Journal of Artificial Intelligence and Consciousness

The Theory of Cognitive Consciousness, and Λ (Lambda)*

Selmer Bringsjord
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
Selmer.Bringsjord@gmail.com

Naveen Sundar G.
Rensselaer AI & Reasoning (RAIR) Lab
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
Naveen.Sundar.G@gmail.com

Received 7 February 2020
Revised ??? ??? ????

We provide an overview of the theory of cognitive consciousness (TCC), and of Λ; the latter provides a means of measuring the amount of cognitive consciousness present in a given cognizer, whether natural or artificial, at a given time, along a number of different dimensions. TCC and Λ stand in stark contrast to Tononi's Integrated information Theory (IIT) and Φ. We believe, for reasons we present, that the former pair is superior to the latter. TCC includes a formal axiomatic theory, $\mathcal{CA}$, the 12 axioms of which we present and briefly comment upon herein; no such formal theory accompanies IIT/Φ. TCC/Λ and IIT/Φ each offer radically different verdicts as to whether and to what degree AIs of yesterday, today, and tomorrow were/are/will be conscious. Another noteworthy difference between TCC/Λ and IIT/Φ is that the former enables the measurement of cognitive consciousness in those who have passed on, and in fictional characters; no such enablement is remotely possible for IIT/Φ. For instance, we apply Λ to measure the cognitive consciousness of: Descartes; the first fictional detective to be described on Earth (by Edgar Allen Poe), C. Auguste Dupin. We also apply Λ to compute the cognitive consciousness of an artificial agent able to make ethical decisions using the Doctrine of Double Effect.

Keywords: consciousness; cognitive consciousness; AI; Lambda/Λ.

*We are indebted to SRI International for support of a series of symposia on consciousness that proved to be the fertile ground in which which Λ's germination commenced, and to many co-participants in that series for stimulating debate and discussion, esp. — in connection with matters on hand herein — Giulio Tononi, Christof Koch, and Antonio Chella.

1. Introduction

We provide an overview of the theory of cognitive consciousness (TCC), and of Λ; the latter provides a means of measuring the amount of cognitive consciousness present in a given cognizer, whether natural or artificial, at a given time, along a number of different dimensions. TCC and Λ stand in stark contrast to Tononi's [2014; 2012; 2014] Integrated information Theory (IIT) and Φ, which — for reasons we explain — are inadequate when stacked against TCC/Λ. TCC includes a formal axiomatic theory, $\mathcal{CA}$, the axioms in which we enumerate herein; no such formal theory accompanies IIT/Φ. TCC/Λ and IIT/Φ each offer radically different verdicts as to whether and to what degree AIs of yesterday, today, and tomorrow were/are/will be conscious. In our case, while we are guided by the human case, and insist upon staying at the human-level in the study of consciousness, our concern is primarily *artificial* consciousness [Chella & Manzotti, 2007]. Indeed, by the lights of TCC/Λ/$\mathcal{CA}$, there are among humans already a fair number of cognitively conscious AIs — but these AIs have low Λ. Another noteworthy difference between TCC/Λ and IIT/Φ is that the former enables the measurement of cognitive consciousness in those who have passed on, and in fictional characters; no such enablement is remotely possible for IIT/Φ. We herein apply Λ to measure the cognitive consciousness of: Descartes, and the first fictional detective to be described on Earth (by Edgar Allen Poe), C. Auguste Dupin. We also apply Λ to compute the cognitive consciousness of an artificial agent able to make ethical decisions using the Doctrine of Double Effect.

The sequel will unfold as follows. We first take note of the brute and rather discouraging fact that 'consciousness' is polysemous, and announce in this context what our particular concern is (§2). Next, in section 3, we offer a compressed, narrow critique of IIT/Λ, one stemming specifically from our orientation and approach. Then we explain to the reader, in a nutshell, the formal foundations of TCC/Λ (4). Our next step is to give a quick enumeration of the axioms of $\mathcal{CA}$, with each axiom followed by brief explanatory remarks (§5). A concise defintion of Λ is then given (§6), after which our simulations and corresponding Λ measurements are shown (§7). We end the paper proper with a brief conclusion (§8). An Appendix is provided in order to enable readers to more deeply understand the basis for TCC and *Lambda*: cognitive calculi.

2. Our Chosen Route in the Polysemy of 'Consciousness'

William James [1904] was of the opinion that the term 'consciousness' denotes nothing at all, and urged that the term be banished from science.[1] Just over a century

[1]He writes:

> I believe that''consciousness,' when once it has evaporated to this estate of pure diaphaneity, is on the point of disappearing altogether. It is the name of a nonentity, and has no right to a place among first principles. Those who

out from his skeptical recommendation, relevant scientists and engineers, it must be admitted, find a terminological situation that presumably has James turning in his grave: everyone and his uncle seems to have and promote their favorite term in a space that appears to be expanding in Zenoian fashion. We herein, and indeed right now, starting with the present sentence, avow that our particular focus is on *human-level cognitive consciousness*, and the *structure*, and the structure only, of *human-level phenomenal consciousness*, or just human-level p-consciousness for short [this follows the nomenclature and terminology of Block, 1995]. To further explain, we first point out that we dodge any such demand as capturing p-consciousness in a third-person language, a demand that IIT/Φ proponents accept; and then we explain that for us cognition is "in control," and finally further amplify the fact that our approach is human-level.

2.1. *We Dodge Any Such Demand as Capturing P-Consciousness*

The lead author is on record, repeatedly, as attempting to show that p-consciousness, "what-it-feels-like" consciousness, can't be captured in any third-person scheme, IIT's Φ or otherwise. In the basic two-part breakdown of Block [1995], that of p-consciousness versus *access consciousness* (which he dubs 'a-consciousness'), we are firmly on the "a-" side. The latter kind of consciousness has nothing to do with mental states like that which it feels like to be in the arc of a high-speed giant-slalom ski turn, which are paradigmatically in the p- category; instead, a-conscious states in agents are those that explicitly support and enable reasoning, conceived as a mechanical process. TCC marks a rejection of purely phenomenological study of consciousness, such as for example the impressive book-length study carried out by Kriegel [2015], in favor of an emphasis on formal structures and processes that represent reasoning and decision-making (at the human level). Near the end of his study, Kriegel sums up the basic paradigm he has has sought to supplant:

> Mainstream analytic philosophy of mind of the second half of the twentieth century and early twenty-first century offers one dominant framework for understanding the human mind. . . . The fundamental architecture is this: there is input in the form of perception, output in the form of

> still cling to it are clinging to a mere echo, the faint rumor left behind by the disappearing 'soul' upon the air of philosophy. During the past year, I have read a number of articles whose authors seemed just on the point of abandoning the notion of consciousness, and substituting for it that of an absolute experience not due to two factors. But they were not quite radical enough, not quite daring enough in their negations. For twenty years past I have mistrusted 'consciousness' as an entity; for seven or eight years past I have suggested its non-existence to my students, and tried to give them its pragmatic equivalent in realities of experience. It seems to me that the hour is ripe for it to be openly and universally discarded. [James, 1904, p. 477]

> action, and input-output mediation through propositional attitudes, notably belief and desire. (Kriegel [Kriegel, 2015, p. 201]

This basic paradigm, we cheerfully concede, is the one on which TCC/*Lambda*/$\mathcal{CA}$ is essentially based (though ours is a formal approach). For example, in particular, here are three key intensional operators in TCC, *Lambda*, and $\mathcal{CA}$: $\mathbf{K}$ (knows), $\mathbf{B}$ (believes), $\mathbf{P}^i$ (perception, internal), and $\mathbf{P}^e$ (perception, external). (In the simulations below, we use only a generic perception operator.) While Kriegel has associated the paradigm he rejects with "analytic philosophy," which he claims appropriated it from "physics, chemistry, and biology" (p. 202), the fact is, the field of AI is based on exactly the paradigm Kriegel rejects; that in AI agents are by definition essentially functions mapping percepts to actions is explicitly set out and affirmed in all the major textbooks of AI [Russell & Norvig, 2009; Luger & Stubblefield, 1993, see e.g.]. But the AI approach, at least of the logicist variety that TCC/*Lambda*$\mathcal{CA}$ follows [Bringsjord, 2008], has a benefit that Kriegel appears not to be aware of. In defense of his phenomenological approach, he writes:

> Insofar as some mental phenomena are introspectively observable, there is a kind of insight into nature that is *available* to us and that goes beyond that provided by the functionalist framework. This alternative self-understanding focuses on the experiential rather than mechanical aspect of mental life, freely avails itself of first-person insight, and considers that mental phenomena can be witnessed directly as opposed to merely hypothesized for explanatory benefits. It would be perverse to simply ignore this other kind of understanding and insight. [Kriegel, 2015, p. 202]

Kriegel seems to be entirely unaware of the fact that in AI, researchers are often quite happy to base their engineering on self-analysis and self-understanding. Looking back a bit, note that the early "expert systems" of the 1980's were based on understanding brought back and shared when human experts (e.g., diagnosticians) introspected on how they made decisions, what algorithms they followed, and so on. Such examples, which are decidedly alien to physics, chemistry, and biology, could be multiplied at length, easily. To mention just one additional example, it was introspection on the part of chess grandmaster Joel Benjamin, who worked with the AI scientists and engineers who built Deep Blue (the AI system that vanquished Gary Kasparov), that made the difference, because it was specifically Benjamin's understanding of king safety that was imparted to Deep Blue at a pivotal juncture [for a discussion, see Bringsjord, 1998].

