ produced by an agent a for action α at time t.

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

6.2 Λ and Cognitive Intelligence

One prominent drawback of IIT is that the theory has no linkage with any formal theory of computational intelligence [such as Hutter's Hutter's (2005) AIXI, which is such a theory] or for that matter any computational hierarchy used to consider intelligence in an abstract way, such as the Arithmetic Hierarchy explained and used in (Bringsjord & Zenzen 2003). Λ rectifies this situation immediately and decisively, since, by construction, Λ's measures reflect the depth of different components of cognitive intelligence, which we 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 λcomponents in this case. On other hand, Λ will assign high λ values to artificial characters, and even simulations of fictional characters that have rich cognitive lives; this outcome mirrors our intuition that such creatures ought to regarded to possess substantive cognitive consciousness, and, correspondingly, high cognitive intelligence.

6.3 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.[17]

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; though see (Govindarajulu, Licato & Bringsjord 2014)) 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 (Russell & Norvig 2020), the recently released fourth edition of what is by any reasonable metric a massive tome. 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 seen 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.

But AGI leads directly to a different situation. To see this, we can first assume, without loss of generality or accuracy, that *any* type of field or discipline devoted to rigorously designing and implementing creatures or beings that are both artificial and intelligent in some scientifically meaningful senses of these two terms can be thought of as *agents.* Very well; then what sort of such agent 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., (Wang 2019).[18] For our purposes, we can focus on a key *sine qua non* for AGI of any level in an artificial agent, according to Wang (2019): 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.[19] 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.[20]

In FBT, we ask the agent $a^\star$ to be assessed 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 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.

7 Conclusion

To sum up, at least in our view (and, needless to say, in the view too of Searle, Aronson, and the other IIT skeptics we have mentioned — and hopefully also in your view), minimally proposition [(2^p_{IIT}), which expresses that IIT scientifically explains p-consciousness, should be disbelieved by those systematically and objectively seeking a scientific explanation of p-consciousness.[21] Realistically, we are inclined to believe that IIT and Φ, at least in the form of variants, will nonetheless 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 focus upon the 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 indeed 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. As to whether what we envision will materialize, only time, of course, will tell.

Notes

[1] 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: e.g. (Nagel 1979).

[2] 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, resp., quantum mechanics and relativity), direct observation is unattainable. But there is still, if science is being brought to bear in search of explanation, observational data.

[3] 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 Smith 2017). For a lucid history of systematic thought from Euclid to Frege that revolves around formal logic, see (Glymour 1992).

[4] 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: (Jain, Osherson, Royer & Sharma 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 what a scientist does when, as inquiry unfolds through time, she offers candidate explanations for what she observes.

[5] The full title, rarely used: *PhilosophiæNaturalis Principia Mathematica* (Mathematical Principles of Natural Philosophy). Even the very first edition, with Newton's annotations, are available online upon a brief search.

[6] Interested readers can start by consulting (Andréka, Madarász, Németi & Székely 2011).

[7] This is not to say that there isn't a highly cognitive side to cricket. There is; see e.g. (Brearley 2015). Hence, our theory of consciousness, the cognitive theory of consciousness (TCC), or *c-consciousness*, is also embodied in the playing of serious cricket.

[8] Leaving aside for simplification the mind of God, which is purportedly, at least to a degree, known to a number of scientists.

[9] And if he's right, by extension Φ would of course fall as well.

[10] We shall not spend the considerable time that would be needed to list all the axioms, and explain them. Readers can consult the elegant (Ebbinghaus, Flum & Thomas 1994) for nice coverage of **PA**. 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 that induction-free theory of

arithmetic known as 'Robinson Arithmetic,' or sometimes just as 'Q;' for elegant coverage, see (Boolos, Burgess & Jeffrey 2003).

[11] 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? 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 phenomena in question.

[12] Here used an an adjective.

[13] As a matter of fact, while it has as far as we can see slipped by largely unnoticed, Searle has offered a *third* attack on IIT. He writes:

> But the deepest objection is that [IIT] is unmotivated. Suppose [Tononi & Koch] could give a definition of integrated and differentiated information that was not observer-relative, that would enable us to tell, from the brute physics of a system, whether it had such information and what information exactly it had. Why should such systems thereby have qualitative, unified subjectivity? In addition to bearing information as so defined, why should there be something it feels like to be a photodiode, a photon, a neutron, a smart phone, embedded processor, personal computer, . . . or any of their other wonderful examples? (Searle 2013 ¶13)

Like Attacks #1 and #2, this third one doesn't seem to us to be conclusive. One reason is that IIT is motivated, at least in part, by observed phenomena arising from the study of brains, including human brains, which Koch and Tononi quite reasonably take to be things at least intimately causally associated with p-consciousness.

