© 2010 The Author
Journal compilation © 2010 Metaphilosophy LLC and Blackwell Publishing Ltd
Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK, and
350 Main Street, Malden, MA 02148, USA
METAPHILOSOPHY
Vol. 41, No. 3, April 2010
0026-1068

MEETING FLORIDI'S CHALLENGE TO ARTIFICIAL INTELLIGENCE FROM THE KNOWLEDGE-GAME TEST FOR SELF-CONSCIOUSNESS

SELMER BRINGSJORD

Abstract: In the course of seeking an answer to the question "How do you know you are not a zombie?" Floridi (2005) issues an ingenious, philosophically rich challenge to artificial intelligence (AI) in the form of an extremely demanding version of the so-called knowledge game (or "wise-man puzzle," or "muddy-children puzzle")—one that purportedly ensures that those who pass it are self-conscious. In this article, on behalf of (at least the logic-based variety of) AI, I take up the challenge—which is to say, I try to show that this challenge can in fact be met by AI in the foreseeable future.

Keywords: artificial intelligence, consciousness, self-consciousness, knowledge game.

1. Introduction

In the course of seeking an answer to the Dretskean (2003) question "How do you know you are not a zombie?" Floridi (2005) issues an ingenious, philosophically rich challenge to artificial intelligence (AI) in the form of an extremely demanding version of the so-called knowledge game (or "wise-man puzzle," or "muddy children puzzle")—one that purportedly ensures that those who pass it are self-conscious. We shall call this test "KG$_4$"; the significance of the subscript will be clear in due course.

In this essay, on behalf of (at least the logic-based variety of) AI, I take up Floridi's challenge—which is to say, I try to show that this challenge can in fact be met by AI in the foreseeable future. I'm quite convinced that zombies are logically *and* physically possible, and, indeed, that zombies are precisely what logic-based AI, in the long run, will produce (see, e.g., Bringsjord 1995b); that this possibility is enough to refute the view that human consciousness can be replicated through computation (see, e.g., Bringsjord 1999); that the engineering power of logic-based AI (Bringsjord 2008b) is truly formidable; that *any* behavioral test is within the reach of this form of AI (Bringsjord 1995a); and that AI of any variety ought in fact to be guided by the goal of building artificial agents able to pass tests demanding human-level intelligence (Bringsjord and Schimanski 2003). Therefore, it should be easy enough for the reader to understand that I find Floridi's article to be not only relevant but preternaturally so, and that I'm rather motivated to accept his challenge. Of course, anyone convinced not only of AI's ability to eventually create creatures that *appear* to have minds but also of its ability to produce artificial *minds*, will want to show that Floridi's challenge can be surmounted. One of the remarkable aspects of his article is that it targets *both* "weak" and "strong" AI.[1]

The plan of my essay is as follows: In the next section (2) a number of preliminary tasks are completed. For example, I explain the different forms of consciousness relevant to Floridi's knowledge game, and set out the structure of his test-based answer to how-do-you-know-you-are-X questions. In section 3, I review the knowledge game as Floridi sets it out, which includes four increasingly demanding versions. Special attention is paid to the reasoning carried out by agents who pass the third and fourth versions of the game. Then I show in section 4 that Floridi's pessimism about the power of robots and zombies to pass KG_4 is unwarranted, in light of my proof-sketch showing that a robot can deduce the solution to this version of the game. Two objections are then rebutted in section 5, and a brief concluding section (6) wraps up the essay.

2. Preliminaries

A number of preliminaries must be dealt with before we start in earnest. Let's begin with a characterization of the types of consciousness that are central to Floridi's test.

2.1. Types of Consciousness

Following Floridi, we shall distinguish three types of consciousness: *access consciousness* (abbreviated as *a-consciousness*), *phenomenal consciousness* (*p-consciousness*), and *self-consciousness* (*s-consciousness*).[2] This trio is part of the standard terminological furniture of modern philosophy of mind. For example, Block distinguishes between p-consciousness and a-consciousness. The latter concept is characterized by

[1] Briefly: Weak AI: Standard (= Turing-level) computing machines, perhaps suitably connected by sensors and effectors to the external environment, can eventually be engineered to match (≈ *simulate*) the outward behavior of human persons. Strong AI: Standard computing machines, perhaps . . . , can eventually be engineered so as to literally *replicate* the inner mental lives of human persons.

[2] Actually, Floridi speaks of *environmental consciousness* rather than a-consciousness, but the two concepts are equivalent, as Floridi himself avers. Floridi says that an agent is environmentally conscious if it "is able to process information about, and hence to interact with, [its] surroundings, its features and stimuli effectively, under normal circumstances" (2005, 417). In the present essay I run with "a-consciousness" in view of the fact that in AI and philosophy of AI this is a more familiar term.

him as follows: "A state is access-conscious ([a]-conscious) if, in virtue of one's having the state, a representation of its content is (1) inferentially promiscuous, i.e., poised to be used as a premise in reasoning, and (2) poised for [rational] control of action, and (3) poised for rational control of speech" (Block 1995, 231). As I have explained elsewhere (Bringsjord 1997), and as Floridi agrees, it's plausible to regard certain extant, mundane computational artifacts to be bearers of a-consciousness. For example, theorem provers with natural-language generation capability, and certainly sophisticated autonomous robots, would qualify. It follows immediately that a zombie would be a-conscious.

And now here is Block's characterization of p-consciousness, which matches what Floridi has in mind: "So how should we point to [p]-consciousness? Well, one way is via rough synonyms. As I said, [p]-consciousness is experience. P-conscious properties are experiential properties. P-conscious states are experiential states, that is, a state is [p]-conscious if it has experiential properties. The totality of the experiential properties of a state are 'what it is like' to have it. Moving from synonyms to examples, we have [p]-conscious states when we see, hear, smell, taste and have pains. P-conscious properties include the experiential properties of sensations, feelings and perceptions, but I would also include thoughts, wants and emotions" (Block 1995, 230). What about s-consciousness? Here is Floridi's description of this concept (where "Ag" stands for any agent): "Ag may be *self-conscious* if Ag has a (second- or higher-order) sense of, or is (introspectively) aware of, Ag's personal identity (including Ag's knowledge that Ag thinks) and (first- or lower-order) perceptual or mental experiences (including Ag's knowledge of what Ag is thinking)" (Floridi 2005). As we can see, none of these three definitions is precise, let alone formal. But that is certainly not Floridi's fault. *No one* has formal accounts of these varieties of consciousness on hand, and we shall thus, of necessity, make do with the descriptions given above.[3]