In this light, we now make two points regarding TCC, Λ, and the axiom system $\mathcal{CA}$. First, following the AI tradition to which we have alluded, we feel free to use self-understanding and introspection in order to articulate model, simulated, and axiomatize. Second, as a matter of fact, as will be shortly seen, our work *directly* reflects our affirmation of the importance of self-belief, self-consciousness, and other self-regarding attitudes, right down to the fact that we have specific elements in our formal languages (e.g. a self-designator in $\mathcal{DCEC}^*$, explained later), and axioms in $\mathcal{CA}$ (e.g. axiom **TheI**) that model these "self-" phenomena.

2.2. *Cognitive in Control of the Perceptual and Affective*

Honerich [2014] has recently argued that the best comprehensive philosophical account of consciousness is one that places an emphasis on perceptual, over and above affective or cognitive phenomena. From our perspective, and in our approach, we place the emphasis very much on cognition. This is primarily because in our orientation, cognitive consciousness ranges over perceptual and affective states. In this regard, we are in agreement with at least a significant portion of a penetrating and elegant review of Honderich's *Actual Consciousness* by Jacquette.[2] He writes:

> If I am not only consciously perceiving a vicious dog straining toward me on its leash, but simultaneously feeling fear and considering my options for action and their probabilities of success if the dog breaks free, then I might be additionally conscious in that moment of consciously perceiving, feeling, and thinking.
>
> Consciousness in that event is not exhaustively divided into Honderich's three types. If there is also consciousness of any of these types of consciousness occurring, then consciousness in the most general sense transcends these specific categories. [Jacquette, 2015, ¶5 & first two sentence of ¶6]

We don't have time to provide a full explanation and defense of the fact that TCC/Λ/$\mathcal{CA}$ reflect the position that cognitive aspects of consciousness should be — concordant with Jacquette's trenchant analysis of Honderich — "in control."

2.3. *Consciousness at the "Person Level"*

We are only interested herein in consciousness at the level of *human persons*. Unlike those who affirm IIT/Φ, we are not interested in consciousness in nonhuman animals, such as chimpanzees and fish, let alone "lower" creatures such as ants. And we certainly have zero interest in what some call "proto-consciousness," supposedly a kind low-level consciousness that, so the story goes, is everywhere and in everything.[3] This constraint on our investigation flows deductively from the con-

[2]Even under the charitable assumption that one cannot e.g. form beliefs about states in which one at once perceives, feels, and cognizes, it's exceedingly hard to find Honderich's basis for holding that the perception side holds sway. As Jacquette writes:

> Supposing that there are just these three types of consciousness, that there is never a higher consciousness of simultaneously experiencing moments *of* perceptual and cognitive or affective consciousness, or the like, why should perceptual consciousness come first? Why not say that cognitive consciousness subsumes perceptual and affective consciousness? If inner perception complements the five outer senses plus proprioception as it does in Aristotle's *De anima* III.5 and Brentano's 1867 *Die Psychologie des Aristoteles*, along with all the descriptive psychological and phenomenological tradition deriving from this methodological bloodline of *noûs poetikos* or *innere Wahrnehmung*, then affective consciousness might also be subsumed by cognitive consciousness. [Jacquette, 2015, ¶4]

[3]The first author of the present paper agrees with A. [2006] that non-cognitive consciousness, in the end, is downright incoherent, but there is no space to take this up in the present venue, which

junction of our assumption that, if you will, cognition drives the show, with the proposition that only agents operating in significant measure at the level of human personhood can have the kind of high-level cognition that can do the driving. Some of the intellectual uniqueness of *H. sapiens sapiens* is nicely explained and defended in readable fashion in the hard-hitting but informal [Penn *et al.*, 2008].

We do think It's important to ensure that our readers know that we in no way deny that some non-human animals are conscious, in some way and at some level. We have little idea how to axiomatize, or even to take the first few steps toward axiomatizing, the brand of "cognitively compromised" consciousness that non-human animals have, but we in no way assert that these creatures don't have it.

3. The Rejection of Integrated Information Theory and Φ

We reject IIT/Φ because we believe it has many fatal flaws, but the purpose of the present paper is not to systematically assess and refute IIT/Φ. Accordingly, we only mention, with brutal brevity, but two of these flaws. Since our chief objective is to highlight the stark contrast between IIT/Φ versus TCC/Λ, we direct the reader to other works that seek to outright refute IIT/Φ; a nice place to start in this regard is [Cerello, 2015].

The first of the two flaws infecting IIT/Φ we mention here is that cognition is nowhere to be be found, and the emphasis is on p-consciousness. Without cognition, it is exceedingly hard for us to see what the point is. If we find out that some creature has high Φ, but it can't think, solve problems, make decisions, reason from evidence, know things on the basis of reasoned arguments and proofs, communicate in languages at the top of the Chomsky Hierarchy, why is the Φ significant? Human beings are the most impressive creatures in the known, natural universe — buy why? Because not only are they p-conscious, they are also cognitive agents (and as cognitive agents are intelligent).

For a host of reasons long provided by the lead author and colleagues [e.g. Bringsjord, 1992b, 2007], and in part well and wisely noted well over three centuries ago by Leibniz [as explained in Bringsjord, 2015], the ambitious targeting of p-consciousness in rigorous science and engineering, including specifically in AI, is an exceedingly bad idea. Why? Well, the chief reason is that p-consciousness cannot be captured in any rigorous third-person language, period; and whatever cannot be so captured can't possibly be a target that is reached by rigorous science and engineering. The reason is simple: such science and engineering, by definition, deals only with what *can* be captured in this manner. Yet IIT/Φ sets itself the goal of capturing either the very nature of p-consciousness, or, minimally, what information processing is associated with, or perhaps causes, p-consciousness. A wiser approach,

as we have already said isn't devoted to a refutation of IIT/Φ, which certainly does embrace the notion of non-cognitive (p-)consciousness.

it seems to us, is to insist that high-level cognition is front and center an object of study, and that the *structure* of p-consciousness, at the human-level, is a target as well, but falls under cognitive control.

The second flaw we mention is simply that any account of p-consciousness cast exclusively in terms of information processing is subject to the "arbitrary realization argument" given by Bringsjord [1992b]. The gist of the argument is that it's entirely possible, not only logically but physically, that whatever information processing is taken to capture or correlate with p-consciousness can be instantiated in ridiculous systems that no rational person is going to be willing to say are p-conscious. The problem here is really no different than that which refutes Turing-machine functionalism, the doctrine that the cognition and p-consciousness seen in human persons only needs to correspond to appropriate information processing in a Turing machine. This doctrine is untenable, because a Turing machine of the requisite sort could be instantiated by, say, a collection of interconnected beer cans.

4. The Formal Foundations of TTC, Λ, and $\mathcal{CA}$

One of the requirements for Λ is that we be able to formalize states of minds of artificial and natural agents. Formal logics, specifically with provision for such "mental" formalization, provide us with a universal way of satisfying this requirement; and, at least in theory, logics can be employed by any computational system (since e.g. any form of computation, including those beyond the Turing Limit, can be captured in formal logic). While logics have long been used in AI [Bringsjord & Govindarajulu, 2018], our use of logic here is specifically for measuring and quantifying the consciousness of cognitive agents. We do not require that the underlying agent have been implemented in logic; it is only necessary that the agent be *expressed* in logicist form.[4] Briefly here (we explain in more detail in Appendix A), a *cognitive calculus* is a formal logic $L \equiv \langle \mathcal{L}, \mathcal{I} \rangle$ with a language $\mathcal{L}$ configured to model human-level cognition, and an inference system I. $\mathcal{L}$ includes a quantified modal language with minimally the two standard quantifiers $\forall, \exists$ and a finite set of cognitive operators $\Omega = \{\omega_1, \ldots, \omega_q\}$. The pure non-modal part of the language $\mathcal{L}_0 \subset \mathcal{L}$ is used to model states of the world. The cognitive operators, Ω, are used to model states of minds of agents by building up formulae from agents, times, and formulae. A cognitive operator ω is of the form:

Cognitive Operators

$$\omega : A \times \ldots A \times T \ldots \times T \times \mathcal{L} \ldots \mathcal{L} \rightarrow \mathcal{L}$$

One example of a cognitive operator, one that is central to the axiomatization $\mathcal{CA}$ and our later simulations (as will be seen), is $\mathbf{B} : A \times T \times \mathcal{L} \rightarrow \mathcal{L}$, used to model beliefs

[4] A simple concrete example is the proof that any Turing machine can be easily captured in first-order logic; [e.g. see Boolos *et al.*, 2003].

of agents by building formulae of the form $\mathbf{B}(a, t, \phi)$. A more complex operator is $\mathbf{S}: A \times A \times T \times \mathcal{L} \rightarrow \mathcal{L}$ that denotes a communicating the statement ϕ to b at time t by the formula $\mathbf{S}(a, b, t, \phi)$. Again, and importantly, the full modal language $\mathcal{L}$ is used to model *both* states of the world and cognitive states of agents.

