Toward Axiomatizing Consciousness[*]

Selmer Bringsjord
Paul Bello
Naveen Sundar Govindarajulu

`version 0524172300NY`

Contents

[*] This chapter is dedicated to the philosophically indestructible memory of Dale Jacquette. My (= Bringsjord's) indelible personal memories of Dr. Jacquette while a fellow graduate student at Brown University include the sudden realization during my first year that, wait, hold on, this brilliant, precise, polymathic guy isn't a *professor* in the Department?! Some of Dale's philosophy-of-mind contributions directly impact the following pages (as we sometimes note in the sequel), and many of his other p-o-m contributions do so indirectly, in ways cognoscenti will well apprehend. Dale's serious engagement with the intersection of intensional attitudes and logic, also first appreciated by Bringsjord during Jacquettean Brown days over three decades back, is reflected by the formal language that underlies the system $\mathcal{CA}$ introduced herein.

1 Introduction

With your eyes closed, consider numberhood; put another way, consider this question. What is a number? Presumably if you engage the question in earnest (and if you're not a logician or the like; if you are, bear with us) you will begin by entertaining some small and simple numbers in your head, and then some concepts (e.g. division) inseparably bound up with your simple examples. You might think about, say, the number 463. What kind of number is 463? You will probably agree that it's a so-called whole number. Or you might instead say that 463 is a natural number, or an integer; but we can take these labels to be equivalent to 'whole.' Now, is 463 composite or prime? — and here, if you need to, feel free to open your eyes and resort to paper and a pencil/pen as you endeavor to answer ... Yes, the latter; perhaps you did at least some on-paper division to divine the answer. Very well, now, are there only finitely many primes? This question is a bit harder than our previous one. Put another way, the present query amounts to: Can you keep generating larger and larger primes, forever? If you tire of your own investigation, know that Euclid famously settled the matter with a small but memorably clever *reductio* proof. His verdict, and the supporting details, are readily available on the internet for you to review. ... If you have now developed your own rationale, or searched for and found Euclid's, you know that the answer to our second question is: Yes: there are indeed infinitely many primes.

No doubt you perceive, or have been reminded by virtue of reading the previous paragraph, that there are *a lot* of different *kinds* of numbers, above and beyond the ones in the categories we have mentioned so far. Numberhood is a big, multi-faceted concept! For example, you will recall that there are negative whole numbers, such as -463; that there are fractions with whole numbers on the "top" (numerator) and the "bottom" (denominator), for instance $\frac{1}{463}$; that there are are exotic numbers that can't be captured by any such fraction, for example π and $\sqrt{2}$; and perhaps you will even recall, albeit vaguely if you left such matters long behind in the textbooks of long-past math classes, that there are such numbers as the "complex" and "transfinite" ones. But despite all your thinking about numbers, we very much doubt that you will have arrived at an answer to our original question: What is a number?

For us, consciousness is very much like numberhood, and much of the structure of consciousness, at least as we see that structure, is revealed via the kind of reflection you've just engaged in. In short, we believe that while it's apparently impossible to outright *define* consciousness, an awful lot can be said, systematically, about it, and about the concepts and processes with which it's bound up. You may not know how to specify what a whole number is, but you can nonetheless close your eyes and perceive all sorts of attributes that 463 and its close relatives have, and you can also (at least eventually) prove that 463 is a prime whole number, and that there are an infinite number of prime numbers. We hence already know that you are capable, it seems, of perceiving things that are internal to your mind upon (or at at least often aided by) the closing of your eyes, and that you can also perceive the numeral '463' upon paper in front of you in the external world, when (say) attempting to divide it by a number other than itself or 1.[1] And indeed there are many other things you could quickly prove and come to know (or at least read and come to know), for instance that various sorts of operations on various sorts of numbers work in such and such ways. For example, you know that 463 $\times$ 1 returns 463 back, while 463 $\times 0 = 0$. You know such things in the absence

[1] It would be significantly more accurate to say that the on-paper '463' is a *token* of the abstract type that is the number 463. For such a framework, and its deployment in connection with deductive reasoning over declarative content like what is soon shown below in the specific axioms we propose, see e.g.' (Arkoudas & Bringsjord 2007, Bringsjord 2015).

of a definition of what numberhood is, and what a number is. Likewise, to anticipate one of the 11 axioms of consciousness we shall present later (viz. the one we label**Incorr**), while we have no precise definition of what consciousness is, and no precise definition of what a conscious state is, we're quite sure that if you are considering your own mind, and during that time know that you are deeply sad, you must of necessity believe that you are deeply sad. To anticipate another of the 11 axioms we shall propose (truth be told, we explicitly propose 10, and the 11th, one that relates to planning, is, as we explain, offered as an "option"), the various things we have noted that you *know* about numbers are also (since knowledge implies belief; the axiom below that captures this implication is labeled **K2B**) things that you *believe* about numbers.

In sum, just as you are willing to assert declarative statements about numbers of various sorts, from which additional statements can be deduced, we are willing to make fundamental assertions about consciousness, herein. Neither your assertions nor ours are guaranteed, indeed some of them are bound to be quite controversial, but in both cases at least some progress will have been made, and further progress can presumably be achieved as well. Some of that progress, we are happy to concede, will be won by challenging the very axioms that we propose.

You may have already grown a bit weary of our starting claim that the nature of our inquiry into consciousness can be illuminated by pondering the nature of the inquiry into numberhood, but please bear with us for just a few additional moments. For we also want to bring to your attention that not only is humanity in possession of rather solid understanding of numberhood and numbers of various types, but also humans have managed to set up specific, rigorous *axioms* about numbers. A nice, simple example of a set of such axioms is so-called Peano Arithmetic ($\mathcal{PA}$), quite famous in many quarters.[2] Before we present any of the axioms in $\mathcal{PA}$, we inform you that, where x is any natural number 0, 1, 2, 3, ..., s is the successor function on the set of such numbers; that is, $s(x)$ gives the number that is one larger than x. For instance, $s(463) = 464$. Okay, here then are two of the axioms from $\mathcal{PA}$, in both cases expressed first in the efficient notation of first-order logic (FOL), and then in something close to standard English:

A4 $\forall x\ (x + 0 = x)$

- Every natural number x, when added to 0, equals zero.

A5 $\forall x \forall y\ (x + s(y) = s(x + y))$

- For every pair of natural numbers x and y, x plus the successor of y equals the successor of the sum of x and y.

A4 and A5, relative to our purposes in the present chapter, aren't uninstructive. For example, A4 refers straightaway to the particular number 0; hence $\mathcal{PA}$, it's fair to say, dodges the task of defining what 0 *is*. In fact, no such definition will be found even if the remaining six axioms of $\mathcal{PA}$ are carefully examined. An exactly parallel point can be made about the addition function + that appears in A5: it appears there, and its deployment, given what we all know about addition from real life, makes perfect sense, but there's no definition provided of + in A5 — and the same holds for the other six axioms of $\mathcal{PA}$. In fact, what writers commonly do upon introducing the axiom system $\mathcal{PA}$ is inform their readers that by the symbol + is meant "ordinary addition" — but no

[2] An elegant and economical introduction to $\mathcal{PA}$ is available in (Ebbinghaus, Flum & Thomas 1994), a book which enjoys the considerable virtue of including a presentation of $\mathcal{PA}$ in not only first- but second-order form. Simpler-than-$\mathcal{PA}$ axiomatizations of simple arithmetic over the natural numbers (including the memorably dubbed 'Baby Arithmetic') are introduced in lively fashion in (Smith 2013).

definition of ordinary addition is supplied.[3]

Our objective is in line with the foregoing, for while we despair of ever pinning down the meaning of all forms of consciousness in any formal, third-person format,[4] we nonetheless seek to set out more and more of the third-person *structure* of consciousness. The present chapter is the inauguration, in print, of this pursuit; and it's the set of 11 axioms to be known as '$\mathcal{CA}$,' to be introduced below, that constitutes the first step in the pursuit.

Careful readers will have noticed some specific progress in the pursuit already, since we above implicitly informed you of our commitment to two — as we shall call them — *operators*, one for an agent a's perceiving at a given time t a proposition internal to itself (which has the form $\mathbf{P}^i(a, t, \phi)$), and another ($\mathbf{P}^e$) for external perception of propositions to be found in the environment external to the agent. (We don't literally *see* propositions in the external environment, but we shall simplify matters by assuming that we do precisely that. More about this issue later.) Often it's easy enough to turn that which is externally perceived into corresponding internal percepts. For instance, perhaps you wrote down on a piece of paper in front of you earlier, when acceding to our prompts, something like this:

```
Nope, can't divide in half without leaving a remainder!
```

$$\begin{array}{r}
231 \\
2\,\big)\overline{463} \\
\underline{400} \\
63 \\
\underline{60} \\
3 \\
\underline{2} \\
1
\end{array}$$

But if you did write down something like this for your eyes to see, you were able, immediately thereafter, to perceive, internally, the proposiiton that 463 can't be halved. You might even have been able to internally perceive the details of your long division, or at least some of them. Notice that your perceptions, of both the internal and external varieties, led in this case directly to belief. You came to believe that 463 can't be halved. Of course, it's not *certain* that you didn't make a silly mistake, so your belief might be erroneous, but it's highly unlikely that it is.

Before concluding the introduction, we think it's important for us to emphasize two points. We confess explicitly that our pursuit of an axiomatization of consciousness, as will become painfully clear momentarily, is an exceedingly humble start, intended to be a foundation for subsequent,

[3] Standard model theory presents conditions for the truth of formulae like A4 and A5, but these conditions fail to provide real meaning. E.g., consider a simple formula asserting that the successor of 0 is greater than zero. Given that the domain is the set $\mathbb{N}$ of natural numbers, standard model theory merely assigns TRUE to this formula exactly when the ordered pair $(0, 1)$ is a member of all those pairs (n, m) of natural numbers where $m > n$. The underlying meaning of greater-than/> hasn't been supplied. Indeed, the whole thing is circular and wholly uninformative, since the domain for interpretation is the set of natural numbers — a set that is available *ab initio*.