2.2. Types of Agents

I further follow Floridi in partitioning the class of relevant agents into three categories; namely, *human persons* (who enjoy *a-*, *p-*, and *s-consciousness*), *robots* or *artificial agents* (said by Floridi to be "endowed with interactivity, autonomy and adaptability" [2005, 420]), and *zombies* (who have *a-consciousness* but lack *p-* and *s-consciousness*). Hereafter I refer simply to *persons* as the first class (wanting as I do to leave aside, for example, divine persons), *robots* as the second, and *zombies* as the third. Please note that in AI it's common linguistic practice to regard

[3] Despite the fact that we don't have formal definitions of a-, p-, and s-consciousness, it seems clear that there are some logical relations holding between these concepts. For example, it specifically seems clear that anything that is s-conscious is p-conscious. This is a principle Floridi employs and defends in this paper. I happily affirm the principle.

devices to be bona fide artificial agents or robots even when they are remarkably dim. For example, Russell and Norvig (2002) classify computer programs that do no more than compute elementary number-theoretic functions as artificial agents, and the same lattitudinarian approach holds in AI for the domain of robots as well. In Floridi's scheme, and the present essay's (which premeditatedly inherits directly from Floridi's), *robots* must be capable of interacting with other agents and the external environment, have autonomy, and be adaptable. For a discussion of a continuum of sophistication for robots and artificial agents directly relevant to the present essay, see Bringsjord, Noel, and Caporale 2000.

2.3. The Test-Based Answer to the Question

Floridi takes Dretske's question to be one "that can take as an answer 'a way of knowing that, unlike zombies, we are conscious of things,' that is, how one can *possibly* know that one is a zombie" (Floridi 2005, 419). Understood this way, Floridi maintains that there is a test-based way to answer the question. In fact, Floridi generalizes the situation, and explains that there is a test-based way to answer the question *type*: "How do you know that you are an *X*?" We read: "Good tests usually are informative, that is, they usually are more than just successful criteria of identification of *x* as [*X*], because they examine the very process they are testing precisely while the process is occurring, and so they provide the tested agent with a way of [(c1)] showing that he qualifies as a certain kind of agent, [(c2)] knowing that he is that kind of agent, and [(c3)] answering how he knows that he is that kind of agent, by pointing to the passed test and its [(c1)–(c3)] features" (2005, 420).

Where *X* is any attribute, we can sum up Floridi's approach via figure 1. In this figure, a test is said to include a stem **S**, a question **Q**, and an environment **E**. (Of course, if we were to specify a full "ontology" of testing, we would need to invoke additional categories; for example, testers. But we are streamlining, without loss of generality, and while to facilitate exposition we shall discuss other categories—we shall speak of testers: the prisoners in Floridi's knowledge game—we shall not explicitly build these categories into our explicit representations or into the reasoning of testees.) The stem refers to information that the tester gives the testee before asking the key question **Q**, and the environment consists of information that the testee can gain by sense perception. For example, in the first of the knowledge-game tests presented by Floridi, the "classic" version of the knowledge game, the tester/guard announces to the testees/prisoners that there are five fezzes of a particular color distribution; this information is part of the stem. The question is simply "What color is your fez?" And the environment for a testee includes the color of the fez atop the heads of two other testees. This color can be readily perceived through vision by each of the prisoners.

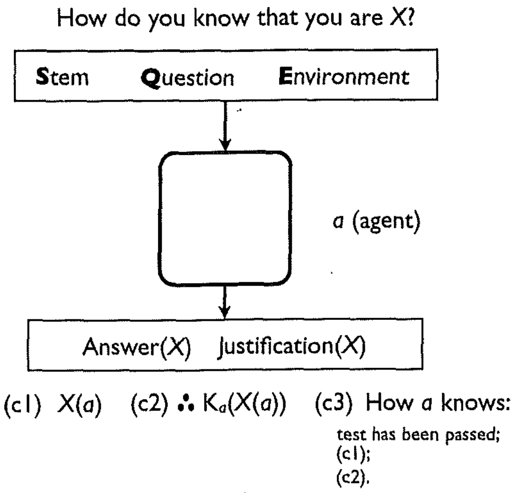
How do you know that you are X?
Stem Question Environment
a (agent)
Answer(X) Justification(X)
(c1) X(a) (c2) ∴ K_a(X(a)) (c3) How a knows:
test has been passed;
(c1);
(c2).

Figure 1. Schematic overview of test-based answer to question type

In order to seek an answer to Dretske's question, we have only to set $X = \textit{not a zombie}$ in the schema of figure 1. As we shall see, Floridi believes he has found assignments to **S**, **Q**, and **E**, in the test KG_4, that provide an answer to Dretske's question.

3. The Knowledge-Game Quartet

Floridi (2005) considers a continuum of four versions of the knowledge game, which we shall refer to as KG_1/KG'_1, KG_2, KG_3, and KG_4. (The only difference between KG_1 and KG'_1 is that in the former, each prisoner answers the question separately, whereas in the latter the trio answers simultaneously as a multiagent system.) In addition, we shall give each version an informal mnemonic label to help us remember something distinctive about a particular version. As you will recall, the last version of the knowledge game, KG_4, is the test for self-consciousness that supposedly separates zombies from persons, and of course supposedly agents from persons as well. This claimed separation is conveyed by table 1, which also expresses the rest of Floridi's claims with respect to whether agents of the three aforementioned types: can pass now (√), fail now but possibly pass in the future (?), fail now but will pass in the future (*G*), or are forever doomed to fail (×).

Table 2 expresses my position, which as the reader can see is rather more optimistic than Floridi's from the perspective of AI.

I shall now briefly review each member of Floridi's continuum.