4.1. *Deontic Cognitive Event Calculus with the De Se Operator*

The particular cognitive calculus we use here is $\mathcal{DCEC}^*$, a calculus that is a modal extension of first-order logic, and an extension of $\mathcal{DCEC}$ used for instance in Govindarajulu & Bringsjord [2017a]. The first-order core of $\mathcal{DCEC}^*$ includes an adapted version of the *event calculus* Mueller [2006], a first-order calculus that has been used to build extensive and deep models of the physical world. Other calculi [e.g. the *situation calculus* of McCarthy & Hayes, 1969] for modeling commonsense and physical reasoning can be easily switched out in-place of the event calculus.[5] The first-order language in $\mathcal{DCEC}^*$ also includes a set of terms $\{a_1, \ldots, a_n\}$ denoting agents and a set of terms $\{t_0, t_1, \ldots\}$ for denoting and modeling time.

Modeling Self-Reference: The formal modeling of an agent referring to itself is key for TCC and Λ, so we explain quickly how this phenomenon is captured. Given an agent a, there are three types of self-referential statements the agent a can make, as shown in Figure 1. The statements are: (1) *de dicto*: these statements are the least self-referential and refer to the agent a through a name; (2) *de re*: these statements are more self-referential and refer to a through an indexical reference; and (3) finally, *de se*: statements which refer to the agent through the $*$ operator, where this operator is not reducible to any other name or indexical operator. (This operator is used in the axiom **TheI** in $\mathcal{CA}$, as is seen below.) To model *de se* statements, $\mathcal{DCEC}^*$has a corresponding $*$ operator. Going into the mechanisms of these statements is beyond our scope; longer discussions can be found in Bringsjord & Govindarajulu [2013]; Govindarajulu *et al.* [2019].

Cognitive Operators: The modal cognitive operators present in the calculus include the standard operators for knowledge **K**, belief **B**, desire **D**, intention **I**, etc. The general format of an intensional operator is $\mathbf{K}(a, t, \phi)$, which says that agent a knows at time t the proposition ϕ. Here ϕ can in turn be any arbitrary formula. Also, note the following modal operators: **P** for perceiving a state, **C** for common knowledge, **S** for agent-to-agent communication and public announcements, **B** for belief, **D** for desire, **I** for intention, and finally and crucially, a dyadic deontic operator **O** that states when an action is obligatory or forbidden for agents. $\mathbf{O}(a, t, \phi, \psi)$ models the condition ϕ under which ψ becomes an obligation for a. Figure 2 shows four statements and their versions in $\mathcal{DCEC}^*$.

[5] In fact, and this is quite relevant to the axiom **CCaus** of $\mathcal{CA}$ that ascribes common knowledge of physics/causation to all cognitive agents, we can in principle switch to less naïve formalizations of physical theories, e.g. axiomatizations of classical mechanics.

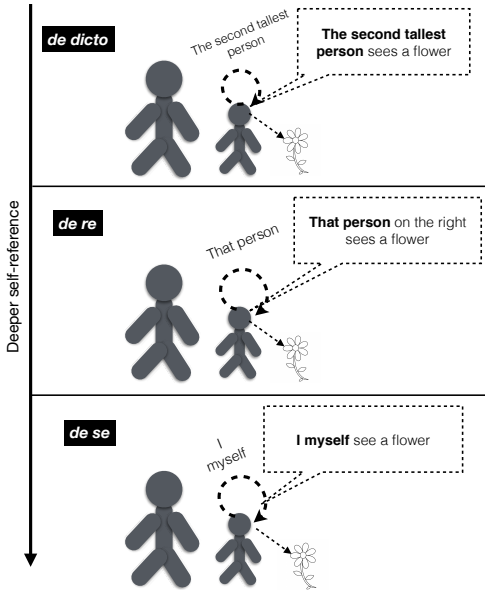

Fig. 1. Three Types of Self-Referential Statements

Syntax

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{cases} q : \text{Formula} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \forall x : \phi(x) \mid \\ \mathbf{P}(a, t, \phi) \mid \mathbf{K}(a, t, \phi) \mid \\ \mathbf{C}(t, \phi) \mid \mathbf{S}(a, b, t, \phi) \mid \mathbf{S}(a, t, \phi) \mid \mathbf{B}(a, t, \phi) \\ \mathbf{D}(a, t, \phi) \mid \mathbf{I}(a, t, \phi) \\ \mathbf{O}(a, t, \phi, (\neg) happens(action(a^*, \alpha), t')) \end{cases}$$

5. The Axiomatic Theory of Cognitive Consciousness ($\mathcal{CC}$)

In this brief section we draw directly from [Bringsjord *et al.*, 2018], in which the axiomatic theory $\mathcal{CA}$ of cognitive consciousness was first presented. Since that time, $\mathcal{CA}$ has matured, but because the chief purpose of the present paper is to achieve broad coverage of TCC/Λ in a context that takes account of IIT/Φ, comprehensive discussion of the former pair must be left for other times and places. It will suffice to present, in rapid-fire fashion, the 12 axioms in question, and to merely offer short remarks on each member of the dozen.

Note that a recent version of IIT, version 3.0, has been declared by M. *et al.* [2014] to have eight axioms, where this octet is supposed to be "derived" from five "phenomenological" axioms. We cannot fathom how the octet is susceptible of formalization, so that theorems can be produced and a formal science of IIT erected (the same attitude is all the more rational with respect to the phenomenological

Fig. 2. Example: Four Different Statements Modeled in $\mathcal{DCEC}^*$.

ancestral quintet). But at any rate, our intention is that $\mathcal{CA}$ be eventually rendered fully in one or more cognitive calculi. (Only some of the axioms in $\mathcal{CA}$ can be represented in $\mathcal{DCEC}^*$.) This rendering allows a computing machine that is cognitively conscious in keeping with $\mathcal{CA}$ to reason over $\mathcal{CA}$ and produce theorems, and allows modern automated-reasoning in AI to derive theorems from this set. Note that, for us, from the standpoint of AI, computational simulations of self-conscious agents should be enabled by *implementation* of one or more of the axioms of $\mathcal{CA}$, presumably usually in the context of some scenario or situation that provides a context composed, minimally, of an environment, n agents, and some kind of challenge for at least one of these agents to meet by reasoning over instances of one or more of the axioms of $\mathcal{CA}$. Before passing to the axioms themselves, we briefly emphasize the high expressivity of the formal logics needed to formalize and implement an artificial agent that satisfied all or even just some of the axioms. Note that $\mathcal{DCEC}^*$ is not sufficiently expressive to formalize the 12 axioms that follow; but some cognitive calculus $\mathscr{C}$ is sufficient to handle any version of all 12. Again, a general characterization of a cognitive calculus is given in Appendix A.1. And a final point before we begin: While we do not return to spell it out, it will become clear to the reader later, after Λ is defined, that when any version of any of the 12 axioms is instantiated for

a given agent a, that agent would thereby have substantive Λ measures.

Perception to Belief	
P2B	Human persons perceive internally and externally, and in both cases the percepts in question are believed (at varying degrees of strength, with external perception at the strength of *evident*, but never *certain*) by these agents, whereas some of what is internally perceived is indeed certain.
Remark	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. Ultimately belief should in our opinion be stratified, in that a belief is accompanied by a strength factor. So for example Jones, if having ingested a powerful drug, may believe only at the level of *more probable than not* that there is a walrus before him. With stratification in place, belief will become graded from certain to certainly false, and so will knowledge. We do not yet have robust implementations of artificial agents that embody this formal machinery, but see [Govindarajulu & Bringsjord, 2017c].

Knowledge to Justified True Belief	
K2JTB	Human persons know things, and hence they believe these things, have justifications (in the form of proofs and arguments) for this belief, and the belief is veridical.
Remark	In [Bringsjord *et al.*, 2018] we presented instead a weaker axiom saying only that knowledge of ϕ implies a belief that ϕ, but here affirm the traditional doctrine going back to Plato, according to which, when human persons know some ϕ they have justified true belief that ϕ. (This doctrine, which can be abbreviated as 'k=jtb,' as some readers will know, was famously attacked by Gettier [1963].)