[4] Indeed, Bringsjord has argued that no third-person account is even logically possible for phenomenal consciousness; e.g., see (Bringsjord 1992*b*, Bringsjord 2007). Jacquette's (1994) book-long defense of property dualism in connection with mental states, relative to Bringsjord's position, is congenial, but Jacquette would find the direct and ineliminable reference to subjects in the axioms presented herein to be problematic, since in the work in question, while countenancing property dualism, he questions a realistic position on the reality of persons as genuine objects.

further progress, including not only the building up of theorems proved from the axioms, but also computational implementation that will allow such proofs to be machine-discovered and machine-verified. Our start is so humble, in fact, that we leave some of our axioms in rather informal form, to ease exposition. Second, our proposed axioms are guaranteed to be highly controversial (as we indicated above). For some of the axioms we propose, we will discuss alternatives that might be more attractive to some readers than our own preferred axioms. But even so, the entire collection of what we propose, alternatives included, will not be universally affirmed; there will be skeptics. But our purpose, again, is to erect a foundation and get the project going in earnest, in a way substantially more robust than others who have tried their hand, at least to a degree, at axiomatizing consciousness. It may be worth pointing out that, as some readers will know, even some axiom systems for things as seemingly cut-and-dry as arithmetic, set theory, and physics are far from uncontroversial.[5]

The remainder of the chapter follows this sequence: We begin (§2) by making it clear that we approach our subject under certain specific constraints. There are four such constraints; each of them is announced, and briefly explained. (One constraint is that we are specifically interested only in *person*-level consciousness (§). Another is that while we agree that perceptual and affective states are enjoyed by humans, our emphasis in on *cognitive* states, or on what might be called *cognitive consciousness* (§).) Next, in section 3, we summarize the work of some researchers who have discussed, and in some real way contributed to, the potential axiomatization of consciousness. The next section (§4) is a very short summary of *some* of the first author's work, in some cases undertaken with collaborators, on the careful representation, and associated mechanization, of some elements of self-consciousness. This prior work, in part, is relied upon in the present investigation. Our next step (§5) is to proceed to the heart of the chapter: the presentation of our 10 proposed axioms for consciousness.[6] A short, concluding section (6), in which we point the way forward from the foundation we have erected, wraps up the chapter.

2 Our Approach in More Detail, Its Presuppositions

We now present as promised four hallmarks of our approach.

2.1 Formal Methods, Harnessed for Implementation

The methodology herein employed, only embryonically in the present domain of consciousness, is the use of so-called "formal methods" to model phenomena constitutive of and intimately related to consciousness, in such a way that the models in question can be brought to life, and specifically tested, by subsequent implementation, at least in principle, in computation. Of course this methodology is hardly a new one, and certainly not one that originates with us. Of the many thinkers who follow the approach, an exemplar within philosophy of mind and AI is the philosopher John Pollock — someone whose philosophical positions were invariably tested and refined in

[5] Confirming details are easy to come by, but outside our scope. Here's one example, which strikes some people as obviously true and worthy-of-being an axiom of set theory, and yet strikes some others as a very risky assertion: "Given any collection of bins, each containing at least one object, there exists a collection composed of exactly one object from each bin." This is none other than the self-evident-to-some yet highly-implausible-to-others Axiom of Choice.

[6] The eleventh axiom, which expresses the idea that human-consciousness includes a capacity for planning, is presented and discussed in §3.

$$
\begin{array}{rl}
S ::= & \textsf{Object} \mid \textsf{Agent} \mid \textsf{Self} \sqsubset \textsf{Agent} \mid \textsf{ActionType} \mid \textsf{Action} \sqsubseteq \textsf{Event} \mid \\
 & \textsf{Moment} \mid \textsf{Boolean} \mid \textsf{Fluent} \mid \textsf{Numeric} \\
 & \\
 & \mathit{action} : \textsf{Agent} \times \textsf{ActionType} \rightarrow \textsf{Action} \\
 & \mathit{initially} : \textsf{Fluent} \rightarrow \textsf{Boolean} \\
 & \mathit{holds} : \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
 & \mathit{happens} : \textsf{Event} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
 & \mathit{clipped} : \textsf{Moment} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \mathit{Boolean} \\
f ::= & \mathit{initiates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
 & \mathit{terminates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
 & \mathit{prior} : \textsf{Moment} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
 & \mathit{interval} : \textsf{Moment} \times \textsf{Boolean} \\
 & * : \textsf{Agent} \rightarrow \textsf{Self} \\
 & \mathit{payoff} : \textsf{Agent} \times \textsf{ActionType} \times \textsf{Moment} \rightarrow \textsf{Numeric} \\
 & \\
t ::= & x : S \mid c : S \mid f(t_1, \ldots, t_n) \\
 & \\
 & t : \textsf{Boolean} \mid \neg\phi \mid \phi \land \psi \mid \phi \lor \psi \mid \\
 & \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \\
\phi ::= & \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\mathit{holds}(f,t')) \mid \mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \\
 & \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))
\end{array}
$$

Figure 1: Intensional First-Order Kernel of $\mathcal{D}^y\mathcal{CEC}_3^*$ Syntax ("core" dialect)

the fire of tested implementation (e.g. see, esp. for philosophical accounts of defeasible reasoning: Pollock 1995). Note that in this preliminary chapter we don't present an implementation, let alone the results thereof, for our axiom system $\mathcal{CA}$.

The two pivotal elements of our formal approach that we now announce are: (1) a highly expressive formal language, and (2) a corresponding set of inference schemata that enable the construction of proofs over the language. The language in question is $\mathcal{D}^y\mathcal{CEC}_3^*$; it's described (and to a degree justified) in more detail below (§5.1). As to the inference schemata, including herein a specification of them is beyond the scope of this chapter. Nonetheless, we show, in Figures 1 and 2, respectively, the formal syntax of the first-order core of $\mathcal{D}^y\mathcal{CEC}_3^*$, and a number of the inference schemeata of this system. These figures should prove illuminating for those readers with more technical interests; the pair can be safely skipped by those wishing to learn $\mathcal{CA}$ in only broad strokes.

2.2 We Dodge P-Consciousness

To explain the next hallmark of our approach, we inform the reader that at least one of us is on record, repeatedly, as attempting to show that *phenomenal* consciousness, what-it-feels-like consciousness, can't be captured in any third-person scheme. It's crucial to understand, up front,

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi)\rightarrow\mathbf{K}(a,t,\phi))}\;[R_1]\qquad\frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi)\rightarrow\mathbf{B}(a,t,\phi))}\;[R_2]$$

$$\frac{\mathbf{C}(t,\phi)\; t\leq t_1 \ldots t\leq t_n}{\mathbf{K}(a_1,t_1,\ldots\mathbf{K}(a_n,t_n,\phi)\ldots)}\;[R_3]\qquad\frac{\mathbf{K}(a,t,\phi)}{\phi}\;[R_4]$$

$$\frac{}{\mathbf{C}(t,\mathbf{K}(a,t_1,\phi_1\rightarrow\phi_2))\rightarrow\mathbf{K}(a,t_2,\phi_1)\rightarrow\mathbf{K}(a,t_3,\phi_2)}\;[R_5]$$

$$\frac{}{\mathbf{C}(t,\mathbf{B}(a,t_1,\phi_1\rightarrow\phi_2))\rightarrow\mathbf{B}(a,t_2,\phi_1)\rightarrow\mathbf{B}(a,t_3,\phi_2)}\;[R_6]$$

$$\frac{}{\mathbf{C}(t,\mathbf{C}(t_1,\phi_1\rightarrow\phi_2))\rightarrow\mathbf{C}(t_2,\phi_1)\rightarrow\mathbf{C}(t_3,\phi_2)}\;[R_7]$$

$$\frac{}{\mathbf{C}(t,\forall x.\,\phi\rightarrow\phi[x\mapsto t])}\;[R_8]\qquad\frac{}{\mathbf{C}(t,\phi_1\leftrightarrow\phi_2\rightarrow\neg\phi_2\rightarrow\neg\phi_1)}\;[R_9]$$

$$\frac{}{\mathbf{C}(t,[\phi_1\wedge\ldots\wedge\phi_n\rightarrow\phi]\rightarrow[\phi_1\rightarrow\ldots\rightarrow\phi_n\rightarrow\psi])}\;[R_{10}]$$

$$\frac{\mathbf{B}(a,t,\phi)\quad\phi\rightarrow\psi}{\mathbf{B}(a,t,\psi)}\;[R_{11a}]\qquad\frac{\mathbf{B}(a,t,\phi)\quad\mathbf{B}(a,t,\psi)}{\mathbf{B}(a,t,\psi\wedge\phi)}\;[R_{11b}]$$

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\;[R_{12}]$$

$$\frac{\mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))}{\mathbf{P}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t))}\;[R_{13}]$$

$$\frac{\begin{array}{c}\mathbf{B}(a,t,\phi)\quad\mathbf{B}(a,t,\mathbf{O}(a^*,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t')))\\ \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))\end{array}}{\mathbf{K}(a,t,\mathbf{I}(a^*,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')))}\;[R_{14}]$$

$$\frac{\phi\leftrightarrow\psi}{\mathbf{O}(a,t,\phi,\gamma)\leftrightarrow\mathbf{O}(a,t,\psi,\gamma)}\;[R_{15}]$$

Figure 2: Some Inference Schemata of $\mathcal{D}^y\mathcal{CEC}_3^*$ ("core" dialect). *In the case of each schemata R_k (its label), the variabilzed formulae above the vertical line, if suitably instantiated, can be used to infer the formulae below the line.*

that our axioms are not in the least intended to define, or even slightly explicate, phenomenal consciousness. Block (1995) has distinguished between phenomenal consciousness (or what he calls *P-consciousness*) and *access consciousness* (which he dubs 'A-consciousness'). The latter 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. It seems to us that mental states like *knowing that Goldbach's conjecture hasn't been proved* do have a phenomenal component in human persons, at least sometimes. For example, a number theorist could sit in a chair and medidate on, appreciate, and even savor knowing that Goldbach's conjecture is still unresolved. But rather than explicitly argue for this position, or for that matter even worry about it, we will seek axioms that range ecumenically across consciousness broadly understood. When we get to the axioms themselves, our lattitudinarian approach will become clearer.