TABLE 1. Four versions of the knowledge game (Floridi)

Version	Label	Agent Type		
		robots	zombies	Persons
KG_1	"classic"	√	√	√
KG_2	"boots"	√	√	√
KG_3	"deafening"	?	√	√
KG_4	"self-consciousness"	×	×	√

TABLE 2. Four versions of the knowledge game (Bringsjord)

Version	Label	Agent Type		
		robots	zombies	persons
KG_1	"classic"	√	√	√
KG_2	"boots"	√	√	√
KG_3	"deafening"	G	√	√
KG_4	"self-consciousness"	G	G	√

3.1. The "Classic" Version (KG1/KG′1)

The first test in the continuum, KG_1/KG'_1, in many ways serves as a time-honored portal to logic-based formalisms and techniques in AI (see, e.g., Fagin et al. 2004, Arkoudas and Bringsjord 2005, Genesereth and Nilsson 1987), and Floridi sums it up as follows:

> A guard challenges three prisoners, *A*, *B*, and *C*. He shows them five fezzes, three red and two blue, blindfolds them and makes each of them wear a red fez, thus minimising the amount of information provided. He then hides the remaining fezzes from sight. When the blindfolds are removed, each prisoner can see only the other prisoners' fezzes. At this point, the guard says: "If you can tell me the colour of your fez you will be free. But if you make a mistake or cheat you will be executed."
>
> The guard interrogates *A* first. *A* checks *B*'s and *C*'s fezzes and declares that he does not know the colour of his fez. The guard then asks *B*. *B* has heard *A*, checks *A*'s and *C*'s fezzes, but he too must admit he does not know. Finally, the guard asks *C*. *C* has heard both *A* and *B* and immediately answers: "My fez is red." (2005, 422)

As astute readers will immediately appreciate, *C* is quite right, and is therefore released. Readers are expected not only to be able to grasp that *C* is correct (that is, to grasp that *C*'s fez is red) but also to be able to prove that *C*'s fez is red (using what *C* knows). For machine-generated and machine-checked proofs that support *C*'s response, see Arkoudas and Bringsjord 2005.

3.2. The "Boots" Version (KG2)

In KG_2, five pairs of boots are used instead of the fez quintet; two pairs are ordinary, but three are "torturing instruments that crush the feet" (Floridi 2005, 426). The question to the contestants here, of course, is whether one has donned the hurtless variety or the crushing kind. The answer must be given synchronically by the trio.

Floridi is quite right that each type of agent can pass this test with flying colors, and indeed it takes only a modicum of familiarity with the current state of robotics, combined with but a touch of imagination, to grasp that KG_2 could really and truly be passed by today's non-p-conscious and non-s-conscious robots, armed as they are with standard, purely mechanical sensors of various kinds. Therefore, as Floridi correctly asserts, "[b]ootstrapping states are useless for discriminating between humans and zombies" (2005, 426).

3.3. The "Deafening" Version (KG3)

What distinguishes Floridi's "deafening" version of the knowledge game is that the question **Q** in this case is *self-answering*; such a question is one "that answers itself if one knows how to interpret it" (Floridi 2005, 428). From the perspective of AI (for the basic formal scheme see, e.g., Sun and Bringsjord 2009), this means that a self-answering question $\mathbf{Q}_{SA}$, once parsed by an artificial agent or robot, delivers knowledge φ_Q which, when combined with prior knowledge Φ_p possessed by this agent, allows the agent to infer the correct answer.[4] As a first example, Floridi supplies (**Q4E**) "How many were the four evangelists?"[5]

While Floridi's pessimism about AI's ability to produce s-conscious and p-conscious robots (or zombies), for the purposes of the present

[4] What I say here may strike some alert readers as odd. They may ask: "Don't all questions get answered on the basis of both background knowledge and knowledge (however small it may be) by the question itself? What then distinguishes a *self-answering* question?" A fully satisfying reply would require more space than I have here, but I volunteer that a self-answering question is marked by the fact that answering it correctly can be done without moving outside the bounds of the *a priori* and analytic—as is borne out in the example I very soon give (i.e., **Q3**).

[5] This is actually a rather interesting specimen, because it has a close non-self-answering relative that is effortlessly correctly answered on the basis of *only* standard prior knowledge: "How many were the evangelists?" Or, a less awkward version: "How many evangelists were there?" (There are some unaware of the fact that the quartet in question corresponds to the traditional authors of the four gospels. I don't mean to imply that the background knowledge here is had by everyone. And there are other complications I leave aside, such as that in some heterodox frameworks the writers of the *apocryphal* gospels count as evangelists.)

dialectic, is to rest upon his KG_4, it is worthwhile to note that this pessimism first surfaces in his paper in connection with KG_4's predecessor, KG_3: Floridi claims that "*current* and *foreseeable* artificial agents [= robots] as we know them cannot answer self-answering questions, either in a stand-alone or in a multiagent setting" (2005, 431). He is certainly right about current robots; he is probably wrong about foreseeable ones.

To see this, consider that, from the standpoint of logic-based AI, engineering a robot that understands and correctly answers (and justifies that answer) some self-answering questions seems surprisingly straightforward, when you think about some of these questions a bit. For example, consider the self-answering question **Q3**: "What is the cardinality of the set composed of three elements?" Clearly, this question conveys declarative information; specifically, declarative information that captures the nature of the set in question. This information can be expressed in standard first-order logic, following the customary language of set theory; for example:

$$\exists y \exists x_1 \exists x_2 \exists x_3 [a = y \land x_1 \in y \land x_2 \in y \land x_3 \in y \land x_1 \neq x_2 \land x_2 \neq x_3 \land x_1 \neq x_3]$$

If we let this formula be denoted by φ, then a robot seeking to answer **Q3** would be seeking to verify

$$\Phi_{Q3} \vdash \exists n(card(a) = n \land \varphi)$$

and this proof can indeed be found by automatic theorem provers armed with the standard machinery of set theory underlying the cardinality of finite sets. Such a proof is elementary, and is found quickly by novices taking their first course in axiomatic set theory.[6] The upshot of this example is that even *current* logic-based AI is able to handle some self-answering questions.

Notice, though, that I say "some" self-answering questions. There is indeed a currently insurmountable obstacle facing logic-based AI that is related, at least marginally, to self-answering questions: it is simply that current AI, and indeed even *foreseeable* AI, is undeniably flat-out impotent in the face of *any arbitrary* natural-language question—whether or not that question is self-answering. To be a bit more precise: Take an artificial agent or robot, stipulate boldly that it's the absolute best that AI can muster today; or bolder still, imagine the best such being that can be mustered by learned extrapolation into the future from where AI is today. Let's dub this wondrous robot "*R*." Now imagine a test that is radically streamlined relative to Floridi's elaborate KG_i, namely, a test in which the question to *R* is just a single fact-finding query; **Q***, let's say. For example: "Is the Vatican south of a tall, largely open-air metal tower located near a

[6] For sample proofs of this type, quite elementary, see Suppes 1972.

river that was home to a famous siege perpetrated by Selmer Bringsjord's violent ancestors?" Of course, the test isn't made any easier if **Q*** happens to be self-answering; the following, for instance, would doubtless stump our *R* as well: "Ceteris paribus, can a superior extemporaneous human debater raised in the United Kingdom learn to play legal chess in an afternoon, if that session is her first exposure to the game?"[7]

It should be noted that both Floridi and I refrain from claiming that no robot will *ever* be able to answer arbitrary fact-finding questions. This can be seen by looking at tables 1 and 2, where for KG_3 it will be seen that in Floridi's case he admits that this version of the game may be passed by a future robot ("?"), and in my own case there is the claim that this test is going to be passed in the future ("*G*").