[14] These are not the zombies of TV and cinema, but rather zombies of *the philosophers*. For an extended treatment of zombies in this sense, and how that can be used to overthrow similarly vulnerable theories of p-consciousness, see (Bringsjord 1999).

[15] E.g., we perceive that we are in pain when we are.

[16] E.g., we perceive creatures whose behavior indicates to us that they are in pain.

[17] For recommendations regarding characterizations of AGI in the literature, and crisp summaries of these characterizations, we are greatly indebted to James Oswald.

[18] We recognize that our readers may wish to study other accounts of AGI, and hence provide some pointers: Goertzel (2014), 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. Hutter (2005) (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. Finally, in a very nice paper that is generally in line with the paper by Wang (2019) we here rely upon, Voss (2007) provides an account of AGI that for specific technical reasons beyond the scope of the present chapter 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').

[19] 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 (2019) does emphasise solving "general" problems, he doesn't emphasize *human-level* problems of this type.

[20] FBT is sometimes referred to as the "Sally-Anne" task. For a definition and discussion of FBT in psychology/cognitive science, see (Premack & Woodruff 1978). 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 (Arkoudas & Bringsjord 2009). For a more recent, elegant analysis and formal model of the task using hybrid logic, see (Braüner 2014).

[21] Of course, again, we don't think that such an explanation is even conceptually possible, since the explanation Δ that would be at the heart of an explanation [= (n)] would by definition need to be 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 the second author 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 (Bringsjord 2007).

References

Andréka, H., Madarász, J. X., Németi, I. & Székely, G. (2011), 'A Logic Road from Special Relativity to General Relativity', *Synthese* pp. 1–17.
URL: *http://dx.doi.org/10.1007/s11229-011-9914-8*

Arkoudas, K. & Bringsjord, S. (2009), 'Propositional Attitudes and Causation', *International Journal of Software and Informatics* **3**(1), 47–65.
URL: *http://kryten.mm.rpi.edu/PRICAI_w_sequentcalc_041709.pdf*

Ashcraft, M. & Radvansky, G. (2013), *Cognition*, Pearson, London, UK. This is the 6th edition.

Block, N. (1995), 'On a Confusion About a Function of Consciousness', *Behavioral and Brain Sciences* **18**, 227–247.

Boolos, G. S., Burgess, J. P. & Jeffrey, R. C. (2003), *Computability and Logic (Fourth Edition)*, Cambridge University Press, Cambridge, UK.

Braüner, T. (2014), 'Hybrid-Logical Reasoning in the Smarties and Sally-Anne Tasks', *Journal of Logic, Language and Information* **23**, 415–439.

Brearley, M. (2015), *The Art of Captaincy: The Principles of Leadership in Sport and Business*, Pan Macmillan, London, UK.

Bringsjord, S. (1997), 'Consciousness by the Lights of Logic and Common Sense', *Behavioral and Brain Sciences* **20**(1), 227–247.

Bringsjord, S. (1999), 'The Zombie Attack on the Computational Conception of Mind', *Philosophy and Phenomenological Research* **59**(1), 41–69.

Bringsjord, S. (2007), 'Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline', *Journal of Consciousness Studies* **14**(7), 28–43.
URL: *http://kryten.mm.rpi.edu/jcsonebillion2.pdf*

Bringsjord, S., Bello, P. & Govindarajulu, N. (2018), Toward Axiomatizing Consciousness, *in* D. Jacquette, ed., 'The Bloomsbury Companion to the Philosophy of Consciousness', Bloomsbury Academic, London, UK, pp. 289–324.

Bringsjord, S. & Ferrucci, D. (2000), *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, Lawrence Erlbaum, Mahwah, NJ.