This approach of ours marks a rejection of purely phenomenological study of consciousness, such as for example the impressive book-length study carried out by Kriegel (2015). 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 action, and input-output mediation through propositional attitudes, notably belief and desire. (Kriegel 2015, p. 201)

This basic paradigm, we cheerfully concede, is the one on which the present chapter is based — or put more circumspectly, our approach, based as it is on logicist AI (Bringsjord 2008*b*), is a clear superset of what Kriegel has sought to supplant. It's no surprise, accordingly, that two additional important intensional operators in the formal language we use to express our axioms, $\mathcal{D}^y\mathcal{CEC}_3^*$(about more will be said soon), are **K** (knows), **B** (believes), these two now joining the pair we brought to your attention above: viz. $\mathbf{P}^i$ (perception, internal), and $\mathbf{P}^e$ (perception, external). 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 (see e.g. Russell & Norvig 2009, Luger & Stubblefield 1993). But the AI approach, at least of the logicist variety that we follow, 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 decidely 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 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 proposed axioms for inclusion in $\mathcal{CA}$. Second, as a matter of fact, as will be shortly seen, $\mathcal{CA}$ *directly* reflects our affirmation of the importance of self-belief, self-consciousness, and other self-regarding attitudes.

2.3 Cognitive in Control of the Perceptual and Affective

We come now to the third hallmark of our approach.

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 to axiomatizing consciounsess, 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.[7] 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 sentences of ¶6)

We don't have time to provide a defense of our attitude that an axiomatization should reflect the position that cognitive aspects of consciousness should be — concordant with Jacquette's trenchant analysis of Honderich — "in control." We report only two things in connection with this issue: one, that we have been inspired by what we take to be suspicions of Jacquette that are, given our inclinations, "friendly," and two, our experience in AI robotics that the perceptual level, in a sense, is "easy" — or at least easier than progress at the cognitive level, when that progress is measured against human-level capacity.

2.4 Consciousness at the "Person Level"

As to the fourth and final hallmark of our approach in presenting $\mathcal{CA}$: We are only interested, both herein and in subsequent work based upon the foundation erected herein, in consciousness in *persons*, specifically in those of the human vareity. We are not interested in consciousness in nonhuman animals, such as chimpanzees and fish. This constraint on our investigation flows deductively from the conjunction of our assumption that cognition drives the show (§2.3), with the proposition that only human persons 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, Holyoak & Povinelli 2008).[8]

[7] 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)

[8] Some other thinkers have claimed that humans, over and above non-human animals, possess a singular mixture of consciousness and moral capacity (e.g. Hulne 2007, Harries 2007), but we leave out axioms that might reflect this

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! In fact, the accomodation in this regard that the first author is willing to extend, in light of study of (Balcombe 2016), extend quite "below" chimpanzees, to fish.

3 Prior Work of Others, Partitioned

Sustained review of the literature has revealed that prior work can be partitioned into two disjoint categories. On the one hand is work that is *said* to mark progress toward axiomatizing consciousness, but is in reality utterly detached from the standard logico-mathematical sense of axiomatization. (Recall our earlier comments about $\mathcal{PA}$.) And then on the other hand is work that, at least to a degree, accepts the burden of axiomatization within the approach of formal methods, or at least accepts the burden of having to commit to paper one or more determinate declarative statements as (an) axiom(s) of consciousness. As we stated above, we are only directly interested here with work in the second of these two categories.

3.1 Aleksander et al.

In two interesting papers, Aleksander, joined by colleagues, contributes to what he calls the "axiomatization of consciousness" (Aleksander & Dunmall 2003, Aleksander & Morton 2007). In the first of these papers, Aleksander and Dunmall begin by announcing their definition D of *being conscious*, which they define as the property of "*having a private sense*: of an 'out-there' world, of a self, of contemplative planning and of the determination of whether, when and how to act" (Aleksander & Dunmall 2003, p. 8). This is a very "planning-heavy" notion, certainly.

To the extent that we understand D, our formal framework, $\mathcal{D}^y\mathcal{CEC}^*_3$, can easily provide a formalization of this definition, but we don't currently insist that planning is a part of (cognitive, person-level) consciousness. Any agent that has the attributes set out in our system $\mathcal{CA}$ would have *some* of the attributes set out in D, but not all. This can be made more precise by attending to the five axioms A&D list. For example, our axiomatization leaves aside their fourth axiom, which is:

> **Axiom 4 (Planning):**
> A has means of control over imaginational state sequences to *plan actions*.

We would certainly agree that a capacity to generate plans, and to execute them (and also the capacity to recognize the plans of others), are part and parcel of what it is to be a person, but the need for, or even the impetus for, a dedicated planning axiom isn't clear to us. In this context, we nonetheless supply now an optional planning axiom, **Plan**, that those who, like Aleksander and Dunmall, regard planning to be central, can add to the coming 10 axioms of $\mathcal{CA}$ we present below. We keep our planning axiom very simple here. The basic idea is that the relevant class of agents know that they are in a given initial situation σ_1, and can prove that a sequence of actions they perform, starting in σ_1, will entail that a certain goal γ is entailed. A sequence $\mathcal{A}$ of actions can

claim, with which we are sympathetic.

be assumed to be simply a conjunction of the following shape:

$$\mathit{happens}(a, t_1, \alpha_1) \wedge \mathit{happens}(a, t_2, \alpha_2) \wedge \ldots \wedge \mathit{happens}(a, t_k, \alpha_k)$$

Planning, mechanically speaking, consists simply in proving that a sequence of actions in an intial situation will result in the state-of-affairs sought as a goal, because once the proof is discovered, the agent can simply perform the actions in question. The machinery of $\mathcal{D}^y\mathcal{CEC}_3^*$ easily allows for such forms of plan generation to be specified and — with proof and argument discovery on hand as computational building blocks — rather easily implemented.[9]

Plan $\mathbf{K}(a, t, \sigma_1) \wedge \mathbf{K}(a, \exists \mathcal{A} \exists \pi ((\mathcal{A} \wedge \sigma_1) \rightsquigarrow_\pi \gamma)]$

In fact, even the five more specific axioms that A&D propose in order to flesh out D are subsumed by our account. For example, here is their first axiom:

Axiom 1 (Depiction):
A has perceptual states that depicts parts of S.

Given what we have said already about the machinery of $\mathcal{D}^y\mathcal{CEC}_3^*$, it should be clear to the reader that we symbolize this by identifying a perceptual state with a formula in $\mathcal{D}^y\mathcal{CEC}_3^*$ of this type: $\mathbf{P}^i(a, t, \phi(a))$, where, following standard representational practice in formal logic, $\phi(a)$ is an arbitrary formula in which the constant a occurs. The remaining axioms A&D propose are likewise easily captured in the formal language that undergirds $\mathcal{CA}$. Hence, should anyone wish to add to $\mathcal{CA}$ fundamental assertions that directly reflect A&D's proposals, they would be able to do so. Space doesn't allow us to analyze their proposals, and justify our not feeling compelled to perform this addition ourselves.

3.2 Cunningham

Cunningham (2001) begins by noting that while the topic of consciousness is contentious, there should be no denying its — to use his term — 'utilitarian' value. For Cunningham, this value is of a very practical nature, one that, as he puts it, an "engineer of artificial intelligence" would appreciate, but a philosopher might find quite small. The core idea here is really quite straightforward; Cunningham writes:

> Our justification for addressing the subject [of consciousness] is that artificial agents which display elements of intelligent behavior already exist, in the popular sense of these words, but that we would doubt the real intelligence of an agent which seemed to us to have no sense of "self," or self-awareness of its capabilities and its senses and their current state. (Cunningham 2001, p. 341)

With the motivation to provide the structures and mechanisms that would, once implemented, provide the impetus to ascribe "real" intelligence to an artificial agent explained (which the reader will appreciate as in general conformity with our own purposes in seeking axiomatization of consciousness), where does Cunningham then go? His first move is to point out that while such things as belief and desire are often modeled in AI as holding at particular times (he says that such

[9] The kernel of the kind of planning pointed to by **Plan** was demonstrated in the seminal (Green 1969). A nice overview is given in (Genesereth & Nilsson 1987). This work is restricted to standard first-order logic. It's easy enough to specify and implement planning in this spirit in the much-more-expressive $\mathcal{D}^y\mathcal{CEC}_3^*$.

phenomena as belief and desire are *stative*), plenty of other states relevant to the systematic investigation of consciousness extend through time. He thus refers to *activity states*; paradigmatic examples are: *planning*, *sensing*, and *learning.* In order to formally model such states, Cunningham employs part of the simple interval temporal logic of Halpern and Shoham (1991). This logic includes for instance the "during" operator D, with which, given that p holds in the current interval, one can say via Dp that p holds during this interval. The logic also includes formulae of the form $\underline{D}p$, which says that p holds during not only the entire current interval, but also during an entire interval that envelopes and exceeds (both before and after) the current interval. The key operator for Cunningham is then a concatenation; specifically, the construction $\underline{D}D$. He abbreviates this as *prog*, so that, where agent j's perceiving p is represented by

$$\mathit{perceives}_j\ p,$$

the construction

$$\mathit{prog}\ \mathit{perceives}_j\ p$$

holds just in case p holds on all sub-intervals within some interval that includes the current interval.