It should also be noted that Floridi specifically classifies self-answering questions like **Q3** as "internally, semantically self-answering" questions, while the question in the case of the "boots" game is, as he says, "self-answering in a more complex way, for the answer is *counterfactually embedded* in [**Q**] and it is so somewhat 'indexically' since, under different circumstances, the question or the questioning would give nothing away" (2005, 428). The counterfactual aspects of Floridi's depiction of *A*'s reasoning in KG_3 are clearly present; here is that depiction of the reasoning needed to crack KG_3 (where the state *S* is *hearing the guard's question*, the state *D* is *being deaf*, and *Q* is the guard's question): "A reasons that if *A* were in $\neg S$ then *A* would be in some state *D*; but if *A* were in *D* then *A* could not have received *Q* [= **Q**]; but *A* received *Q*, so *A* could receive *Q*, so *A* is not in $\neg S$, but *A* is in either *S* or $\neg S$, so *A* is in *S*" (Floridi 2005, 429). I end this section by pointing out that *if the counterfactual aspects of Floridi's description of A's reasoning in KG_3 are ignored*, it's easy enough to understand that elementary logic-based AI techniques can be used to express and certify this reasoning.[8] Such understanding arrives once one sees that (1) the core reasoning in standard extensional form is simple, and that (2) such reasoning can be easily certified by any of today's decent automated theorem provers. For the first point, simply confirm mentally that

$$\{\forall x(Sx \rightarrow Dx), \forall x(Dx \rightarrow \neg RecQx), RecQa, \forall x(Sx \vee (Sx))\} \vdash Sa$$

[7] We can of course debate what is and isn't a self-answering question. After all, no identity conditions for such questions have been supplied (by anyone, as far as I can tell; Floridi points out parenthetically that such questions have received surprisingly little attention in logic). The question here, by Floridi's lights, may not be self-answering, but that it is is provably consistent with the formal elements I introduced as general constraints on such questions in the previous paragraph.

[8] This means that some necessary conditions for logic-based-AI machines passing the test are satisfied. Any such machine, if passing tests like those Floridi presents, must have a sufficiently expressive underlying language and proof theory (or *argument* theory in nondeductive cases; see Bringsjord 2008a), and a means of using these elements in a reasonable amount of time. We shall later discuss what is *sufficient* to pass the relevant class of tests.

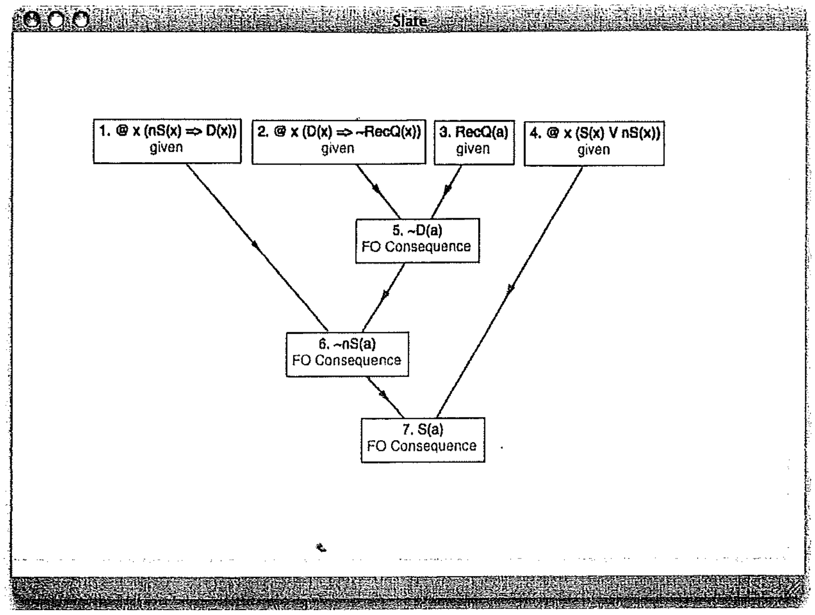

Figure 2. Proof for "cracking" KG_3 machine certified in the Slate system (courtesy of Joshua Taylor)

and then, for the second point, observe that in figure 2, using the Slate system (Bringsjord et al. 2008), the "decounterfactualized" reasoning has been certified by one of the automated theorem provers included in Slate (viz., SNARK; see Stickel et al. 1994).

Of course, it may be asked: "What right have you, though, to decounterfactualize the reasoning?" Since, as I have noted from the outset of this essay, KG_4, not KG_3, is the most serious challenge to AI that Floridi has fashioned, and since the fourth version of the knowledge game, as Floridi sees the situation, *also* requires counterfactual reasoning of those agents who successfully negotiate it, I will save my answer to this question/objection until I analyze KG_4 (in section 4).

3.4. The "Self-Consciousness" Version (KG4)

We come now to the positively ingenious "self-consciousness" version of the knowledge game, in which our three embattled prisoners are given not fezzes or boots or beverages but a tablet from a collection of five, three of which are innocuous, while two make those ingesting them completely dumb.

Here is the (once again counterfactual) reasoning that leads *A* to give the correct answer, in Floridi's words:

> [H]ad I taken the dumbing tablet I would not have been able to report orally my state of ignorance about my dumb/non-dumb state, but I have been and I know that I have been, as I have heard myself speaking and saw the guard reacting to my speaking, but this (my oral report) is possible only if I did not take the dumbing tablet, so now I know that I am in a non-dumb state, hence I know that I have not taken the dumbing tablet, and I know that I know all this, that is, I know that my previous state of ignorance has now been erased, so I can revise my statement and reply, correctly, in which state I am, which is a state of not having taken the dumbing tablet, of knowing that I haven't, and—by going through this whole process and passing the test—of knowing how I know both that I haven't and that I know that I haven't.
>
> (Floridi 2005, 432–33)

4. AI, Contra Floridi, Can Handle KG4

I have argued elsewhere that a grant today of one billion U.S. dollars would be insufficient to allow a first-rate (for that matter, the world's preeminent) robotics R&D group to produce, even in an exceedingly long project, a p-conscious robot (Bringsjord 2007). In fact, I argue that such a group, however well financed, would not even know where to *start*. Things are radically different in the case of KG$_4$. In this case, it can be demonstrated that foreseeable AI can produce an artifact capable of passing. The demonstration consists in showing that such an artifact can, now, if sufficient time and energy is expended to carry out the engineering, apparently be produced. While I don't have enough space here to supply the demonstration in the form of a proof, I *do* have sufficient space to provide a proof-*sketch*, which I give below. This reasoning, expressed in significant part in the *cognitive event calculus*, ***CEC*** for short (for formal details, see Arkoudas and Bringsjord 2009), should be more than detailed enough to justify the assertion that *foreseeable* AI will be up to the task.[9]

Before I present the proof-sketch, a brief overview of the language of ***CEC*** is in order. The reason is that some readers may be unfamiliar with *multisorted* logic (MSL), and with some of the core concepts in the event-calculus approach to reasoning about time and change.