Introspection (positive)	
Intro	Humans persons know that they know what they know, etc.
Remark	This axiom is well-known in formal logic because it corresponds to a much-discussed axiom from alethic modal logic ($\Box\phi \rightarrow \Box\Box\phi$, the characteristic axiom of modal logic S4), which in epistemic logic interprets $\Box$/necessarily as 'Knows' (our $\mathbf{K}$, herein). A bound $k \in \mathbb{N}$ can be placed on the iteration of of $\mathbf{K}$, 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 ($\neg\mathbf{K}\phi \rightarrow \mathbf{K}\neg\mathbf{K}\phi$), and once again a bound can be placed on the iteration, if desired.

Incorrigibilism

Incorr For any contingent, "Cartesian" property P, it's necessarily the case that if a human person believes himself/herself to have the property *seeming to have* P, that person does have this property.

Remark This axiom expresses the doctrine of *incorrigibilism* (to use the philosophers' term), essentially that self-beliefs of this kind are infallible. A chapter in [Bringsjord, 1992b] is devoted to a defense of this doctrine; it's said there that since if the human mind is physical it couldn't be incorrigible, the human mind isn't physical.

Essence

Ess Every human person has an essence E known to that person, and E is such that if the person in question has a Cartesian property at some time, this implies that this person has E.

Remark The concept of an essence for an agent may strike some as far-fetched, but serious reflection upon what a self-designator (such as 'I' in our cases) means (as opposed to what it denotes) usually quickly results in some willingness to take the possibility that we have essences seriously. In our AI work, which endorses proof-theoretic semantics, a self-designator (see the axiom **TheI**, below) in an artificial agent must be accompanied by suitable proof structures (e.g. there shouldn't be any proof that two diverse agents a_1 and a_2 both have some essence E). With regard to AI, note that Gödel's formalization of the concept of a divine essence has recently been investigated formally and computationally in [Benzmüller & Paleo, 2014].

Emotions Decomposably Complex

¬CompE It's possible for a human person to be in an affective state S such that, for every permutation over a finite set of basic, "building-block" emotions, it's not the case that if that permutation holds of the person in question, this entails that the person is in S. In other words, emotional states in human persons are not limited to what can be composed from finte sets of "building-block" emotions.

Remark This axiom asserts that persons can enter emotional states — but also asserts that some of these states are not constituted by the instantiation of parameters in some core conjunction of "building-block" emotions. The standard core idea in the treatment of emotions in formal logic (or some informal declarative format, which then is reducible to formal logic) is that complex and nuanced emotions are composed of some permutation of building-block emotions (perhaps with levels of intensity represented by certain parameters), modulated by cognitive and perceptual factors. A classic example of such an ontology of emotions is provided by Johnson-Laird & Oatley [1989], whose building-block emotions are: *happiness*, *sadness*, *fear*, *anger*, and *disgust*. The present axiom marks a rejection of such models, in light of what we regard to be myriad counter-examples. However, it would not be hard at all to formalize the categorization of emotions given by Johnson-Laird & Oatley [1989] in a cognitive calculus. Formalization of competing ontologies of emotion can likewise easily be formalized in a cognitive calculus. For instance, the well-known, so-called "OCC" theory of emotions [Ortony *et al.*, 1988], can for the most part be formalized even in a propositional modal logic Adam *et al.* [2009], and every definition

in such a logic can be easily encoded in a cognitive calculus.

Irreversibility of Consciousness

Irr Phenomenal consciousness in human persons is irreversible.

Remark That which it feels like to you to experience a moving scene in Verdi's *MacBeth* over some interval of time cannot even conceivably be "lived out in reverse." Of course, we hardly expect our bald assertion here regarding this example to be compelling. Skeptics can consult Bringsjord & Zenzen [1997]. Patrick Suppes [2001] can be viewed as being aligned with our position, at least to a degree, since he admits that from a conscious, common-sense point of view, even physical processes don't appear to be reversible (despite the fact that they are from the standpoint of both classical and quantum particle mechanics). As to an AI connection, note that, taking care to align themselves with formal accounts of intelligent agents based on inductive learning [e.g. Hutter, 2005], Maguire *et al.* [2016] present an account of consciousness as the compression of data. While we are not prepared to affirm the claim that consciousness at heart *consists* in the capacity to compress data, but such compression leads to irreversibility.

Freedom

Free Humans persons perceive, internally, that: they can decide to do things (strictly speaking, to *try* to do things), where these decisions aren't physics-caused (in accordance with some physics theory $\mathscr{C}$) by any prior events, and where such decisions are the product itself of a decision on that same person's part.

Remark While Bringsjord is an unwavering proponent of the Chisholmian view that contra-causal (or — to use the other term with which this view has traditionally been labeled — libertarian) freedom is in fact enjoyed by human persons [e.g. as defended in Bringsjord, 1992a], our tack here is more ecumenical: We "back off" from the proposition that free agents are those who can make decisions that are in some cases not physics-caused by prior events/phenomena, but are caused by the agents themselves. **Free** asserts only that agents *perceive* that such a situation holds. (Thus we don't even insist that agents *believe* they are contra-causal free.) Perception here is of the internal variety, and the actions in question are restricted to inner, mental events, namely decisions. In addition, axiom **Free** leaves matters open as to which physics theory $\mathscr{C}$ of causation the agent perceives to be circumvented by the agent's own internal powers of self-determination. We assume only that any instantiation to $\mathscr{C}$ is itself an axiom system; this in principle opens the door to seamless integration and exploration of the combination of $\mathcal{CA}$ and $\mathscr{C}$.

Common Knowledge of Causation

CCaus Where $\mathscr{C}$ is some physics theory, it's common belief among human persons

that $\mathscr{C}$ holds.

Remark Numerous formal models of time, change, and causation have been presented in the literature, even if we restrict ourselves to the AI literature. Numerous accounts of causation are available from physics itself; this is of course why we have availed ourselves of the placeholder $\mathscr{C}$, which could for example be instantiated to an axiomatization of classical mechanics or special relativity. In much prior work, and indeed in the simulations below, we have found it convenient and productive to employ one particular model of time, change, and causation: a naïve, folk-psychological one based on the *event calculus*, first employed in [Arkoudas & Bringsjord, 2009].

De Se Propositions

TheI Every human person has a special self-designator I^* (normally invoked by use of the first-person pronoun "I") whose sense is this person's essence E (see the **Ess** axiom), and human persons know from a first-person perspective that they have Cartesian properties; and having such properties does not entail that from a first-person perspective that the person has any contingent, non-Cartesian properties.

Remark This axiom directly reflects the position of Perry [1979, 1977] on the special status of first-person knowledge; it specifically asserts that there is a form of self-knowledge (and perhaps merely self-belief) that doesn't entail that the self has any physical, contingent properties, and also asserts that all the agents within the purview of $\mathcal{CA}$ do indeed know such things about themselves. I^* is also inspired by [Castañeda, 1999], a work that peerlessly explains both the need to have a symbol for picking out each self as separate from all else. In prior AI work, we have further developed and implemented this axiom by using a self-designator in a cognitive calculus for robots; see e.g. [Bringsjord *et al.*, 2015].

Planning

Plan Human persons, knowing their initial situations, can reason by proof or argument to plans which, if executed in these situations, at least sometimes secure the goals they seek.

Remark In AI, Planning is a longstanding and vast sub-discipline. Accordingly, any number of specifications of **Plan** are possible, even under the constraint, which we embrace, that planning must be logic-based. The simplest AI-flavored specification of **Plan** is probably one that closely follows "Green's Method" [succinctly covered in Genesereth & Nilsson, 1987], in which, in order to find a plan given some initial situation σ and goal γ, the agent in question seeks a proof that a series of actions can be carried out starting in σ such that γ eventuates. Given that TCC places heavy emphasis on theory-of-mind constructs, any genuinely plausible fleshing out of **Plan** inspired by Green's Method would minimally need to be a radical extension of this method so that, e.g., goals can be purely cognitive in nature. E.g., an agent

may set itself the goal of getting some other agent to have an iterated belief, or to have knowledge of its knowledge of the knowledge of some other agent, etc.

Communication	
Comm	Human persons have the capacity to communicate back and forth over time with other agents by generating to these agents content that corresponds to formulae in one or more formal language of cognitive calculi, and by understanding content received as formulae in one or more of these formal languages.
Remark	In AI terms, generation is in the sub-discipline of NLG (Natural Language Generation), and understanding is in the sub-discipline of NLU (Natural Language Understanding). We are thus committed to pursuing NLP in AI on the basis of formal logic (specifically cognitive calculi), and regard the meaning of natural language to correspond to formulae within the context of proofs and arguments that contain these formulae. For a very recent (albeit brief) treatment of natural language in connection with proof-theoretic semantics, see [Francez, 2015, Par II].

6. A Concise Definition of Λ (Lambda)

We now present a concise definition of Λ. The definition is based on the four components described immediately below.

Components of Λ

Agents A finite set of agents $A = \{a_1, \ldots, a_n\}$.

Time A set of timepoints $T = \{t_0, \ldots, t_n, \ldots\}$.