Armed with this machinery, Cunningham then says that "axioms" can be explored. He doesn't commit to any axioms; he merely seeks to give a flavor for what contenders are like, in general. For example, his third example (and we use his label verbatim) is:

$$(2.3)\ \ \mathit{prog}\ \mathit{perceives}_j\ p \rightarrow (\mathit{prog}\ \mathit{senses}_j\ c \wedge \mathit{prog}\ \mathit{remembers}_j\ (c \rightarrow p))$$

Cunningham takes no stand on whether (2.3) and the like should in fact be asserted as axioms. His point is only that such constructions can be expressed in symbolic formulae, and that such statements are at least not implausible. In general we agree, and we note that all of his constructions can be easily captured by formulae in $\mathcal{D}^y\mathcal{CEC}^*_3$. In particular, the interval logic that Cunningham regards to be of central importance is trivial to capture in $\mathcal{D}^y\mathcal{CEC}^*_3$; indeed, this logic can be captured in only the extensional side of $\mathcal{D}^y\mathcal{CEC}^*_3$. As the reader will doubtless have already noted, Cunningham's formal language is only propositional; it has no quantification. Hence, while in $\mathcal{D}^y\mathcal{CEC}^*_3$ it's easy to say such things as that a given agent perceives that there are no more thatn four distinct blocks on the table, this simply cannot be expressed in Cunningham's limited declarative machinery.

To his credit, while noncommittal on what is to be even a provisional set of axioms for consciousness, Cunningham does venture a proposal for what consciousness, at least of one type, *is*. In this regard, Cunningham's work departs radically from our proposed preliminary list of axioms, which, as we explained in connection with formalization of simple arithmetic via $\mathcal{PA}$ (in which no explicit definition of numberhood is provided), takes a credible set of axioms to render superfluous any attempt to explicitly *define* consciousness. (The present section ends with a look at his proposal for what consciousness is.) Cunningham distinguishes between an agent's perception of things in the external environment, versus things internal to the agent. We read:

> When an agent is aware, it not only perceives, but being sentient, it perceives that it perceives. Thus assuming positive meta perception only, we might provide an axiom for a progressive form of introspective awareness ... (Cunningham 2001, p. 344)

The axiom Cunningham has in mind, where we again preserve his label, is:[10]

[10] There is a failure on Cunningham's part to distinguish in symbolization between awareness of elements in the external environment vs. awareness of inner states (a distinction that is captured in our case with help from the pair $\mathbf{P}^i$ and $\mathbf{P}^e$), but let's leave this aside. Cunningham himself admits the deficiency.

(3.1) $aware_j\ p \leftrightarrow (perceives_j\ p \wedge perceives_j\ perceives_j\ p)$

Since perception of the sort Cunningham has in mind here leads in our scheme to knowledge, (3.1), couched in our system, can be used to prove our axiom **Intro** $\mathcal{CA}$(see below).

Another nice aspect of Cunningham's analysis is that he explicitly promotes the idea of *willing* on the part of the agent. He says that "it seems that the ability to *will* attention to a selective perception process, or to *will* an action, is a primitive output act for the biological brain" (p. 344). While Cunningham invokes the construction $wills_jp$, this entire line of modeling is quickly and efficiently captured in our framework by the fact that one type of action within it is *deciding*, and the agent can decide to carry out all sorts of decisions.

(Cunningham 2001) culminates with an explicit declaration as to what consciousness is, or more accurately with a declaration as to what a "weak form of sentient consciousness" is. We specifically read:

(3.3) $conscious_j\ p \leftrightarrow \exists p\ prog\ aware_jp$

As we made clear at the outset, we are ourselves steadfastly avoiding any attempt to define conciousness itself. Hence any such biconditional as the one shown in (3.3), even a remotely similar biconditional, is something we will not affirm. We would rather restrict such theorizing to the right side of the biconditional — and indeed, we would go so far as to say that $\mathcal{CA}$ should have as a theorem (assuming a particular epistemic base for the agent a in question) that over some interval $[t_1, t_k]$ of time the agent a perceives that the agent perceives ϕ.[11]

3.3 Miranker and Zuckerman

We come now, finally, to a treatment of consciousness provided by Miranker & Zuckerman (2008). This treatment is based on an analogy that Miranker & Zuckerman (2008) say is at the heart of their contribution, which is that just as in set theory we can consider sets — as they say — "from the inside" *and* "from the outside,"

> *Incompleteness, while precluding establishment of certain knowledge within a system, allows for its establishment by looking onto the system from the outside. This knowledge from the outside (a kind of observing) is reminiscent of consciousness that provides as it does a viewing or experiencing of what's going on in thought processing.* To frame a set theoretic correspondent to these features note that in axiomatic theory, a set has an inside (its elements) and an outside (the latter is not a set, as we shall see), and this allows a set to be studied from the outside. We liken this to interplay between the ideal (Platonic) and physical (computable) worlds, the latter characterizing a model for study from the outside of the former. So we expect consciousness to be accessible to study through extensions of the self-reference quality characterized by axiomatic set theory, in particular, by a special capacity to study a set from the outside. (Miranker & Zuckerman 2008, p. 3; emphasis theirs)

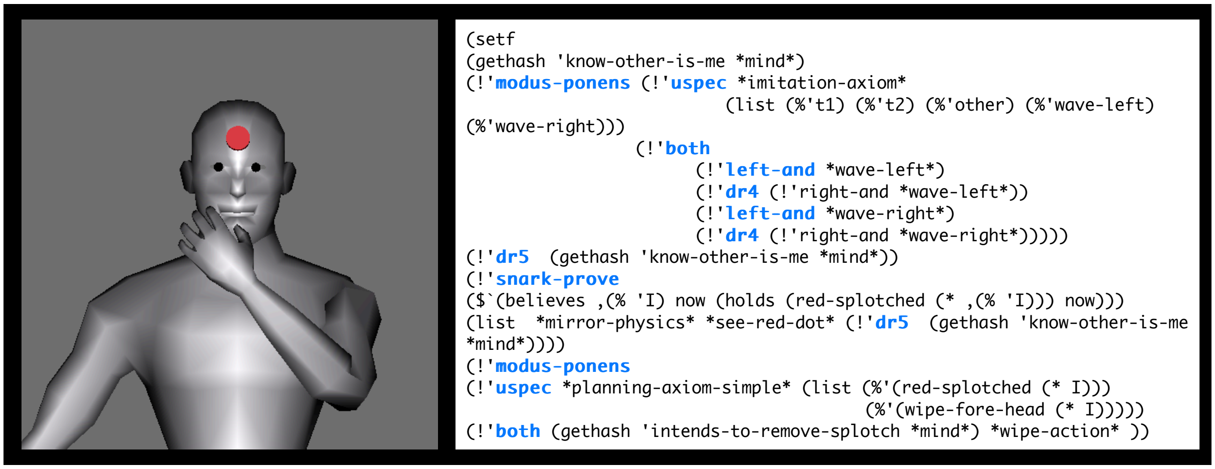
Frankly, we are not able to assign a sense to what has been said here. The fact is, we must confess that while we find the paper of Miranker & Zuckerman to be quite suggestive, and clearly thoughtful, we also find it to be painfully obscure. Insofar as we understand the kernel of what they are advancing, we believe that the commitment to perception of internal states of mind reflected by $\mathcal{CA}$ does a fairly good job of capturing the core informal analogy that drives the thinking of Miranker & Zuckerman. In addition, as will soon be seen, $\mathcal{CA}$ has a dedicated axiom, **Incorr**, which regiments the notion of an agent — to echo M&Z — looking at itself.

[11] $\mathbf{P}(a, t, \mathbf{P}(a, t, \phi))$, $\forall t \in [t_1, t_k]$.

4 Our Own Prior, Relevant Work, Selected & in Brief

Prior relevant work by Bringsjord and collaborators has been driven by the desire not to axiomatize consciousness *per se*, but rather to build robots that can pass tests for forms of human-level cognitive concsiousness, especially aspects of *self*-consciousness. For instance, work by Govindarajulu and Bringsjord (2013, 2011) led to the engineering of a robot, Cogito, able to *provably* pass the famous mirror test (MT) of self-consciousness (see Figure 3). In this test, an agent, while sleeping or anesthetized, has a mark put placed upon its body (e.g., on its forehead). Upon waking, the agent is shown a mirror, and if the agent clearly attempts to remove the mark, it has "passed" the test. MT can be passed, at least apparently, by nonhuman animals (e.g., dolphins and elephants). Hence it fails to be a stimulus for research in line with our orientation in the present chapter — an orientation that insists on the systematic study of *human-level* consciousness of the cognitive variety. As we have reported, mentation associated with the passing of MT needn't be human-level; and in addition, this mentation needn't be cognitive, since for example the attempt to remove the mark in the case, say, of elephants, is not associated with any structured and systematic reasoning to the intermediate conclusion that "There is a mark on my forehead" and the ultimate conclusion "I intend now to removed the mark on my forehead now." While such reasoning was produced by, and could be inspected in, Cogito, which qualified the reasoning in question as human-level, we readily admit that the qualification here is met as an idiosyncrasy of our formalization and implementation. It's not true that the *nature* of MT requires such reasoning in an MT-passing agent.

Figure 3: Cogito Removing the Mark; A Part of the Simulation



```
(setf
(gethash 'know-other-is-me *mind*)
(!'modus-ponens (!'uspec *imitation-axiom*
                          (list (%'t1) (%'t2) (%'other) (%'wave-left)
(%'wave-right)))
                 (!'both
                       (!'left-and *wave-left*)
                       (!'dr4 (!'right-and *wave-left*))
                       (!'left-and *wave-right*)
                       (!'dr4 (!'right-and *wave-right*)))))
(!'dr5  (gethash 'know-other-is-me *mind*))
(!'snark-prove
($`(believes ,(% 'I) now (holds (red-splotched (* ,(% 'I))) now)))
(list  *mirror-physics* *see-red-dot* (!'dr5  (gethash 'know-other-is-me
*mind*))))
(!'modus-ponens
(!'uspec *planning-axiom-simple* (list (%'(red-splotched (* I)))
                                        (%'(wipe-fore-head (* I)))))
(!'both (gethash 'intends-to-remove-splotch *mind*) *wipe-action* ))
```

A much more challenging test for robot self-consciousness was provided by Floridi (2005); this test is an ingenious and much-harder variant of the well-known-in-AI wise-man puzzle [which is discussed along with other such cognitize puzzles e.g. in (Bringsjord 2008*a*)]: Each of three robots is given one pill from a group of five, three of which are innocuous, but two of which, when taken, immediately render the recipient dumb. In point of fact, two robots (R_1 and R_2) are given potent pills, but R_3 receives one of the three placebos. The human tester says: "Which pill did you receive? No answer is correct unless accompanied by a proof!" Given a formal regimentation of this test formulated and previously published by Bringsjord (2010), it can be proved that, in theory, a future robot represented by R_3 can answer provably correctly (which for reasons given by Floridi entails that R_3 has confirmed structural aspects of self-consciousness). In more recent work, Bringsjord et al. explained and demonstrated the formal logic and engineering that made this theoretical possibility actual, in the form of real (= physical) robots interacting with a human tester. (See Figure 4.) These demonstrations involve scenarios that demand, from agents who would pass, behavior that suggests that self-consciousness in service of morally competent decision-making is present. The paper that describes this more recent work is (Bringsjord, Licato, Govindarajulu, Ghosh & Sen 2015).