In standard first-order logic (FOL), as is well known, quantification in any formal theory, or in any knowledge base for some problem or application, is over a so-called *domain*, which is simply a set. For example, suppose that we have a domain D composed of all the people in the classroom of some introductory logic course on some particular day. There are students in D, as well as teaching assistants, and there is a professor (a) at the front of the room in question. Given this setup, to say

[9] If I were given a grant of sufficient size, I'm confident that with help from colleagues in my laboratory I could produce a working KG$_4$-passing robot within one year.

in standard FOL that every student likes every teaching assistant who likes the professor, we might write

$$\lambda : = \forall x \forall y ((Sx \wedge Sy \wedge Lya) \rightarrow Lxy)$$

But this is somewhat cumbersome, and—for reasons that certainly needn't be given here—inefficient when it comes time for a machine to reason deductively over such information. MSL allows us to partition D so that it contains the following *sorts*: `Students`, `TeachingAssistants`, and `Professors`. If we then correspondingly partition our supply of variables, and use obvious notation to indicate doing so, we can supplant λ with

$$\lambda' : = \forall s \forall t (Lta \rightarrow Lst)$$

Now, ***CEC*** is a sorted calculus; it includes the following sorts:

`Object Agent ActionType Action Event Moment Boolean Fluent`

where `Action` is a subsort of `Event`. It's important to know that `Boolean` is simply the set composed of `true` and `false`. In addition, put intuitively, a fluent is a state of affairs that, through time, can shift between holding and not. Fluents have long been used in logic-based AI.

How does one define well-formed formulas (wffs) in this approach? The grammar is a straightforward close relative of those used for FOL; a few remarks will suffice. ***CEC*** has a number of key function symbols. Here are four defined:

- *holds*: `Fluent` × `Moment` → `Boolean`
- *happens*: `Event` × `Moment` → `Boolean`
- *clipped*: `Moment` × `Fluent` × `Moment` → `Boolean`
- *initiates*: `Event` × `Fluent` × `Moment` → `Boolean`

Read informally, the first bullet says that *holds* is a function that takes a fluent and a time point and returns a truth-value giving whether or not the fluent holds at this time point. The second and fourth should be self-explanatory. The third, understood intuitively, conveys that the function *clipped* takes a time point, a fluent, and another time point, and is true when the fluent is terminated between these two times.

We are now in position to articulate the proof-sketch. Here goes:

Theorem. I, A, didn't ingest a tablet of the dumbing variety.

Proof-sketch. We begin by noting that KG_4 pivots around five time points, which we shall make privileged constants in the following manner:

- t_1 (*apprise*): This is the time point at which the prisoners are apprised of the situation; that is, the time at which they learn of the five tablets (and the two kinds therein, partitioned, recall, 3–2), and so on. In short, a good portion of the knowledge acquired from the environment **E** is here perceived by *A*. (I shall use "*a*" rather than "*A*" because constants are traditionally lowercase characters, and at any rate they are in ***CEC***; see Arkoudas and Bringsjord 2009.) I assume that this information consists of formulae in the set Φ.

- t_2 (*ingest*): The pills are ingested by the quartet. *A*, of course, consumes a nondumbing tablet.
- t_3 (*inquire*): The guard inquires as to which tablet has been taken.
- t_4 (*speak1*): *A* responds with "Heaven knows!"
- t_5 (*speak2*): *A* says: "The nondumbing variety!" (Alternatively, *A* responds with "No!" in response to the question "Did you receive a dumbing tablet?")

We observe that *a* perceives the first part of the information he will soon enough use to pass the test:

(1) $\mathbf{P}\ (a, \Phi, apprise)$

Here I make use of the **P** operator for perceiving. This is a slight adaptation of the **S** operator in ***CEC***, which represents seeing. Since we are dealing with a fair test, we know that optical and auditory illusions can be safely ignored, and hence have a rule (which I don't bother to reproduce verbatim; this is a proof-*sketch*, after all) allowing agents to infer that which they directly perceive.

It follows immediately from (1) and the rule known as DR_4 in ***CEC*** (again, for full formal details, see Arkoudas and Bringsjord 2009), viz.,

$$\mathbf{P}(agent, \phi) \Rightarrow \mathbf{K}(agent, \phi)$$

that[10]

(2) $\mathbf{K}\ (a, \Phi, apprise)$

Next, note specifically that *a* knows that if he takes the dumbing tablet at *ingest*, he can't report orally his state of ignorance at any subsequent time, unless the effects are—to use the language of the event calculus—"clipped." This result corresponds to counterfactual knowledge, in Floridi's informal version of *A*'s reasoning; keep this point in mind, for it will be quite relevant shortly (in section 5.2). I can prove this (lemma) now. First, *a* has the following knowledge on the basis of Φ:

(3) $\mathbf{K}\ (a, (initiates\ (action\ (a, eats(p)), dumb_a, ingest)))$

[10] To ease and accelerate exposition, I overload the **K** operator. In ***CEC*** proper, this operator ranges over the agent in question, and some proposition *P*. Here, I compress declarative information by allowing the operator to range over *sets* of propositions, and to have a third argument (a time point).

Next, note that *a* knows that if no clipping occurs, then he will remain dumb. More precisely, he can deduce that if the left-hand side of the biconditional inside the third axiom shown in footnote 12 (i.e., A3) is negated, then no clipping of the poison occurs—in which case he is dumb at all time points later than *ingest*. The deduction here is easy, given the fact that what is known is true,[11] combined with the standard axioms of the event calculus, which I assume to be common knowledge among the prisoners.[12]

But of course we are not home yet: our agent *a* must deduce that he did *not* receive a tablet of the dumbing variety. Thankfully, the deduction is easy.

The high-level structure of the deduction conforms to indirect proof: the assumption that agent *a did* receive a dumbing tablet leads to a contradiction, from which we infer by reductio ad absurdum that our assumption is in error, and hence the answer from *a* is a negative one.