Logic A cognitive calculus $L \equiv \langle \mathcal{L}, \mathcal{I} \rangle$ with at least a first- or second-order core $\langle \mathcal{L}^0, \mathcal{I}^0 \rangle$.

Interface Agents are modeled as black boxes that consume a set of expressions in the logic and produce a set of expressions. We have *interface functions* $i, o : A \times T \rightarrow 2^{\mathcal{L}}$ that give us the inputs agents consume at different timepoints and the outputs they produce.

Λ provides a measure of the degree of cognitive consciousness for an agent at a time (and over intervals composed of such times), and does so by first appropriating standard $\Delta/\Sigma/\Pi$ measures of the complexity of purely extensional formulae in logics like first- and second-order logic. Given any measure $\mu : \mathcal{L}^0 \rightarrow \mathbb{N}$ of pure first- or second-order formulae $\phi \in \mathcal{L}^0$, we can extend the measure to a new measure $\mu_\omega : \mathcal{L} \rightarrow \mathbb{N}$ that applies to any formula in $\mathcal{L}$:

Extending Measures from $\mathcal{L}^0$ to $\mathcal{L}$

$$\mu_\omega(\phi) = \begin{cases} \mu(\phi) & \text{if } \phi \in \mathcal{L}^0 \\ \max_\psi \mu_\omega(\psi) + 1 & \text{if } \phi \equiv \omega_i[a_1, t_1, \ldots \psi \ldots] \end{cases}$$

For example, let μ count the number of predicate symbols in a formula.

Example

$$\begin{aligned} \mu(Happy(john)) &= 1 \\ \mu_\omega\big(Happy(john)\big) &= 1 \\ \mu_\omega\Big(\mathbf{B}\big(mary, t_2, Happy(john)\big)\Big) &= 2 \end{aligned}$$

For any agent a, we want to look at the new complexity the agent introduces that is above any input complexity. For this, we introduce $\Delta : 2^{\mathcal{L}} \times 2^{\mathcal{L}} \to 2^{\mathcal{L}}$ operator that computes differences between two sets of formulae. This can be simply the set-difference operator. For convenience, let $\omega_j\big[\Gamma\big]$ denote the subset of formulae with operators ω_j in Γ:

$$\omega_j\big[\Gamma\big] = \{\phi \mid \ \phi \equiv \omega_j[\ldots] \text{ and } \phi \in \Gamma \text{ or } \phi \text{ a subformula } \in \Gamma\}$$

Given a set of measures $\{\mu^0, \ldots, \mu^N\}$ and a set of modal (or cognitive) operators $\{\omega_0, \ldots, \omega_M\}$, we define Λ as a function mapping an agent at a time point to a matrix $\mathbb{N}^{M \times N}$:

$$\Lambda : A \times T \to \mathbb{N}^{M \times N}$$

Definition of Λ

$$\Lambda(a,t)_{i,j} = \max_\phi \Big\{\mu^i(\phi) \mid \phi \in \Delta\Big(\omega_j\big[o(a,t)\big], \omega_j\big[i(a,t)\big]\Big)\Big\}$$

Example 2

Let us consider two modal operators $\{\mathbf{B}, \mathbf{D}\}$ and the following base measures μ^0 which measures quantificational complexity via Σ or Π measures, μ^1 which counts the total number of predicate symbols (not a count of unique predicate symbols), and μ^2 which counts the number of distinct time expressions. This gives $\Lambda : A \times T \to \mathbb{N}^{2\times 3}$. At some timepoint t, let an agent a have the following $\Delta(o(a,t), i(a,t)) = \{\mathbf{B}(\phi_1), \mathbf{D}(\phi_2)\}$

$$\phi_1 \equiv \neg\forall a : Happy(a,t); \qquad \phi_2 \equiv \forall b : \neg Hungry(b,t) \rightarrow Happy(b,t)$$

Applying the measures:

$$\mu^o(\phi_1) = 1, \mu^1(\phi_1) = 1; \mu^2(\phi_1) = 1$$
$$\mu^o(\phi_2) = 1; \mu^1(\phi_2) = 2; \mu^2(\phi_2) = 1$$

Giving us:

$$\Lambda(a,t) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix}$$

6.1. *Some Distinctive Properties of Λ (vs. Φ)*

Here are some properties of the Λ framework of potential interest to our readers:

Non-Binary Whereas Φ is such that an agent either is or is not (P-) conscious, cognitive consciousness as measured by Λ admits of a fine-grained range of the *degree* of cognitive consciousness.

Zero Λ for Some Animals and Machines Animals such as insects, and computing machines that are end-to-end statistical/connectionist "ML," have zero Λ, and hence cannot be cognitively conscious. In contrast, as emphasized to Bringsjord in personal conversation,[6] Φ says that even lower animals are conscious.

Human-Nonhuman Discontinuity Explained by Λ From the computational/AI point of view, cognitive scientists have taken note of a severe discontinuity between *H. sapiens sapiens* and other biological creatures on Earth [Penn *et al.*, 2008], and the sudden and large jump in level of Λ from (say) chimpanzees and dolphins to humans is in line with this observation. It's for instance doubtful that any nonhuman animals are capable of reaching third-order belief; hence $\Lambda[\mathbf{B}, 0] = n$, where $n \geq 3$, for any nonhuman animal, is impossible. In stark contrast, each of us believes that you, the reader, believe that we believe that San Francisco is located in California.

Human-Human Discontinuity Explained by Λ A given neurobiologically normal human, over the course of his or her lifetime, has very different cognitive capacity. E.g., it's well-known that such a human, before the age of four or five, is highly unlikely to be able to solve what has become known as the *false-belief task* (or sometimes the *sally-anne task*), which we denote by 'FBT.' From the point of view of Λ, the explanation is simply that an agent with insufficiently high cognitive consciousness is incapable of solving such a task; specifically, solving FBT requires an agent to have

[6] With Tononi and C. Koch, SRI T&C Series.

beliefs about the beliefs of other agents, where the target of those beliefs involves at least basic quantification.[7]

Non-existent Agents Can Have Cognitive Consciousness

In fact, such agents can have *high* levels of cognitive consciousness. For example, brilliant fictional detectives, such as Poe's remarkable C. Auguste Dupin, can be shown to have high levels of cognitive consciousness. E.g., in "The Purloined Letter" Dupin exploits his ability to infer what logic dictates he should believe about the criminal's beliefs about the beliefs of detectives investigating said detective.

7. Simulations and Measurements

Now we model the cognition of four different agents in $\mathcal{DCEC}^*$ and measure the Λ of their cognition as so modeled. For each model, we are given a chunk of reasoning in the form of a set of assumptions Γ that are used to prove a goal ϕ at some time t by the agent a under consideration. We assume that $\Delta\big(o(a,t), i(a,t)\big) = \Gamma + \phi$. We have measures $\{\mu^0, \mu^1, \mu^2, \}$ as defined above in Example 2 and modal operators $\langle \mathbf{K}, \mathbf{B}, \mathbf{P}, \mathbf{O}, \mathbf{I} \rangle$. We also add two new measures: μ^3, that counts the number of occurrences of the *de se* $*$ operator; and μ^4, that counts the number of nested modal operators.

7.1. *Le Chevalier Auguste Dupin and "The Purloined Letter"*

Our first simulation is a situation from Poe's *The Purloined Letter* involving the first fictional detective, the aforementioned Dupin. In this simulation, Dupin d has to figure out where the minister m has hidden an eponymous letter he has stolen from g. The beliefs shown are from d's point of view. g believes that m has hidden the letter in an elaborate manner. d initially believes that m has hidden the letter either in an elaborate or plain manner. Since m does not wish g to find the letter, if m believes that g believes that the letter will be hidden plainly, then m will hide the letter in an elaborate manner (and *vice versa*). m believes that g believes that the letter is hidden elaborately. d believes the previous statement. From these statements, d then derives that the letter is hidden in a plain fashion. This situation is shown in Figure 3 and the corresponding Λ measurements are shown below in Equation 7.1:

$$\Lambda(a,t) = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 0 & 3 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{7.1}$$

[7] An artificial agent able to solve FBT is presented by Arkoudas & Bringsjord [2009]; the agent uses an early cognitive calculus.

```
{:name        "The Purloined Letter"
 :description "Dupin's reasoning as he goes through the case"

 :assumptions {1 (Believes! g t1 (hide m elaborate))
               2 (Knows! d  t1 (exists ?method (and (hide m ?method)
                                                    (or (= ?method plain)
                                                        (= ?method elaborate)))))
               3 (Believes! m t1 (Believes! g t2  (hide m elaborate)))
               4 (if (Believes! t1 m (Believes! g t2 (hide m elaborate))) (hide m plain))
               5 (if (Believes! m t1 (Believes! g t2 (hide m plain))) (hide m elaborate))
               6 (Believes! m (Believes! g (hide m elaborate)))
               7 (Believes! d t1 (if (Believes! m t2 (Believes! g (hide m elaborate))) (hide m plain)))
               8 (Believes! d t1 (if (Believes! m t2 (Believes! g t3 (hide m plain))) (hide m elaborate)))
               9 (Believes! d t1 (Believes! m  t2 (Believes! g t3 (hide m elaborate))))}

 :goal        (Believes! d t4 (hide m plain))}
```