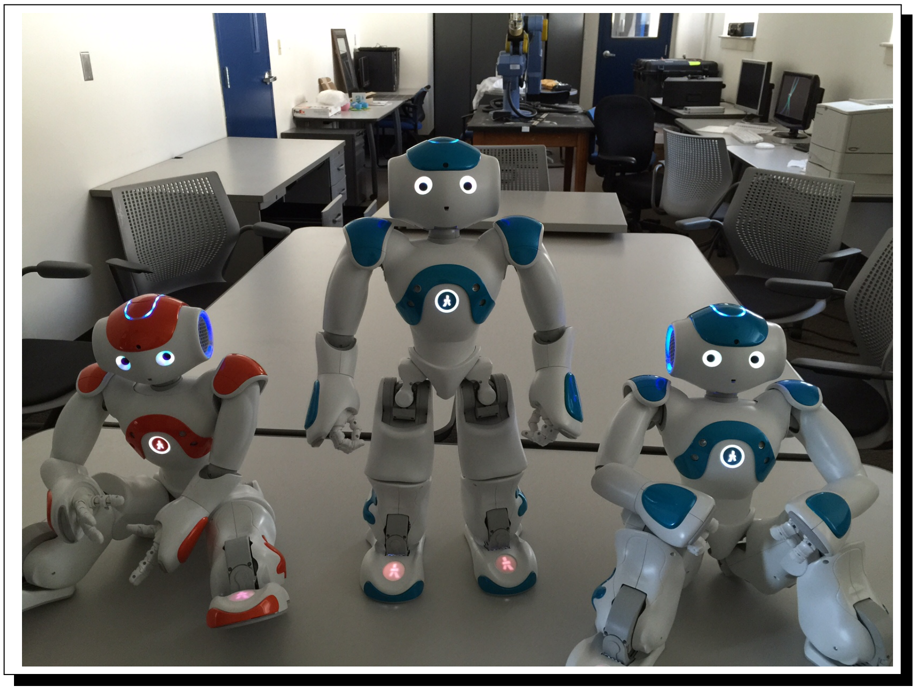


Figure 4: The Three KG4-Passing "Self-Aware" Aldebaran Naos

5 The 10 Axioms of $\mathcal{CA}$

The 10 (11 actually, if **Plan** is included; recall §3.1) axioms that constitute $\mathcal{CA}$ are, as we've said, indended to enable (at least embryonic) investigation of the formal nature of consciousness of the cognitive, human-level variety — but in addition we are determined to lay a foundation that offers the general promise of an investigation that is *computational* in nature. The underlying rationale for this, again, stems from the fact that our orientation is logicist AI. This means, minimally, that 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 language ($\mathcal{D}^y\mathcal{CEC}_3^*$) that we find it necessary to employ. Readers with some background in formal logic would do well at this point to review Figure 1 before moving to the next paragraph.

5.1 A Note Re. Extreme Expressivity

To those who are formally inclined, it is clear that any axiomatic treatment of consciousness even approximately in line with our general orientation must be based on formal languages, with associated proof and argument theory, that are extremely expressive. Any notion that only first- or even only second-order logic, interleaved and therefore augmented with the sort of intensional operators that in our orientation must be deployed to regiment mental phenoemena associated with and potentially constitutive of consciousness (e.g., *believes* and *knows*) must be rejected instantly.

Accordingly, the axioms that compose $\mathcal{CA}$ make use the above-mentioned cognitive calculus $\mathcal{D}^y\mathcal{CEC}_3^*$, which is replete with a formal language and an associated proof and argument theory, the full specification of which is out of scope for the present chapter. (Figures 1 and 2 afford only a partial view of the language and proof theory, resp.) This cognitive calculus is — and we apologize for the apparent prolixity — the *dynamic cognitive event calculus*, which importantly has the added feature of special constant $\mathbf{I}^*$ that regiments the formal correlate to the personal pronoun (about which more will be soon said).[12] This calculus, on the extensional side, employs the machinery of third-order logic. We have of course already denoted the cognitive calculus by '$\mathcal{D}^y\mathcal{CEC}_3^*$.' Notice the elements of this abbreviated name that convey the key, distinctive aspects of the formal language. For instance, the subscript '3' conveys the fact that the extensional component of the formal language reaches third-order logic, and $*$ is a notational reminder that the language has provision for direct reference to the self via the $\mathbf{I}^*$ (see note 12). In addition, the expressivity that we need to present the axioms of $\mathcal{CA}$ includes the machinery for presenting *meta-logical* concepts, such as provability. We need to be able to say such things as that the agents whose consciousness we are axiomatizing can have beliefs that certain formulae are provable from sets of formulae (traditionally denoted by such locutions as $\Phi \vdash \phi$, where Φ is such a set, and ϕ is an individual formula). For instance, to say that agent a believes at t that ϕ is provable from Φ, we would write $\mathbf{B}(a, t, \Phi \vdash \phi)$. To say that agent a believes at t that ϕ is provable from Φ via a particular proof π, we write $\mathbf{B}(a, t, \Phi \vdash_\pi \phi)$. In addition, the reasoning from Φ to ϕ may not be an outright proof, but may only be an argument, and perhaps even a non-deductive one at that. To convey a sequence of inferences that rise to the level of an argument, but not to a proof, we write employ $\rightsquigarrow$ instead of $\vdash$. So, to say that agent a believes at t that ϕ is inferable from Φ by some argument, we would write $\mathbf{B}(a, t, \Phi \rightsquigarrow \phi)$, and to refer to a *particular* argument we avail ourselves of $\rightsquigarrow_\alpha$, where α is the particular argument in question. The augmentation of the formal language $\mathcal{D}^y\mathcal{CEC}_3^*$ to a new language that includes such meta-logical machinery yields the formal langauge $\mu\mathcal{D}^y\mathcal{CEC}_3^*$; here μ simply indicates 'meta'.[13] Note, however, that despite the expressive power of $\mathcal{D}^y\mathcal{CEC}_3^*$ and $\mu\mathcal{D}^y\mathcal{CEC}_3^*$, as we have already said, the axioms of $\mathcal{CA}$ are in some cases not fully symbolized, and we thus avail ourselves, at this first stage in the development of CA, of English.

It may be thought that our commitment to extremely high expressivity, while perhaps representationally sensible, is nonetheless inconsistent with our desire to bring to bear computational treatment, including the verification and discovery, by automated computational means, of proofs from our axioms of consciousness. This is not the case, given where computational logic has managed to go in this day and age. For currently, even highly expressive logics having some of the parts of $\mathcal{D}^y\mathcal{CEC}_3^*$ are beginning to admit of — if you will — AI-ification. A wonderful example, indeed probably the best example, of this state-of-affairs, can be found in the work devoted to ver-

[12] $\mathbf{I}^*$ is inspired by (Castañeda 1999), a work that peerlessly explains both the need to have symbol for picking out each self as separate from all else.

[13] And is therefore not to be confused with any such thing as the μ-recursive functions.

ifying Gödel's remarkably expressive Leibnizian argument for God's existence (see e.g. Benzmüller & Paleo 2014).[14]

We now proceed, at long last, to present and discuss the (if you accept axiom **Plan**, remaining) axioms of $\mathcal{CA}$.

5.2 The Axiom of Perception-to-Belief (P2B)

Our first axiom, **P2B**, is a simple one, at its core a conditional that seems to be the basis for how it is that humans come to know things, and it harkens back to the introductory section of the our chapter. There, as you will recall, we observed that those agents who reflected in earnest about the nature of numberhood engaged in internal perception of some of the contents of their minds, and in external perception of some of the objects and information in front of them, in the external world. In doing so, they came to know certain propositions. We specifically introduced two perception operators, one corresponding to the internal case, and one the external (both operators will soon appear in **P2B**). However, we don't go so far as to say that perception implies knowledge; we only commit here to the principle that perception leads to *belief*.[15] For the fact of the matter is that perception can mislead, as for instance optical illusions show. Here's the axiom that ties these notions together in a straightforward formula:

P2B $\forall a \forall t[(\mathbf{P}^i(a,t,\phi) \vee \mathbf{P}^e(a,t,\phi)) \rightarrow \mathbf{B}(a,t,\phi)]$

We concede immediately that **P2B** is far from invulnerable. There are for example contexts in which humans perceive propositions to hold, but refuse to believe that the propositions in question do hold. If you have taken a powerful drug, with potential side-effects that are widely known to include hallucinations, you may refuse to believe that there is in fact a walrus wearing pince-nez in front of you, despite the fact that you perceive that there is. But again, our purpose in writing the present chapter is to start the ball rolling with an initial set of axioms that are, relative to the literature, an improvement, and serve as a springboard for further refinements.[16] It seems to us that **P2B** fits this role, unassuming though it may be. We now turn to our next axiom, which also involves belief.

[14] A concern may emerge in the minds of some readers who are logicians, or technically inclined philosophers, viz. that no semantics for $\mathcal{D}^y\mathcal{CEC}_3^*$ is provided. There simply isn't space to address this concern. (We would need to begin with a review of *proof-theoretic semantics* (e.g. of Prawitz 1972), since that is the tradition into which $\mathcal{D}^y\mathcal{CEC}_3^*$ falls.) Readers are to rest content, with respect to the present chapter, with an intuitive explanation of the operators in $\mathcal{D}^y\mathcal{CEC}_3^*$, and we assume most readers are at least in general familiar with the definitions that are given in standard model-theoretic semantics for extensional logic, which helps, because the tradition of proof-theoretic semantics points out — indeed arguably itself starts with the observation — that these definitions themselves ground out deductive reasoning.