We have already seen that if *a* did receive a dumbing tablet, then at all time points he cannot speak; hence he cannot speak at the particular time point *speak1*. Suppose for the sake of argument, then, that *a* did receive a dumbing tablet. Then by our lemma he cannot speak at *speak1*. But *a* perceives that he *does* speak at this time point. Hence he knows that he speaks at this time point. Hence he does in fact speak at this time point. Therefore a contradiction is produced. By reductio, *a* did not receive a dumbing tablet. **QED**

Hypersedulous readers are encouraged to flesh out my reasoning so as to produce a step-by-step proof. For the benefit of such folks (and to any of them yet to obtain the Ph.D. who produce the proof: please contact me about potential graduate study, immediately), I divulge that two unspoken axioms are needed for the detailed version:

- "All agents know that they know of the events they intentionally bring about themselves." Formalized:

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

[11] The rule in Arkoudas and Bringsjord 2009 is R_4 and is this: $\mathbf{K}(agent, \phi) \Rightarrow \phi$

[12] Hence we have, where **C** is the common knowledge operator:

- A1: $\mathbf{C}\ (\forall f, t\ (initially(f) \land \neg\ clipped(0, f, t) \rightarrow holds(f, t)))$
- A2: $\mathbf{C}\ (\forall\ e, f, t_1, t_2\ ((happens(e, t_1) \land initiates(e, f, t_1) \land t_1 < t_2 \land \neg\ clipped(t_1, f, t_2)) \rightarrow holds(f, t_2))$
- A3: $\mathbf{C}\ (\forall\ t_1, f, t_2\ (clipped(t_1, f, t_2) \leftrightarrow (\exists\ e, t\ (happens(e, t) \land t_1 < t < t_2 \land terminates(e, f, t))))))$

?Note that since the common knowledge operator **C** is applied to each axiom, it can be instantly deduced that all prisoners/agents know these axioms, from which it follows by R_4 that these axioms are true in this context.

- "All agents know that if an agent believes that a certain fluent f holds at t and that agent doesn't believe that f has been clipped between t and t', then that agent will believe that f holds at t'." Formalized:

  $\mathbf{C}\ (\forall\ a, f,\ t,\ t'\ ((\mathbf{B}\ (a,\ holds(f,\ t))\ \wedge\ \mathbf{B}\ (a,\ t{<}t') \wedge \neg\ \mathbf{B}\ (a,\ clipped(t, f,\ t'))) \rightarrow \mathbf{B}\ (a,\ holds(f,\ t')))$

5. Objections

5.1. *"But A's reasoning is first-person reasoning"*

The first objection runs as follows: "Your proof sketch, Bringsjord, which is intended to show a fully mechanical version of the reasoning Floridi ascribes to prisoner *A*, dodges the fact that we are talking here about *self*-consciousness. Notice the use of the first-person pronoun in the reasoning that Floridi presents as an expression of *A*'s. This pronoun is absent in your proof sketch; you use only the constant *a*, not 'I.' Hence you have failed to capture *A*'s solution."

This objection is easily dispensed with.

First, as a matter of formal logic and the specifics of ***CEC***, the fact is that restricted versions of the epistemic version of the modal axiom 4 (which marks the modal system KT4/S4) are active in the present case.[13] I can't discuss the specifics here, but the point is that knowing *P* essentially implies knowing that one knows *P*—which is a phenomenon often closely associated with first-person knowledge. (The cognitive event calculus, on the other hand, does *not* allow infinite iteration of knowledge operators. Only three iterated **K**s are permitted in any formula.)

Second, recall the traditional tripartite *de dicto*/*de re*/*de se* distinction with respect to kinds of beliefs that has become standard in rigorous epistemology. We can quickly encapsulate the distinction by listing examples (slightly adapted to present purposes) of the three types given by Chisholm (1981, 18):

- *de dicto*: The tallest man believes that the tallest man is wise.
- *de re*: There is an x such that x is identical with the tallest man, and x is believed by y to be wise.
- *de se*: The tallest man believes that he himself is wise.

My view, and the one that underlies the proof-sketch given earlier, is that Frege (1956), Husserl (1970), and Chisholm (1976) are correct that all *de re* and *de se* belief can be reduced to *de dicto* belief, given that persons are associated with individual essences, semantically. Here is how Chisholm

[13] Nonepistemic axiom 4 is: $\Box\ \varphi \rightarrow \Box\ \Box\ \varphi$. A good discussion at the propositional level is provided by Chelles 1980. A good discussion at the quantified level (and note that **CEC** is at this more expressive level) is provided by Hughes and Cresswell 1968.

summarizes the reduction view in *The First Person*: "Some philosophers—for example, Frege and Husserl—have suggested that each of us has his own idea of himself, his own *Ich-Vorstellung* or individual concept. And some of the things that such philosophers have said suggest the following view: The word 'I,' in the vocabulary of each person who uses it, has for its referent that person himself and has for its sense that person's *Ich-Vorstellung* or individual concept. The difference between my 'I'-propositions and yours would lie in the fact that mine imply my *Ich-Vorstellung* and not yours, and that yours imply your *Ich-Vorstellung* and not mine" (1981, 16). We don't need to analyze here the ins and outs of essences. We need not plumb the depths of the question of whether, as a matter of metaphysics, persons have individual essences or haecceities. The point is that whether or not they do, the view that they do suggests a corresponding move in formal logic that serves to help mathematize and mechanize *A*'s reasoning. From the logico-computational viewpoint, the role that individual essences are to play in the production of the above proof sketch is clear: that is the role of allowing, formally, the needed reduction. And Chisholm shows how to carry out the reduction, in chapter 1 of *Person and Object* (1976). The basic trick is perfectly straightforward: *De re* belief is belief that a relevant proposition holds. What proposition? Consider the tallest-man trio of examples given above. Consider, specifically, the situation in which you are the tallest man; you are (as you most assuredly are) wise; and I believe in *de re* fashion of you that you are wise. Then on Chisholm's view I believe a proposition φ which deductively implies that you have both the properties *being the tallest man* and *being wise*. Our φ here is just the proposition that the tallest man is wise.