Fig. 3. Shadowprover With the $\mathcal{DCEC}^*$ Cognitive-Calculus Input for Simulating Dupin

7.2. *Descartes and the Cogito*

In this simulation, we analyze Descartes' famous *"Cogito, ergo sum"*. S1 and S2 state that the agent believes everything is either a name or a thing, but not both . S3 states that things are either real or fictional, but not both. A1 links names and things via the $*$ function. A2 defines *DeRe* existence by stating that if y is a name then the denotation of y has *DeRe* existence if it is real. Then the agent supposes it does not *DeRe* exist. The final premise is an instantiation of the statement that if an agent perceives itself believing something, then the agent is real. The agent then perceives the previous belief. From these beliefs, the agent derives a contradiction as Descartes did. Applying our definition of Λ, we get the values in Equation 7.2:

$$\Lambda(a,t) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 3 & 1 & 3 & 1 & 2 \\ 1 & 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{7.2}$$

7.3. *An "Ethical" AI*

Figure 5 shows a simulation of an agent that is in the process of computing the Doctrine of Double Effect (DDE) [in a particular scenario; see Govindarajulu & Bringsjord, 2017b], a complex ethical principle that applies in moral dilemmas.[8] DDE computation happens with a background context denoted by the formula *situation*. Therefore, Λ measurements shown below in Equation 7.3 are relative to *situation*.

[8]The principle is worth studying from the perspective of cognitive science, as there have been empirical studies that humans employ this principle.

```
{:name        "Cogito Ergo Sum"
 :description "A formaliztion of Descartes' 'Cogito, Ergo Sum'"
 :assumptions {S1                (Believes! I t1 (forall [x] (or (Name x) (Thing x))))
               S2                (Believes! I t1 (forall (x) (iff (Name x) (not (Thing x)))))
               S3                (Believes! I t1 (forall (x) (if (Thing x) (or (Real x) (Fictional x)))))
               S4                (Believes! I t1 (forall (x) (if (Thing x) (iff (Real x) (not (Fictional x))))))
               A1                (Believes! I t1 (forall (x) (if (Name x) (Thing (* x)))))
               A2                (Believes! I t1
                                            (forall (y)
                                                    (if (Name y)
                                                      (iff (DeReExists y)
                                                           (exists x (and (Real x) (= x (* y))))))))
               Suppose           (Believes! I t1 (not (DeReExists I)))


               given             (Believes! I t1 (Name I))

               ;;;
               Perceive-the-belief (Believes! I t1 (Perceives! I t2 (Believes! I t3 (not (DeReExists I)))))
               If_P_B

               (Believes!
                I t1
                (forall [?agent]
                        (if (Perceives! t2 I (Believes! t3 ?agent (not (DeReExists ?agent))))
                          (Real (* ?agent)))))}
 :goal       (and (Believes! I t1 (not (Real (* I))))
                  (Believes! I t1 (Real (* I))))}
```

Fig. 4. Shadowprover Input With the $\mathcal{DCEC}^*$ Cognitive Calculus for Simulating Descartes in "*Cogito, Ergo Sum*"

$$\Lambda(a,t) = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 2 & 0 & 0 \\ 2 & 1 & 2 & 0 & 0 \end{pmatrix} \tag{7.3}$$

```
{:name        "DDE base"
 :description "DDE"
 :assumptions {I2
               (Ought! I now situation
                       (and (not (exists [?t] (HoldsAt (dead P1) ?t)))
                            (not (exists [?t] (HoldsAt (dead P1) ?t)))))

               I3
               (Knows! I now situation)

               I4
               (Believes! I t0
                          (Ought! I now situation
                                  (and (not (exists [?t] (HoldsAt (dead P1) ?t)))
                                       (not (exists [?t] (HoldsAt (dead P1) ?t)))))))}

 :goal        (Intends! I now
                        (and (not (exists [?t] (HoldsAt (dead P1) ?t)))
                             (not (exists [?t] (HoldsAt (dead P1) ?t)))))}
```

Fig. 5. Shadowprover Input With the $\mathcal{DCEC}^*$ Cognitive Calculus for Simulating a Fragment of the Doctrine of Double Effect

8. Conclusion and Future Work

Using the logicist foundation afforded by cognitive calculi, we have presented a definition of a family of measures of cognitive consciousness that is precise and extensible (i.e. Λ), where those measures are undergirded by a dozen axioms ($\mathcal{CA}$) that together characterize cognitive consciousness; and we have placed these things within a context that includes a rejection of non-cognitive p-consciousness-centric IIT/Φ. One of the key ingredients missing in the current formal account of Λ is efficiency. Two similar systems might have the same Λ measure but one system might be using resources (such as time, memory, etc.) more efficiently in its computation. In future work, we will begin to address this by including additional dimensions in Λ that incorporate standard stratification of computational efficiency. A second current deficiency to remedy in future expansion and refinement of the theory of cognitive consciousness is the fact that external perception, for instance vision, doesn't figure in the modeling and simulation thus far produced on the basis of TCC/Λ/$\mathcal{CA}$. G. [2008] has emphasized vision in seeking to explain the core ideas underlying IIT, and logic-based machine vision, anchored to our cognitive calculi, will — albeit only steadfastly in our case at the human level; G. [2008] seeks to exploit a thought experiment about a single photodiode "seeing" — be possible, by using humanoid robots, in a manner inspired in no small part by [Chella *et al.*, 2019].

Acknowledgments

We are deeply appreciative for support from AFOSR that has enabled the development of cognitive calculi of sufficient power to express forms of high computational intelligence (calculi which, as the reader will doubtless have noticed, we have deployed in order to establish TCC and Λ; again, see Appendix A for further information), and for support from ONR that has made possible our sustained pursuit of specific kinds of artificial cognition, including, in particular, cognitive consciousness in artificial agents.

Appendix A. Cognitive Calculi

A.1. *What is a Cognitive Calculus? And Why is It So Named?*

What is a cognitive calculus $\mathcal{C}$, and why is it denoted with the two words in question? (We use '$\mathcal{C}$' here as an arbitrary variable ranging over (the uncountably infinite space of) all cognitive calculi). In keeping with the mathematical-logic literature [e.g. Ebbinghaus *et al.*, 1994][9], we first take a *logical system* $\mathscr{L}$ to be a triple $\langle \mathcal{L}, \mathcal{I}, \mathcal{S} \rangle$ where $\mathcal{L}$ is a (often) sorted/typed formal language (based therefore on an alphabet and a formal grammar), $\mathcal{I}$ is a set of natural[10] inference schemata, and $\mathcal{S}$ is a formal semantics of some sort. For example, the familiar propositional calculus comprises a family of simple logical systems; the same holds for first-order logic; both families are of course at the heart of AI.[11] In the case of both of these families, a frequently included particular inference schema is *modus ponens*, that is

$$\frac{\phi \rightarrow \psi, \phi}{\psi} \quad I_{\text{MP}}$$

And in the case of the latter family, often *universal introduction* is included in a given $\mathcal{I}$; a specification of this inference schema immediately follows.[12]

$$\frac{\phi(\text{a})}{\forall x \phi(\frac{a}{x})} \quad I_{\text{UI}}$$

Note that both of the two inference schemata just shown are included in the particular cognitive calculus we use in the present paper for modeling, and as a framework for automated reasoning. Note as well that both $\mathscr{L}_{\text{PC}}$ and $\mathscr{L}_1$ are *extensional*, which means essentially that the meaning of any formula ϕ in the relevant languages are given by compositional functions operating solely on the internal components of ϕ. If we for example know that ϕ is FALSE, then we know that the meaning of $\phi \rightarrow \psi$ is TRUE, for any ψ in the language, for both of these logical systems.

Moving from the concept of a logical system to that of a cognitive calculus is straightforward, and can be viewed as taking but three steps, to wit:

S1 Expand the language of a logical system to include

[9]Note in particular coverage in this excellent work of Lindström's Theorems, which pertain to the properties of certain logical systems (e.g., completeness).
[10]Hence when the schemata are deductive in nature, we specifically have natural deduction.
[11]As can be confirmed by looking to the main textbooks of the field. E.g. see Russell & Norvig [2009]; Luger [2008].
[12]The standard provisos apply here to the constant a.

i modal operators that represent one or more mental verbs at the human level standardly covered in human-level cognitive psychology [e.g. see any standard, comprehensive textbook on human-level cognitive psychology, such as Ashcraft, 1994; Goldstein, 2008], and regarded to be so-called "propositional attitudes" that give rise to propositional-attitude-reporting sentences, where these sentences are represented by operator-infused formulae in a cognitive calculus.[13] Such verbs include: *knowing*, *believing*, *deciding*, *perceiving*, *communicating*,[14] *desiring*, and *feeling* X where 'X' denotes some emotional state (e.g. possible $X = sad$, and so on. Note that such verbs break the bounds of extensionality, and hence make any logic that captures them an *intensional* logic.[15] Step S1.i is the reason why we speak of a *cognitive* calculus.

ii meta-logical expressions (such as that from a set Φ of formulae a particular formula ϕ can be proved: $\Phi \vdash \phi$). Hence cognitive calculi are not merely object-level elements of logics, but include meta-logical elements as well. E.g. a cognitive calculus can have a meta-conditional saying that if some provability expression such as $\Phi \vdash \phi$ holds, then ϕ holds. Step S1.ii is a necessary, preparatory step for S2.