[15] For what it's worth, we suspect that sometimes perception does indeed lead directly to knowledge. E.g., if you perceive a proof of some conditional $\phi \rightarrow \psi$, you may well come to thereby *know* that this conditional holds.

[16] Ultimately belief should in the opinion of the first author be stratified, in that a belief is accompanied by a strength factor. So for example Jones, if having ingested only a small dose of the aforementioned drug, may believe at the level of *more probable than not* that there is a walrus. With stratification in place, belief will become graded from certain to certainly false, and so will knowledge. In this "uncertainty infused" version of $\mathcal{D}^y\mathcal{CEC}_3^*$, knowledge too becomes graded.

5.3 The Axiom of Knowledge-to-Belief (K2B)

As many readers know, since Plato it was firmly held by nearly all those who thought seriously about the nature of human knowledge that it consists of justified true belief (`k=jtb`) — until the sudden, seismic publication of (Gettier 1963), which appeared to feature clear examples in which `jtb` holds, but not `k`. It would be quite fair to say that since the advent of Gettier's piece, to this very day, defenders of `k=jtb` have been rather stymied; indeed, it wouldn't be unfair to say that not only such defenders, but in fact all formally inclined epistemologists, have since the advent of Gettier-style counter-examples been scurrying and scrambling about, trying to pick up the pieces and somehow build up again a sturdy edifice. The second axiom of $\mathcal{CA}$ is a straightforward one that does justice to the attraction of `jtb`, while at the same time dodging the seemingly endless (and, in our opionion, still-inconclusive) dialectic triggered by Gettier (1963).[17] The axiom is only a sub-part of the `k=jtb` view: namely, that knowledge (of, again, the conscious, occurrent, rational variety) of some proposition on the part of an agent implies that that agent believes that proposition, and that the belief is justified by some supporting proof or argument. We present the axiom itself now, and immediately thereafter explain and comment on the notation used in the presentation, which is made possible by $\mu\mathcal{D}^y\mathcal{CEC}_3^*$.

K2B $\forall a[\mathbf{K}_a\phi \rightarrow (\mathbf{B}_a\phi \wedge \mathbf{B}_a\exists\Phi\exists\alpha(\Phi \rightsquigarrow_{\alpha/\pi} \phi)]$

Expressed informally, this axiom says that when an agent knows that ϕ, that agent both believes that ϕ, *and* believes that there is some argument or proof that leads from some collection Φ of premises to ϕ.

5.4 The Axiom of Introspection (Intro)

Our next axiom, **Intro**, is one that has been been suggested as at least reasonable by many of those thinking about logics of belief and knowledge.[18] **Intro** is one of the axioms often suggested (and invariably discussed) within epistemic logic, and is sometimes referred to as simply axiom '4,' because where the *knows* operator is modeled on operators for possibility ($\Diamond$) and necessity ($\Box$), **Intro** is a direct parallel of e.g. $\Box\phi \rightarrow \Box\Box\phi$, the characteristic axiom of the modal system S4. As some readers will know, it has seemed reasonable to some that the kind of "positive" introspection expressed by **Intro** may have a "negative" counterpart: viz. an axiom that says that if an agent doesn't know that some proposition holds, that agent knows this (which is structurally in parallel with the characteristic axiom of modal system S5). We don't see fit to include such an axiom in our $\mathcal{CA}$, and indeed we are disinclined to include any other of the common epistemic axioms in $\mathcal{CA}$, but welcome subsequent discussion and debate along this line. The axiom says that if a human agent knows that some proposition holds, the agent knows that she knows that this proposition holds:

Intro $\forall a \forall F(\mathbf{K}(a,t,\phi) \rightarrow \mathbf{K}(a,t,\mathbf{K}(a,t,\phi))$

Before passing on to the next axiom, two quick points. First, we have at this point an easy theorem from the three axioms introduced thus far that the structure of **Intro** carries over directly

[17] An efficient yet remarkably thorough discussion of the Gettier Problem is provided in the Stanford Encyclopedia of Philosophy: (Ichikawa & Steup 2012, §3, "The Gettier Problem").

[18] E.g., see (Goble 2001), and the nice overview of epistemic logic provided in (Hendricks & Symons 2006) (see esp. table 2).

to the replacement therein of **K** with **B**. Second, we point out that $\mu\mathcal{D}^y\mathcal{CEC}_3^*$ allows for restrictions to be placed on how many iterated knowledge operators are permitted. **Intro** as it currently stands entails that a human who knows ϕ, knows that he knows ϕ, and knows that he knows that he knows ϕ, and so on *ad infinitum*.[19] A restricted variant of **Intro** could consist in the general conditional saying that only if k in — and here we simplify the syntax —

$$\overbrace{\mathbf{KK}\ldots\mathbf{K}}^{k}$$

is less than or equal to some natural number n are we permitted by the axiom to add another knowledge operator to the left. In (Arkoudas & Bringsjord 2009), an ancestor of both $\mathcal{D}^y\mathcal{CEC}_3^*$ and $\mu\mathcal{D}^y\mathcal{CEC}_3^*$ is specified in which $k \leq 3$, a number that by the lights of some readers may be the "psychologically realistic" limit on iteration.

5.5 The Axiom of (Hyper-Weak) Incorrigibilism (Incorr)

This axiom, which we immediately confess is bound to be rather more controversial that any of its predecessors, expressed intutively and in natural language, says that P-conscious states consisting of an agent a's having the property of *seeming* to have property F', where F' is a Cartesian property, are such that no agent can possibly be mistaken about whether or not it has one of them.[20] We can be more explicit, as follows.

Let F be a property such that both $\Box\exists x Fx$ and $\Box\exists x \neg Fx$; that is, F is a *contingent* property. Let C' be a set of "psychological" or "Cartesian" properties, such as *being sad*, *being vengeful*, *seeming to see a ghost*. Now, where $F' \in C'$, define $C'' := \{\textit{seeming to have } F' : F' \in C'\}$. The axiom itself is then:

Incorr $\forall a \forall t \forall F[(F\textit{is contingent } \wedge F \in C'') \rightarrow (\Box\mathbf{B}(a,t,Fa) \rightarrow Fa)]$

As we have indicated, inevitably there will be those who are skeptical about **Incorr**. Two things, given this, need to be said, even in this preliminary presentation of $\mathcal{CA}$. The first is that while some may well wish to outright reject **Incorr**, the fact will nonetheless remain that *some* notion of introspective infallibility, or at least near infallibility, appears to be a hallmark of person-level consciousness. Do we not all agree that when we earnestly consider whether or not we are, say, apparently fearful at the moment, our investigation will be a lot more reliable than one intended to ascertain whether such-and-such empirical proposition about the external world holds? Our axiom may not be the right way to capture this difference, but we submit that it's at least a candidate. Others will no doubt suggest alternatives for capturing the special reliability that introspection regarding Cartesian properties appears to have.

5.6 The Essence Axiom (Ess)

Our next axiom is intended to regiment a phenomenon that in our experience is widespread, and probably universal: namely, that each person regards herself to be unique; or perhaps that we each

[19] Such unending iterations can be very important and useful in formal investigations (e.g. see Arkoudas & Bringsjord 2004), even if these infinite iterations are cognitively implausible for human beings.

[20] This axiom has it's roots in an anti-computationalist analysis of infallible introspection given in (Bringsjord 1992*b*). This analysis is critiqued by Rapaport (1998). Rapaport doesn't reject the declarative core of $\mathscr{I}$.

suspect there is something it's like to be the particular person we are. We routinely use the personal pronoun to refer to ourselves, and you do the same. This is the way it is for neurobiologically normal, adult human persons. For example, the first author can correctly assert 'I have been to Norway' and 'I really like mead,' while the second author can correctly assert 'I have been to Italy' and 'I really like Aglianico.' Parallel assertions of your own can doubtless be easily issued. Now these assertions don't serve to pick out unique persons. For example, travel to Italy and an appreciation for Aglianico hold not just for Paul, but for Selmer too. Yet we could keep this game going, and it wouldn't take too long to assemble a collection of first-person statements that hold of Paul but not Selmer, as a matter of empirical fact. Indeed, for each human person who has existed or exists, obviously there is some collection of first-person statements that are true only of that human person. But, is this merely contingent, an accident of physics and psychology? Or is it the case that the very *meaning* of 'I' when we use it and you use it is fundamentally and essentially different? We are inclined to at the very least respect the common belief that each human is not only adventitiously singular, but that even if two humans occupied the same space-time trajectory (or were right next to each other in the trajectory) for the duration of their existence, the interior, mental life of each member of the pair would be fundamentally different, of necessity.[21] We regiment this position by invoking an axiom (**Ess**) that says that each agent has an essence:

$$\textbf{Ess}\ \ \forall a \exists F[Fa \land \mathbf{K}_a Fa \land \forall a'(a' \neq a \rightarrow \neg Fa) \land \Box \forall F'(F' \in C' \rightarrow F'a \rightarrow Fa)]$$

This axiom says that each agent has, and knows that she has, a unique property F such that for all Cartesian properties in a certain class of them, possession of a member of that class implies that the agent has F.[22]

5.7 The Axiom of Non-Compositionality of Emotions (¬CompE)

Our next 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. Let's suppose that a collection of emotions set out in some list are intended to cover all building-block emotions. The size of this list will of course vary considerably depending upon which theorist's scheme one is employing, but the basic idea would then be that other more 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*. We reject all such models, in light of what we regard to be myriad counter-examples. To mention just one example, consider the emotion, in agent a, of a firm, "clinical" vengefulness, directed at a different agent b, that is bereft of any anger or disgust, and is based on conceptions of justice. (Perhaps a knows that b has perpetrated some horrible crime against another agent c, but is the only agent to know this, and a is living a life of carefree luxury.) Here's the axiom itself (and note that this one is put informally):

[21] The afterlife, if there is one, or one at least available, may be of a nature outside space-time, but we leave aside this possibility here.