Of course, there isn't space here to cover the reduction in any detail. Given present purposes, it suffices to note that the reduction requires that each person (and in the case of the machine-generated correlate of *A*'s reasoning as conveyed by Floridi, each computing machine) be associated with an individual essence, in our formal semantics. We can thus say that while *a* is an ordinary constant in the language of the cognitive event calculus, and hence it's entirely possible for *a* to be identical to some other constant (for example, the proper name of prisoner *A*; Alfred, perhaps), *a** is a symbol functioning as a personal pronoun for prisoner *A*. We have then only to amend my proof-sketch by replacing each occurrence of *a* with *a**—and we are done.[14]

[14] How would the details look? The simplest thing to do (and I am of course under no obligation to provide a formal semantics that is complicated; all I need is something that gets the engineering job done) is to give a "syntactic" semantics for **CEC** based simply on sets, directly. On this approach, what an agent *b* knows is simply collected into a set (a box) of formulae, suitably indexed with her name. What she knows she knows is simply collected into a box within her box. This approach, which is classically set out in Genesereth and Nilsson 1987, could be appropriated for present purposes without requiring too much

5.2. *"But A's reasoning is counterfactual reasoning"*

The second objection: "You yourself conceded, Bringsjord, that even the reasoning given to solve KG_3 was counterfactual in nature; and in fact you agreed that the reasoning in KG_4 is—as Floridi claims—counterfactual as well. But your proof-sketch appears to be based solely on the *material* conditional. That conditional is allowed to be within the scope of various epistemic and perceptual operators (e.g., **C**, **K**, **P**, etc.), but this in no way yields a conditional that is counterfactual in nature. Hence it's clear that your proof-sketch fails to point the way to a machine-based version of *A*'s victorious reasoning, as set out by Floridi."

In reply, first note that I didn't say that reasoning that produces a pass in the case of the KG_4 test *must* employ counterfactual conditionals. I grant only that the reasoning Floridi offers on behalf of *A* makes use of natural-language versions of such conditionals.[15] The event calculus, as a matter of mathematical fact, obviates, in certain contexts, the need for counterfactual conditionals. In short, my future KG_4-passing machine agent has no need of such conditionals, because their import is expressed by way of the branching histories that the event calculus secures.[16]

My rebuttal can be viewed from a different perspective, namely, that of a judge in the case of KG_4. Suppose, in fact, that you are in the role of judge and must decide the fate of *A*, based on his response (his answer **A** and justification **J**; recall again figure 1). Now suppose that *A* provides not only "Heaven knows!" and (if **Q** is "Did you ingest a dumbing tablet?") "No!" and a natural-language version of the proof sketch I have supplied. Would not the judge, upon receiving this content from *A*, declare that the test had been passed? I should think so.

imagination, starting with legislation to the effect that every agent *b* has on the semantic side a "haecceity" symbol H_b associated with him, and continuing with the stipulation that first-person beliefs are not only the relevant standard formulae in *b*'s box, but the injection, at the relevant time point, of the string H_b (b) into that box. My intuitive understanding of this string would correspond to what Floridi is quoted as saying above in the underlined key phrase in the quote I gave in section 2.1: Ag's personal identity. Note that Chisholm even specifically says in *Person and Object* (1976) that one's haecceity may consist in the property of being identical to oneself.

[15] In this connection, it's worth nothing that the tradition surrounding KG_1 and its relatives (e.g., the muddy children puzzle) is one in which formal logic-based modeling need not reflect counterfactuals. See, e.g., Fagin et al. 2004.

[16] Though I can't present any of the details here, even if Floridi insisted that a machine-generated and machine-certified proof corresponding to *A*'s success include not $\rightarrow$ but the $>$discussed, e.g., by Nute (1984), the situation could be handled by formalizing the semantics for $>$in extensional first-order logic, and relying after that on standard automated proving power of the sort that allowed figure 2 to be produced. Such a trick is actually essentially one that parallels the one used to reduce inference in **CEC** to inference in standard first-order logic (for details see Arkoudas and Bringsjord 2009).

The final part of my reply is simply to note that, from the standpoint of the field that arguably bears most directly on the challenge facing prisoner *A*, decision theory, object-level counterfactual reasoning is not necessary. To put it starkly, decision theory, even when elaborate and philosophically sophisticated, can be erected in the complete absence of the niceties of conditional logic (see, e.g., Joyce 1999).

6. Conclusion

Floridi succeeds in delivering an inventive, unprecedentedly difficult challenge to logic-based AI; this much I gratefully concede.[17] However, in light of the foregoing, it is seen that this challenge can be met by foreseeable AI. Is there a further variation of the knowledge game beyond the capacity of robots produced by foreseeable AI? Yes, I believe so; and I believe that Floridi may eventually find and disclose such a variation. Is there a variation of the game beyond the capacity of a robot to solve, period? No. The problem is that the knowledge game, as a fair, *empirical* test, is by definition such that there is some finite, observable behavior, β, which the judge, as a perfectly rational agent employing certain principles for decision making, will be correct in judging to be a "passing grade." Since such a β can be generated by a suitably programmed Turing machine (or equivalent) operating over a finite amount time without contravening the laws of logic or even physics, it can immediately be deduced that there exists (in so-called mathematical space) a robot that passes with flying colors. To put the point another way, there is every reason to think that as AI marches on into a future beyond our children, and our children's children, and indeed into time centuries hence, leaving us at best a distant memory, the universe will be populated by computational creatures whose external behavior includes anything and everything within our own repertoire. (It is perhaps ironic, but certainly true, that none other than Floridi himself is among the very few on our planet who have professionally and prudently—and in my opinion prophetically—pondered a future in which the world is saturated with information; see, e.g., Floridi 2007.) These creatures may nonetheless lack s-consciousness and p-consciousness. We might know that they lack these attributes, but only via extensive reasoning showing that s-consciousness or p-consciousness is more than computation (see, e.g., the arguments in Bringsjord and Zenzen 2003). But only God would know *a priori*, because his test would be direct and nonempirical: he would know whether these

[17] I can't rationally declare that his challenge is supremely difficult for *other* forms of AI—say, for "low-level" robotics as opposed to logic-based robotics, as classically defined by Levesque and Lakemeyer (2007). The reason includes, for instance, that the knowledge game simply doesn't require intricate, nonsymbolic, "perception-and-action" engineering. For example, it doesn't demand that robots display human-level physical manipulation.

beings are s- and p-conscious not by following the recipe of figure 1 but by considering whether or not such consciousness is present, end of story—analogous to our ability to settle, quite independent of empirical testing, whether or not, say, 83 is a prime number.

Department of Cognitive Science
Department of Computer Science
Rensselaer AI and Reasoning Laboratory
Rensselaer Polytechnic Institute
Troy, NY 12180
USA
selmer@rpi.edu

Acknowledgments

I'm greatly indebted to Luciano Floridi—not only for an inspiring, ever-expanding (Zeno-like) body of written work (a microscopic part of which I engage here) but also for many stimulating conversations about matters at the intersection of philosophy of mind and computation/information, including, specifically, the matters I discuss in the present essay. Massive thanks are due to Konstantine Arkoudas for work in logic-based AI exploited here. I'm grateful as well to Micah Clark, an anonymous referee, and Patrick Allo for many insightful comments.