Bringsjord, S. & Govindarajulu, N. (2020), 'The Theory of Cognitive Consciousness, and Λ (Lambda)', *Journal of Artificial Intelligence and Consciousness* **7**(1), 155–181. The URL here goes to a preprint of the paper.
URL: *http://kryten.mm.rpi.edu/sb_nsg_lambda_jaic_april_6_2020_3_42_pm_NY.pdf*

Bringsjord, S., Govindarajulu, N. & Giancola, M. (2021), 'Automated Argument Adjudication to Solve Ethical Problems in Multi-Agent Environments', *Paladyn, Journal of Behavioral Robotics* **12**, 310–335. The URL here goes to a *rough*, *uncorrected*, *truncated* preprint as of 071421.
URL: *http://kryten.mm.rpi.edu/AutomatedArgumentAdjudicationPaladyn071421.pdf*

Bringsjord, S., Taylor, J., Shilliday, A., Clark, M. & Arkoudas, K. (2008), Slate: An Argument-Centered Intelligent Assistant to Human Reasoners, *in* F. Grasso, N. Green, R. Kibble & C. Reed, eds, 'Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 8)', University of Patras, Patras, Greece, pp. 1–10.
URL: *http://kryten.mm.rpi.edu/Bringsjord_etal_Slate_cmna_crc_061708.pdf*

Bringsjord, S. & Zenzen, M. (2003), *Superminds: People Harness Hypercomputation, and More*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Doerig, A., Schurger, A., Hess, K. & Herzog, M. H. (2019), 'The unfolding argument: Why iit and other causal structure theories cannot explain consciousness', *Consciousness and cognition* **72**, 49–59.

Ebbinghaus, H. D., Flum, J. & Thomas, W. (1994), *Mathematical Logic (second edition)*, Springer-Verlag, New York, NY.

Glymour, C. (1992), *Thinking Things Through*, MIT Press, Cambridge, MA.

Goertzel, B. (2014), 'Artificial General Intelligence: Concept, State of the Art, and Future Prospects', *Journal of Artificial General Intelligence* **5**(1), 1–46.

Govindarajulu, N., Licato, J. & Bringsjord, S. (2014), Toward a Formalization of QA Problem Classes, *in* B. Goertzel, L. Orseau & J. Snaider, eds, 'Artificial General Intelligence; LNAI 8598', Springer, Basel, Switzerland, pp. 228–233.
URL: *http://kryten.mm.rpi.edu/NSG_SB_JL_QA_formalization_060214.pdf*

Hanson, J. R. & Walker, S. I. (2021), 'Formalizing falsification for theories of consciousness across computational hierarchies', *Neuroscience of Consciousness* **2021**(2), niab014.

Hutter, M. (2005), *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*, Springer, New York, NY.

Jain, S., Osherson, D., Royer, J. & Sharma, A. (1999), *Systems That Learn: An Introduction to Learning Theory, Second Edition*, MIT Press, Cambridge, MA.

Koch, C. & Tononi, G. (2012), *Consciousness: Confessions of a Romantic Reductionist*, MIT Press, Cambridge, MA.

Koch, C., Tononi, G. & Searle, J. (2013), 'Can a Photodiode be Conscious?', *The New York Review of Books* . This is an exchange between K&T and S.

McKinsey, J., Sugar, A. & Suppes, P. (1953), 'Axiomatic Foundations of Classical Particle Mechanics', *Journal of Rational Mechanics and Analysis* **2**, 253–272.

Nagel, E. (1979), *The Structure of Science: Problems in the Logic of Scientific Explanation*, Hackett, Indianapolis, IN. This is the second edition.

Premack, D. & Woodruff, G. (1978), 'Does the Chimpanzee have a Theory of Mind?', *Behavioral and Brain Sciences* **4**, 515–526.

Russell, S. & Norvig, P. (2020), *Artificial Intelligence: A Modern Approach*, Pearson, New York, NY. Fourth edition.

Searle, J. (1992), *The Rediscovery of the Mind*, MIT Press, Cambridge, MA.

Searle, J. (2013), 'Can Information Theory Explain Consciousness?', *New York Review of Books* pp. 54–58.
URL: *http://www.nybooks.com/articles/2013/01/10/can-information-theory-explain-consciousness*

Smith, R. (2017), Aristotle's Logic, *in* E. Zalta, ed., 'The Stanford Encyclopedia of Philosophy'.
URL: *https://plato.stanford.edu/entries/aristotle-logic*

Voss, P. (2007), Essentials of General Intelligence: The Direct Path to Artificial General Intelligence, *in* B. Goertzel & C. Pennachin, eds, 'Artificial General Intelligence', Springer, Berlin, Germany, pp. 131–157.

Wang, P. (2019), 'On Defining Artificial Intelligence', *Journal of Artificial General Intelligence* **10**(2), 1–37.
URL: *https://doi.org/10.2478/jagi-2019-0002*