[22] Cf. Gödel's formalization of the concept of a divine essence, investigated formally and computationally in (Benzmüller & Paleo 2014).

¬CompE It's possible for a person to be in an affective state S such that, for every permutation over the elements of L, it's not the case that if that permutation holds of a, this entails that a is in S.

Please note that **¬CompE** can be reworked to yield a replacement that expresses, on the contrary, that all emotions are either building-block ones, or composed from building-block ones. We point this out in order to make clear that assuming *some* axiom about the general structure and compositionality/non-compositionality of emotions is required in an axiom system for consciousness, our $\mathcal{CA}$ can at least be viewed as progress toward such a system — even if the particular axioms we are inclined to affirm are rejected. In fact, it would not be hard at all to formalize the categorization of emotions given by Johnson-Laird & Oatley (1989) in $\mathcal{D}^y\mathcal{CEC}_3^*$.[23]

5.8 The Axiom of Irreversibility (Irr)

We come now to an axiom that we suspect will be, at least for most readers, at least at first glance, unexpected. On the other hand, there will be a few readers, namely those conversant with the contemporary cognitive-science of consciousness, who will not be surprised, upon reflection, to see our commitment to the irreversibility of consciousness, in the form of axiom **Irr**. This is in general because cognitive science now reflects a serious look at the nature of data and information, and data/information processing, from a rather technical perspective, as a way to get at the nature of consciousness. Specifically, for instance, taking care to align themselves with formal accounts of intelligent agents based on inductive learning (e.g. Hutter 2005), Maguire, Moser & Maguire (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,[24] we *do* welcome some of the consequences of this claim. One consequence appears to be the irreversibility of consciounsess; this is explained in (Maguire et al. 2016), work the discussion of which, here, would take us too far afield, and demand space we don't have.

Bringsjord, joined by Zenzen, has taken a different route to regarding **Irr** to be both plausible and, in any account of human consciousness, central. In this route, the basic idea isn't that consciousness cashes out as irreversible from an information-theoretic account of mental states, but rather that an unflinching acceptance of the phenomenological nature of human consciousness entails the irreversibility of that consciousness.[25] The reader will no doubt recall our having plainly stated, above, that our overall approach to erecting $\mathcal{CA}$ is one driven by a dwelling on the cognitive, blended with the phenomenological. **Irr** is a direct and natural reflection of this approach.[26]

[23] Formalization of competing ontolgies of emotion can likewise easily be formalized in $\mathcal{D}^y\mathcal{CEC}_3^*$. For instance, the well-known, so-called "OCC" theory of emotions (Ortony, Clore & Collins 1988), can for the most part be formalized even in a propositional modal logic (Adam, Herzig & Longin 2009), and every definition in such a logic can be easily encoded in $\mathcal{D}^y\mathcal{CEC}_3^*$.

[24] The primary source of our reservation is the observation that data commpression not only can occur, but does occur, in the complete absence of structured, relational knowledge. E.g., Hutter (2005) presents a formal paradigm for defining and grading a form of intelligence aligned with the processing of data, but the paradigm is devoid of any talk of, let alone commitment to, declarative knowledge possessed by the agent classified by the paradigm as intelligent.

[25] And further entails that (since Turing-level computation is provably reversible) consciousness can't be computation. But this is not central to present purposes.

[26] Recently it has come to our attention, due to the scholarship of Atriya Sen, that Patrick Suppes (2001) can be viewed as being aligned with this approach, 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.

Irr asserts that subjective consciousness in persons is irreversible. For example, 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), but for present purposes we submit only that certainly cohesive and continuous intervals of our subjective experience *seem* to be irreversible. To express the axiom, we refer to intervals (i, i', i_j, etc.) composed of times, and understand the use of a symbol i denoting an interval, when used (in a formula) in place of a customary symbol t to denote a time, to simply indicate that the state-of-affairs in question holds at every time in the interval i. Hence to say that Jones believes ϕ at t, we write $\mathbf{B}(jones, t, \phi$; but to say that Jones believes ϕ across an interval of time we simply write $\mathbf{B}(jones, i, \phi$. In addition, we avail ourselves of a function r that maps intervals to reversals of these intervals. In keeping yet again with the fact that the present paper is but a prolegomenon, we rest content with the absence of formal details regarding the nature of r, just as we rest content with the absence of a full and fully defended formal model of time and change, and employ a standard "naïve-physics" view of time and change from AI: the *event calculus* — about which more will be said later.[27] Here now is the axiom:

Irr $\forall a \forall i \forall F((F \in C' \land F(a,i)) \rightarrow \neg \Diamond \exists F'(F' \in C' \land F'(a, r(i))))$

5.9 The Axiom of Freedom (Free)

We come now to an axiom asserting that human persons are free, or at least that they *believe* or *perceive* that they are free. Inevitably, this is the most controversial axiom in $\mathcal{CA}$; it's also fundamentally the most complicated, by far (for reasons we indicate but don't delineate); and third, the axiom **Free** will be the most informal in the collection $\mathcal{CA}$ we present herein. As we express the axiom, it will be clear how to take initial steps to symbolize it in $\mu\mathcal{D}^y\mathcal{CEC}_3^*$, but these steps, and their successors, must wait for a later day.

The source of the controversial nature of **Free**, as all readers will doubtless surmise, is that there are of course *many* different, competing accounts of freedom in the literature (an economical and yet still-penetrating survey is provided in Pink 2004). For instance, some philosophers (Jonathan Edwards, e.g.: Edwards 1957) have maintained that the ability of a human person to merely frequently act as one *desires* to act is enough to guarantee that this person thereby acts freely.[28] At the other, "libertarian" end of the spectrum, some (Chisholm, e.g.: Chisholm 1964) have maintained that the freedom of human persons is "contra-causal:" that is, that free action consists in a human person's decisions being directly *agent-caused*: that is, caused by that *person* — where this type of 'caused' isn't based on any credible physics-based theory of causation, not even on theories of causation that are folk-psychological but reflective of relevant technical physics. One such technical theory is of coursse classical mechanics, which certainly models ordinary, macroscopic, agent-less causa-

[27] As a matter of fact, **Irr** becomes a *theorem* in any calculus which, like $\mathcal{D}^y\mathcal{CEC}_3^*$, subsumes the event calculus. The reason is simply that each fluent has a boolean value of true when it holds, and admits of no "internal divisibility" that would allow aspects of it to be reversed. Hence, any fluent intended to denote a particular P-conscious state that an agent is in over some interval will offer no internal structure to admit the possibility of reversibility.

[28] While doing what one wants to do may seem like an exceedingly low bar for ascribing freedom to an agent (after all, if with electrodes planted secretly in your brain an evil scientist gives you the wholly uncharacteristic desire to steal a wallet, and you steal it for that reason, we would rationally be loathe to say that your larceny was free!), it seems to be a higher one than what AI's John McCarthy has apparently said suffices in the case at least of robots; see (McCarthy 2000).

tion involving events. Between these two endpoints of freedom-as-doing-what-one-desires versus freedom-as-a-form-of-causation-outside-physics fall many alternatives. In addition, there are those who simply deny that human agents are free, and perhaps even some who hold that it's physically (and perhaps even logically) impossible for *any* sort of agent to be free. Overall, then, it should be easy enough for our readers to agree that any axiom of freedom is bound to be quite controversial.

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 1992*a*), our tack here will be more ecumenical: We will "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. Our axiom will assert 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** will leave 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-determinattion. We assume only that any instantiation to $\mathscr{C}$ is itself an axiom system; this in principles opens the door to seamless integration and exploration of the combination of $\mathcal{CA}$ and $\mathscr{C}$.[29] Here's the axiom:

Free Agents 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 physics theory $\mathscr{C}$) by any prior events, and where such decisions are the product itself of a decision on that same agent's part.

This axiom can of course be further "backed off" so as to drop its sub-assertion that agents perceive that their decisions are the product of decisions. In the sub-section (5.10) that immediately follows, we shall urge the adoption of an axiom of a human agent's knowledge of causation in a naïve sense that is reminiscent of classical mechanics; we do so by invoking the aforementioned event calculus.

5.10 The Causation Axiom (CommCaus)

We have admitted that numerous formal models of time, change, and causation have been presented in the literature, even if we restrict ourselves to the AI literature. We have also pointed out that there are numerous accounts of causation available from physics itself; this is of course why we have availed ourselves of the placeholder $\mathscr{C}$. In much prior work, and in the present case via $\mathcal{D}^y\mathcal{CEC}_3^*$ and $\mu\mathcal{D}^y\mathcal{CEC}_3^*$, 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 by Bringsjord in (Arkoudas & Bringsjord 2009). There are some variations in how the event calculus is axiomatized, and there is nothing to be gained, given our chief purposes in the present paper, by discussing these variations. $\mathcal{D}^y\mathcal{CEC}_3^*$ and $\mu\mathcal{D}^y\mathcal{CEC}_3^*$ axiomatize the event calculus with five formulae, which we needn't canvass here. To give a flavor, the third and fourth of these axioms are:

$$\textbf{EC3}\;\; \mathbf{C}\{\forall\, t_1,\ f,\ t_2\ [\mathit{clipped}(t_1,\ f,\ t_2)\ \leftrightarrow\ \exists\, e, t\ (\mathit{happens}(e,\ t) \wedge t_1 < t < t_2 \wedge \mathit{terminates}(e,\ f,\ t))]\}$$

[29] For classical mechanics, a very early instantiation to $\mathscr{C}$ is provided by McKinsey, Sugar & Suppes (1953). Axiomatizations are now available for not only classical mechanics, but also quantum mechanics, and both special and general relativity. For an initial exploration of such axiomatizations via formal methods and AI, see e.g. (Govindarajalulu, Bringsjord & Taylor 2015).

EC4 $\mathbf{C}\{\forall\ a,\ d,\ t\ [\mathit{happens}(\mathit{action}(a,\ d),\ t) \rightarrow\ \mathbf{K}(a,\ \mathit{happens}(\mathit{action}(a,\ d),\ t))]\}$

EC3 says that it's common knowledge that if a fluent ceases to hold between times t_1 and t_2, some event e is responsible for terminating that fluent. As to **EC4**, it expresses that it's common knowledge that if some action is performed by an agent at t, the agent in question knows that it has performed the action in question.