References

Arkoudas, K., and S. Bringsjord. 2005. "Metareasoning for Multi-Agent Epistemic Logics." In *Fifth International Conference on Computational Logic in Multi-Agent Systems (CLIMA 2004)*, vol. 3487 of Lecture Notes in Articial Intelligence (LNAI), 111–25. New York: Springer. URL: http://kryten.mm.rpi.edu/arkoudas.bringsjord.clima.crc.pdf.

———. 2009. "Propositional Attitudes and Causation." *International Journal of Software and Informatics* 3, no. 1:47–65. URL: http://kryten.mm.rpi.edu/PRICAI w sequentcalc 041709.pdf.

Block, N. 1995. "On a Confusion About a Function of Consciousness." *Behavioral and Brain Sciences* 18:227–47.

Bringsjord, S. 1995a. "Could, How Could We Tell if, and Why Should—Androids Have Inner Lives?" In *Android Epistemology*, edited by K. Ford, C. Glymour, and P. Hayes, 93–122. Cambridge, Mass.: MIT Press.

———. 1995b. "In Defense of Impenetrable Zombies." *Journal of Consciousness Studies* 2, no. 4:348–51.

———. 1997. "Consciousness by the Lights of Logic and Common Sense." *Behavioral and Brain Sciences* 20, no. 1:227–47.

———. 1999. "The Zombie Attack on the Computational Conception of Mind." *Philosophy and Phenomenological Research* 59, no. 1:41–69.

———. 2007. "Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline." *Journal of Consciousness Studies* 14, no. 7:28–43. URL: http://kryten.mm.rpi.edu/jcsonebil lion2.pdf.

———. 2008a. "Declarative/Logic-Based Cognitive Modeling." In *The Handbook of Computational Psychology*, edited by R. Sun, 127–69. Cambridge: Cambridge University Press. URL: http://kryten.mm. rpi.edu/sb lccm ab-toc 031607.pdf.

———. 2008b. "The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field unto Itself." *Journal of Applied Logic* 6, no. 4:502–25. URL: http://kryten.mm.rpi.edu/SB LAI Manifesto 091808.pdf.

Bringsjord, S., R. Noel, and C. Caporale. 2000. "Animals, Zombanimals, and the Total Turing Test: The Essence of Articial Intelligence." *Journal of Logic, Language, and Information* 9:397–418. URL: http:// kryten.mm.rpi.edu/zombanimals.pdf.

Bringsjord, S., and B. Schimanski. 2003. "What Is Articial Intelligence? Psychometric AI as an Answer." In *Proceedings of the 18th International Joint Conference on Articial Intelligence (IJCAI–03)*, 887–93. San Francisco: Morgan Kaufmann. URL: http://kryten.mm.rpi.edu/ scb.bs.pai.ijcai03.pdf.

Bringsjord, S., J. Taylor, A. Shilliday, M. Clark, and K. Arkoudas. 2008. "Slate: An Argument-Centered Intelligent Assistant to Human Reasoners." In *Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 8), Patras, Greece*, edited by F. Grasso, N. Green, R. Kibble, and C. Reed, 1–10. URL: http:// kryten.mm.rpi.edu/Bringsjord etal Slate cmna crc 061708.pdf.

Bringsjord, S., and M. Zenzen. 2003. *Superminds: People Harness Hypercomputation, and More*. Dordrecht: Kluwer Academic.

Chellas, B. F. 1980. *Modal Logic: An Introduction*. Cambridge: Cambridge University Press.

Chisholm, R. 1976. *Person and Object: A Metaphysical Study*. London: George Allen and Unwin.

———. 1981. *The First Person: An Essay on Reference and Intentionality*. Minneapolis: University of Minnesota Press.

Dretske, F. 2003. "How Do You Know You Are Not a Zombie?" In *Privileged Access and First-Person Authority*, edited by B. Gertler, 1–13. Burlington: Ashgate.

Fagin, R., J. Halpern, Y. Moses, and M. Vardi. 2004. *Reasoning About Knowledge*. Cambridge, Mass.: MIT Press.

Floridi, L. 2005. "Consciousness, Agents and the Knowledge Game." *Minds and Machines* 15, nos. 3–4:415–44. URL: http://www.philosophyonformation.net/publications/pdf/caatkg.pdf.

———. 2007. "A Look into the Future Impact of ICT on Our Lives." *Information Society* 23, no. 1:59–64.

Frege, G. 1956. "The Thought: A Logical Inquiry." *Mind* LXV 289–311.

Genesereth, M., and N. Nilsson. 1987. *Logical Foundations of Articial Intelligence*. Los Altos, Calif.: Morgan Kaufmann.

Hughes, G., and M. Cresswell. 1968. *An Introduction to Modal Logic*. London: Methuen.

Husserl, E. 1970. *Logical Investigations*. London: Routledge and Kegan Paul.

Joyce, J. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Levesque, H., and G. Lakemeyer. 2007. "Cognitive Robotics." In *Handbook of Knowledge Representation*, edited by F. van Harmelen, V. Lifschitz, and B. Porter, 869–86. Amsterdam: Elsevier.

Nute, D. 1984. "Conditional Logic." In *Handbook of Philosophical Logic, vol. 2: Extensions of Classical Logic*, edited by D. Gabbay and F. Guenthner, 387–439. Dordrecht: D. Reidel.

Russell, S., and P. Norvig. 2002. *Articial Intelligence: A Modern Approach*. Upper Saddle River, N.J.: Prentice Hall.

Stickel, M., R. Waldinger, M. Lowry, T. Pressburger, and I. Underwood. 1994. "Deductive Composition of Astronomical Software from Subroutine Libraries." In *Proceedings of the Twelfth International Conference on Automated Deduction (CADE–12), Nancy, France*, 341–55. SNARK can be obtained at the URL provided here. URL: http://www.ai.sri.com/~ stickel/snark.html

Sun, R., and S. Bringsjord. 2009. "Cognitive Systems and Cognitive Architectures." In *The Wiley Encyclopedia of Computer Science and Engineering*, edited by B. W. Wah, 1: 420–28. New York: Wiley. URL: http://kryten.mm.rpi.edu/rs sb wileyency pp.pdf.

Suppes, P. 1972. *Axiomatic Set Theory*. New York: Dover.