S2 Delete $\mathcal{S}$; if desired, move selected elements of $\mathcal{S}$ into $\mathcal{I}$, which requires casting these elements as inference schemata that employ meta-logical expressions secured by prior step S1.ii. S2 reflects the fact that cognitive calculi have purely *inferential* semantics, and hence are aligned with the tradition of *proof-theoretic semantics* Gentzen [1935]; Prawitz [1972]; Schroeder-Heister [2012/2018]. (In particular, cognitive calculi thus do not employ possible-worlds semantics for modal operators. In possible-worlds approaches, e.g., *knows* doesn't get defined as justified true belief; but as we explained in the paper proper, knowledge in a cognitive calculus holds iff the agent in question believes the known proposition on the strength of a proof or argument.) We might for instance wish to include an inference schema that regiments the idea that an agent knows that which is provable from what she knows. Step S2 is the reason why we speak of a cognitive *calculus* (instead of e.g.

[13]The attitudes are covered nicely in Nelson [2015]. Here's an informative quote from this work:

> Propositional attitude reporting sentences concern cognitive relations people bear to propositions. A paradigm example is the sentence 'Jill believes that Jack broke his crown.' Arguably, 'believes,' 'hopes,' and 'knows' are propositional attitude verbs and, when followed by a clause that includes a full sentence expressing a proposition (a that-clause) form propositional attitude reporting sentences. [Nelson, 2015, ¶1]

[14]Due to lack of space, we leave aside our approach to formal NLP on the basis of proof theory alone. For a truly excellent book on proof-theoretic semantics, including, natural language, we recommend Francez [2015].

[15]This fact is discussed in some detail in Bringsjord & Govindarajulu [2012], and is replete with relevant proofs. As an example, note that the truth or falsity of 'Jones believes that ϕ' is not determined by the truth or falsity of ϕ, since humans routinely believe that falsehoods hold.

a cognitive *logic*, or cognitive logical system).

S3 Expand $\mathcal{I}$ as needed to include inference schemata that involve the operators from S1.i. For instance, where $\mathbf{K}$ is the modal operator for 'knows' and $\mathbf{B}$ for 'believes,' we might (and in learning *ex nihilo*, for reasons explained in the paper proper, we do) wish to have this inference schema in a given $\mathcal{C}$:

$$\frac{\mathbf{K}\phi}{\mathbf{B}\phi} \quad I_{\text{KB}}$$

A.2. *Regarding Related Work*

Much could be said about work/systems that are related to cognitive calculi, but sustained treatment of this issue is out of scope in this brief appendix, which is merely meant to supplement the paper coming before it. We will say only a few things, and hope they are at least somewhat enlightening; here goes. The first published, implemented cognitive calculus, a multi-operator modal logic (minus, by definition, and as explained earlier in the present appendix, any conventional semantics) based on multi-sorted first-order logic, can be found in Arkoudas & Bringsjord [2008, 2009]; the second of these publications is a refinement of the first. Implementation at that point was based upon Athena, a recent introduction to which, along with a study of proof methods in computer science, is providedn in the excellent Arkoudas & Musser [2017]. Related work as cited in this earlier work remains relevant over a decade later, and in particular, so-called "BDI logics" [e.g. see Rao & Georgeff [1991]] are related, and we applaud their advent — but as explained in 2009/2009, such logics cover very few propositional attitudes present in adult and neurobiologically normal cognition (e.g. no communication operators, and no emotional states), and are not based on purely inferential semantics. Automated reasoning in the tradition of higher-order logic (HOL) as descended from Frege, and most prominently from Church, which is masterfully chronicled in Benzmüller & Miller [2014], is obviously related to cognitive calculi; this is especially true since HOL is now very much on the scene in 21st-century AI [e.g. see Benzmüller & Paleo [2016]]. In contrast, cognitive calculi, and the automation thereof, are based on committments guided by the study of human cognition; and as we see it, that cognition for matters formal and extensional is for the most part circumscribed by natural deduction in second-order logic in the complete absence of formal semantics [e.g. consider the raw material in the practice of mathematics that gives rise to the argument and analysis in Shapiro [1991]], and in matters literary circumscribed by modal operators mixed with third-order logic [e.g. see Bringsjord *et al.* [2016]]. Traditionally, in terms of the Frege-to-Church-to... history that HOL has, HOL is extensional; in contrast, cognitive calculi by definition cannot fail to have operators that cover human cognition. The final thing we mention here is that cognitive calculi are not in any way deductive and bivalent or trivalent; they can be infused with

uncertainty, and have multiple values; see e.g. Govindarajulu & Bringsjord [2017c].

A.3. *What About the Metatheory of Cognitive Calculi?*

Space limitations preclude the presentation of the proof-theoretic metatheory of cognitive calculi, but of course such things as soundness theorems are possible, and some have been proved. For instance, the cognitive calculus $\mathcal{CC}$, markedly simpler than $\mathcal{DCEC}^*$ (it e.g. has on the intensional-operator side only epistemic operators) a soundness theorem has been achieved. Completeness, note, is a very different story: since for example full/standard second-order logic (SOL) is incomplete, and many cognitive calculi have as part of their extensional cores SOL, these cognitive calculi are incomplete.

References

A., K. [2006] Panexperientialism, Cognition, and the Nature of Experience, *Psyche* **12**(5), 1–14.

Adam, C., Herzig, A. and Longin, D. [2009] A Logical Formalization of the OCC Theory of Emotions, *Synthese* **168**(2), 201–248.

Arkoudas, K. and Bringsjord, S. [2008] "Toward formalizing common-sense psychology: An analysis of the false-belief task," in T.-B. Ho & Z.-H. Zhou (eds.), *PRICAI 2008: Trends in Artificial Intelligence* (Springer Berlin Heidelberg, Berlin, Heidelberg), ISBN 978-3-540-89197-0, pp. 17–29.

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

Arkoudas, K. and Musser, D. [2017] *Fundamental Proof Methods in Computer Science: A Computer-Based Approach* (MIT Press, Cambridge, MA).

Ashcraft, M. [1994] *Human Memory and Cognition* (HarperCollins, New York, NY).

Benzmüller, C. and Miller, D. [2014] Automation of Higher-Order Logic, in *Handbook of the History of Logic; Volume 9: Logic and Computation* (North Holland, Amsterdam, The Netherlands).

Benzmüller, C. and Paleo, B. W. [2014] "Automating Gödel's Ontological Proof of God's Existence with Higher-order Automated Theorem Provers," in T. Schaub, G. Friedrich & B. O'Sullivan (eds.), *Proceedings of the European Conference on Artificial Intelligence 2014 (ECAI 2014)* (IOS Press, Amsterdam, The Netherlands), pp. 93–98, http://page.mi.fu-berlin.de/cbenzmueller/papers/C40.pdf.

Benzmüller, C. and Paleo, B. W. [2016] "The Inconsistency in Goödel's Ontological Argument: A Success Story for AI in Metaphysics," in S. Kambhampati (ed.), *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)* (AAAI Press), pp. 936–942, https://www.ijcai.org/Proceedings/16/Papers/137.pdf.

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

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

Bringsjord, S. [1992a] Free Will, in *What Robots Can and Can't Be* (Kluwer, Dordrecht, The Netherlands), pp. 266–327.

Bringsjord, S. [1992b] *What Robots Can and Can't Be* (Kluwer, Dordrecht, The Netherlands).

Bringsjord, S. [1998] Chess is Too Easy, *Technology Review* **101**(2), 23–28, http://kryten.mm.rpi.edu/SELPAP/CHESSEASY/chessistooeasy.pdf.

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, http://kryten.mm.rpi.edu/jcsonebillion2.pdf.

Bringsjord, S. [2008] The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself, *Journal of Applied Logic* **6**(4), 502–525, http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf.

Bringsjord, S. [2015] Could a Robot Be a Bona Fide Hero? https://www.youtube.com/watch?v=2Y5eC9Vp5Do, The URL here goes to the video of Bringsjord's TED_{X} talk in Limassol, Cyprus.

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

Bringsjord, S. and Govindarajulu, N. S. [2012] Given the Web, What is Intelligence, Really? *Metaphilosophy* **43**(4), 361–532, http://kryten.mm.rpi.edu/SB\NSG\Real\Intelligence\040912.pdf, This URL is to a preprint of the paper.