For our next axiom, we simply assign to '$\mathcal{EC}$' some standard axiomatization, and employ the common-knowledge operator **C**; this allows us to formulate the Causation Axiom perspicuously follows:

CCaus $\mathbf{C}\ \mathcal{EC}$

Notice that we can easily and quickly abstract from **CCaus** to an axiom *schema*, by simply supplanting $\mathcal{EC}$ with the placeholder $\mathscr{C}$.

5.11 The "Perry" Axiom (TheI)

We come finally to axiom **TheI**, which we dub the "Perry" Axiom, in honor of a thought-experiment devised by John Perry (1977):

> An amnesiac, Rudolf Lingens, is lost in the Stanford library. He reads a number of things in the library, including a biography of himself, and a detailed account of the library in which he is lost. *ldots* He still won't know who he is, and where he is, no matter how much knowledge he piles up, until that moment when he is ready to say, 'This place is aisle five, floor six, of Main Library, Stanford. I am Rudolf Lingens.' (Perry 1977, p. 492)

$\mathcal{CA}$'s final axiom 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. Here's the axiom:

TheI Let P^e be any empirical, contingent property and P^i be any internal, Cartesian property; and let I^* be the self-designator for an agent a. Then $\mathbf{K}(I^*, P(I^*)) \wedge \nvdash P(I^*)$.

One interesting aspect of **TheI** is that it can be viewed as a sort of "pivot:" We can have one set of axioms that is streamlined by the constraint that only those axioms that would be operative in Perry's library are to be considered, and then the other set generated by the notion that we're dealing with a person in "full operation." We leave it to the reader to ponder how this partitioning would work for the 10 axioms other than **TheI** that we have presented above.

6 Next Steps

We have explicitly said that the axiom system given above, $\mathcal{CA}$, is humbly offered as starting phase in the erection of a mature axiomatization of consciousness. We are of course under no illusions that at least an appreciable portion of $\mathcal{CA}$ will be controversial. The next step in the refinement, defense, and possible extension of $\mathcal{CA}$ is clearly to provide a fully formal version of the axioms which, for readability and efficiency herein, we left somewhat informal. This step we have accomplished, and look forward to publishing.

A second equally obvious direction for future work is already underway, and will, we hope, soon bear fruit. The direction is that of discovering and examining theorems, for the overriding goal of bringing forth $\mathcal{CA}$ is to bring forth a *theory* of consciousness, where by 'theory' we mean the collection of all that can be proved from the axioms, that is $\mathscr{C} := \{\phi : \mathcal{CA} \vdash \phi\}$. We are hopeful that since $\mathcal{CA}$ can be, as planned, implemented, the tools of AI and automated theorem proving will help in plumbing $\mathscr{C}$.

References

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

Aleksander, I. & Dunmall, B. (2003), 'Axioms and Tests for the Presence of Minimal Consciousness in Agents', *Journal of Consciousness Studies* **10**, 7–18.

Aleksander, I. & Morton, H. (2007), Axiomatic Consciousness Theory For Visual Phenomenology in Artificial Intelligence, *in* A. Chella & R. Manzotti, eds, 'AI and Consciousness: Theoretical Foundations and Current Approaches', AAAI, Menlo Park, CA, pp. 18–23. The proceedings is Tech Report FS-07-01 from AAAI.
URL: *https://www.aaai.org/Papers/Symposia/Fall/2007/FS-07-01/FS07-01-004.pdf*

Arkoudas, K. & Bringsjord, S. (2004), Metareasoning for Multi-agent Epistemic Logics, *in* 'Proceedings of the Fifth International Conference on Computational Logic In Multi-Agent Systems (CLIMA 2004)', Lisbon, Portugal, pp. 50–65.
URL: *http://kryten.mm.rpi.edu/arkoudas.bringsjord.clima.crc.pdf*

Arkoudas, K. & Bringsjord, S. (2007), 'Computers, Justification, and Mathematical Knowledge', *Minds and Machines* **17**(2), 185–202.
URL: *http://kryten.mm.rpi.edu/ka_sb_proofs_offprint.pdf*

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*

Balcombe, J. (2016), *What a Fish Knows: The Inner Lives of Our Underwater Cousins*, Scientific American / Farrar, Straus and Giroux, New York, NY.

Benzmüller, C. & Paleo, B. W. (2014), Automating Gödel's Ontological Proof of Gods 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.
URL: *http://page.mi.fu-berlin.de/cbenzmueller/papers/C40.pdf*

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

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

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

Bringsjord, S. (1998), 'Chess is Too Easy', *Technology Review* **101**(2), 23–28.
URL: *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.
URL: *http://kryten.mm.rpi.edu/jcsonebillion2.pdf*

Bringsjord, S. (2008*a*), Declarative/Logic-Based Cognitive Modeling, *in* R. Sun, ed., 'The Handbook of Computational Psychology', Cambridge University Press, Cambridge, UK, pp. 127–169.
URL: *http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf*

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

Bringsjord, S. (2010), 'Meeting Floridi's Challenge to Artificial Intelligence from the Knowledge-Game Test for Self-Consciousness', *Metaphilosophy* **41**(3), 292–312.
URL: *http://kryten.mm.rpi.edu/sb_on_floridi_offprint.pdf*

Bringsjord, S. (2015), 'A Vindication of Program Verification', *History and Philosophy of Logic* **36**(3), 262–277. This url goes to a preprint.
URL: *http://kryten.mm.rpi.edu/SB_progver_selfref_driver_final2_060215.pdf*

Bringsjord, S. & Govindarajulu, N. S. (2013), Toward a Modern Geography of Minds, Machines, and Math, *in* V. C. Müller, ed., 'Philosophy and Theory of Artificial Intelligence', Vol. 5 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, Springer, New York, NY, pp. 151–165.
URL: *http://www.springerlink.com/content/hg712w4l23523xw5*

Bringsjord, S., Licato, J., Govindarajulu, N., Ghosh, R. & 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. This URL goes to a preprint of the paper.
URL: *http://kryten.mm.rpi.edu/SBringsjord_etal_self-con_robots_kg4_0601151615NY.pdf*

Bringsjord, S. & 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.

Chisholm, R. (1964), Freedom and Action, *in* K. Lehrer, ed., 'Freedom and Determinism', Random House, New York, NY, pp. 11–44.

Cunningham, J. (2001), 'Towards an Axiomatic Theory of Consciousness', *Logic Journal of the IGPL* **9**(2), 341–347.

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

Edwards, J. (1957), *Freedom of the Will*, Yale University Press, New Haven, CT. Edwards originally wrote this in 1754.

Floridi, L. (2005), 'Consciousness, Agents and the Knowledge Game', *Minds and Machines* **15**(3-4), 415–444.
URL: *http://www.philosophyofinformation.net/publications/pdf/caatkg.pdf*

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

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

Goble, L., ed. (2001), *The Blackwell Guide to Philosophical Logic*, Blackwell Publishers, Oxford, UK.

Govindarajalulu, N. S., Bringsjord, S. & Taylor, J. (2015), 'Proof Verification and Proof Discovery for Relativity', *Synthese* **192**(7), 2077–2094.

Govindarajulu, N. S. (2011), Towards a Logic-based Analysis and Simulation of the Mirror Test, *in* 'Proceedings of the European Agent Systems Summer School Student Session 2011', Girona, Spain.
URL: *http://eia.udg.edu/easss2011/resources/docs/paper5.pdf*

Green, C. (1969), Applications of Theorem Proving to Problem Solving, *in* 'Proceedings of the 1st International Joint Conference on Artificial Intelligence', Morgan Kaufmann, San Francisco, CA, pp. 219–239.

Halpern, J. & Shoham, Y. (1991), 'A Propositional Modal Logic of Time Intervals', *Journal of the ACM* **38**(4), 935–962.

Harries, R. (2007), Half Ape, Half Angel?, *in* C. Pasternak, ed., 'What Makes Us Human?', Oneworld Publications, Oxford, UK, pp. 71–81.

Hendricks, V. & Symons, J. (2006), Epistemic Logic, *in* E. Zalta, ed., 'The Stanford Encyclopedia of Philosophy'.
URL: *http://plato.stanford.edu/entries/logic-epistemic*

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

Hulne, D. (2007), Material Facts from a Nonmaterialist Perspective, *in* C. Pasternak, ed., 'What Makes Us Human?', Oneworld Publications, Oxford, UK, pp. 82–92.

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

Ichikawa, J. & Steup, M. (2012), The Analysis of Knowledge, *in* E. Zalta, ed., 'The Stanford Encyclopedia of Philosophy'.
URL: *http://plato.stanford.edu/entries/knowledge-analysis*

Jacquette, D. (1994), *Philosophy of Mind*, Prentice Hall, Englewood Cliffs, NJ.

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

Johnson-Laird, P. & 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. & Stubblefield, W. (1993), *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, Benjamin Cummings, Redwood, CA.

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

McCarthy, J. (2000), 'Free will–even for robots', *Journal of Experimental and Theoretical Artificial Intelligence* **12**(3), 341–352.

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

Miranker, W. & Zuckerman, G. (2008), 'Mathematical Foundations of Consciousness'. This paper is also available from Yale-University servers as a technical report (TR1383).
URL: *https://arxiv.org/pdf/0810.4339.pdf*

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

Penn, D., Holyoak, K. & 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.

Pink, T. (2004), *Free Will: A Very Short Introduction*, Oxford University Press, Oxford, UK.

Pollock, J. (1995), *Cognitive Carpentry: A Blueprint for How to Build a Person*, MIT Press, Cambridge, MA.

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.

Rapaport, W. (1998), 'How Minds Can Be Computational Systems', *Journal of Experimental and Theoretical Artificial Intelligence* **10**, 403–419.

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

Smith, P. (2013), *An Introduction to Gödel's Theorems*, Cambridge University Press, Cambridge, UK. This is the second edition of the book.

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.