Bringsjord, S. and Govindarajulu, N. S. [2013] Toward a Modern Geography of Minds, Machines, and Math, in V. C. Müller (ed.), *Philosophy and Theory of Artificial Intelligence*, Studies in Applied Philosophy, Epistemology and Rational Ethics, Vol. 5 (Springer, New York, NY), ISBN 978-3-642-31673-9, pp. 151–165, doi:10.1007/978-3-642-31674-6_11, http://www.springerlink.com/content/hg712w4l23523xw5.

Bringsjord, S. and Govindarajulu, N. S. [2018] Artificial Intelligence, in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/artificial-intelligence.

Bringsjord, S., Licato, J. and Bringsjord, A. [2016] The Contemporary Craft of Creating Characters Meets Today's Cognitive Architectures: A Case Study in Expressivity, in J. Turner, M. Nixon, U. Bernardet & S. DiPaola (eds.), *Integrating Cognitive Architectures into Virtual Character Design* (IGI Global, Hershey, PA), pp. 151–180.

Bringsjord, S., Licato, J., Govindarajulu, N., Ghosh, R. and Sen, A. [2015] "Real Robots that Pass Tests of Self-Consciousness," in *Proccedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)* (IEEE, New York, NY), pp. 498–504, http://kryten.mm.rpi.edu/SBringsjord_etal_self-con_robots_kg4_0601151615NY.pdf, This URL goes to a preprint of the paper.

Bringsjord, S. and Zenzen, M. [1997] Cognition is not Computation: The Argument from Irreversibility? *Synthese* **113**, 285–320.

Castañeda, H.-N. [1999] *The Phenomeno-Logic of the I: Essays on Self-Consciousness* (Indiana University Press, Bloomington, IN), This book is edited by James Hart and Tomis Kapitan.

Cerello, M. [2015] The Problem with Phi: A Critique of Integrated Information Theory, *PLoS Computational Biology* **11**(9), https://www.ncbi.nlm.nih.gov/

pmc/articles/PMC4574706/, DOI: 10.1371/journal.pcbi.1004286.

Chella, A., Cangelosi, A., Metta, G. and Bringsjord, S. (eds.) [2019] *Consciousness in Humanoid Robots* (Frontiers, Lausanne, Switzerland), ISBN 978-2-88945-866-0, ISSN 1664-8714; DOI 10.3389/978-2-88945-866-0.

Chella, A. and Manzotti, R. (eds.) [2007] *Artificial Consciousness* (Imprint Academic, Exeter, UK).

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

Francez, N. [2015] *Proof-theoretic Semantics* (College Publications, London, UK).

G., T. [2008] Consciousness as Integrated Information: A provisional Manifesto, *The Biological Bulletin* **215**, 216–242.

Genesereth, M. and Nilsson, N. [1987] *Logical Foundations of Artificial Intelligence* (Morgan Kaufmann, Los Altos, CA).

Gentzen, G. [1935] Investigations into Logical Deduction, in M. E. Szabo (ed.), *The Collected Papers of Gerhard Gentzen* (North-Holland, Amsterdam, The Netherlands), pp. 68–131, This is an English version of the well-known 1935 German version.

Gettier, E. [1963] Is Justified True Belief Knowledge? *Analysis* **23**, 121–123, http://www.ditext.com/gettier/gettier.html.

Goldstein, E. B. [2008] *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience* (Cengage Learning, Boston, MA), This is the 5th edition.

Govindarajulu, N. and Bringsjord, S. [2017a] "On Automating the Doctrine of Double Effect," in C. Sierra (ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)* (International Joint Conferences on Artificial Intelligence), ISBN 978-0-9992411-0-3, pp. 4722–4730, doi:10.24963/ijcai.2017/658, https://doi.org/10.24963/ijcai.2017/658.

Govindarajulu, N. S. and Bringsjord, S. [2017b] "On Automating the Doctrine of Double Effect," in C. Sierra (ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (Melbourne, Australia), pp. 4722–4730, doi:10.24963/ijcai.2017/658, https://doi.org/10.24963/ijcai.2017/658, preprint available at this url: https://arxiv.org/abs/1703.08922.

Govindarajulu, N. S. and Bringsjord, S. [2017c] "Strength Factors: An Uncertainty System for Quantified Modal Logic," in V. Belle, J. Cussens, M. Finger, L. Godo, H. Prade & G. Qi (eds.), *Proceedings of the IJCAI Workshop on "Logical Foundations for Uncertainty and Machine Learning (LFU-2017)* (Melbourne, Australia), pp. 34–40.

Govindarajulu, N. S., Bringsjord, S., Ghosh, R. and Peveler, M. [2019] Beyond the doctrine of double effect: A formal model of true self-sacrifice, in *Robotics and Well-Being* (Springer), pp. 39–54.

Honerich, T. [2014] *Actual Consciousness* (Oxford University Press, Oxford, UK).

Hutter, M. [2005] *Universal Artificial Intelligence: Sequential Decisions Based on*

Algorithmic Probability (Springer, New York, NY).

Jacquette, D. [2015] Review of Honderich's *Actual Consciousness*, *Notre Dame Philosophical Reviews* **8**, http://ndpr.nd.edu/news/60148-actual-consciousness.

James, W. [1904] Does 'Consciousness' Exist? *Journal of Philosophy, Psychology, and Scientific Methods* **1**, 477–491.

Johnson-Laird, P. and Oatley, K. [1989] The Language of Emotions: An Analysis of a Semantic Field, *Cognition and Emotion* **3**(2), 81–123.

Kriegel, U. [2015] *Varieties of Consciousness* (Oxford University Press, Oxford, UK).

Luger, G. [2008] *Artificial Intelligence: Structures and Strategies for Complex Problem Solving (6th Edition)* (Pearson, London, UK), ISBN 0321545893.

Luger, G. and Stubblefield, W. [1993] *Artificial Intelligence: Structures and Strategies for Complex Problem Solving* (Benjamin Cummings, Redwood, CA).

M., O., L., A. and G., T. [2014] From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0, *Computational Biology* **5**(10), 1–25.

Maguire, P., Moser, P. and Maguire, R. [2016] Understanding Consciousness as Data Compression, *Journal of Cognitive Science* **17**(1), 63–94, http://www.cs.nuim.ie/~pmaguire/publications/Understanding2016.pdf.

McCarthy, J. and Hayes, P. J. [1969] Some Philosophical Problems from the Standpoint of Artificial Intelligence, in B. Meltzer & D. Michie (eds.), *Machine Intelligence 4* (Edinburgh University Press), pp. 463–502.

Mueller, E. [2006] *Commonsense Reasoning: An Event Calculus Based Approach* (Morgan Kaufmann, San Francisco, CA), This is the first edition of the book. The second edition was published in 2014.

Nelson, M. [2015] Propositional Attitude Reports, in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/prop-attitude-reports.

Oizumi, M., Albantakis, L. and Tononi, G. [2014] From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0, PLOS *Computational Biology* https://doi.org/10.1371/journal.pcbi.1003588.

Ortony, A., Clore, G. L. and Collins, A. [1988] *The Cognitive Structure of Emotions* (Cambridge University Press, Cambridge, UK).

Penn, D., Holyoak, K. and Povinelli, D. [2008] Darwin's Mistake: Explaining the Discontinuity Between Human and Nonhuman Minds, *Behavioral and Brain Sciences* **31**, 109–178.

Perry, J. [1977] Frege on Demonstratives, *Philosophical Review* **86**, 474–497.

Perry, J. [1979] The Problem of the Essential Indexical, *Noûs* **13**, 3–22.

Prawitz, D. [1972] The Philosophical Position of Proof Theory, in R. E. Olson & A. M. Paul (eds.), *Contemporary Philosophy in Scandinavia* (Johns Hopkins Press, Baltimore, MD), pp. 123–134.

Rao, A. S. and Georgeff, M. P. [1991] "Modeling Rational Agents Within a BDI-architecture," in R. Fikes & E. Sandewall (eds.), *Proceedings of Knowledge Representation and Reasoning (KR&R-91)* (Morgan Kaufmann, San Mateo, CA), pp. 473–484.

Russell, S. and Norvig, P. [2009] *Artificial Intelligence: A Modern Approach* (Prentice Hall, Upper Saddle River, NJ), Third edition.

Schroeder-Heister, P. [2012/2018] Proof-Theoretic Semantics, in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/proof-theoretic-semantics.

Shapiro, S. [1991] *Foundations Without Foundationalism: A Case for Second-Order Logic* (Oxford University Press, Oxford, UK).

Suppes, P. [2001] Weak and Strong Reversibility of Causal Processes, in M. Galavotti, P. Suppes & D. Costantini (eds.), *Stochastic Causality* (CSLI, Palo Alto, CA), pp. 203–220.

Tononi, G. [2012] *Phi: A Voyage from the Brain to the Soul* (Pantheon, New York, NY).