The Logicist Manifesto:

At Long Last Let Logic-Based Artificial Intelligence Become a Field Unto Itself[*]

Selmer Bringsjord
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
selmer@rpi.edu

version 9.18.08

Contents

[*] I’m indebted to a number of energetic non-logicists who have debated more primitive versions of this manifesto with me (and who are perhaps starting to see the error of their ways through a dialectic partially reported in the present paper), and to a number of sympathetic-if-not-outright-soulmates for specific, logically astute comments on earlier drafts of this paper. I’m also deeply thankful to three anonymous referees for insightful objections and suggestions; all that they have said has been subjected to sustained reflection on my part, which has in turn led to corresponding text in the paper itself. In addition, were it not for Turner-Eden sessions at CAP conferences dedicated to philosophy of computer science, which weave together logic, technical philosophy, computer science, computational cognitive science, and AI, this manifesto would never have come to be. Finally, I owe a massive debt to my colleague Konstantine Arkoudas, whose vision for and use of logic in AI and Cog Sci exemplifies in large part that which I call for herein.

1 Introduction

This is a premeditatedly polarizing paper — as the title, with help from a little logic, allows you to immediately infer. Starting with the original DARPA-sponsored conference in 1956, AI has been a steadfastly ecumenical field, yet things haven't exactly gone swimmingly; the dreams of Turing, Newell, Simon, McCarthy, Minsky et al. have to this point been repeatedly dashed. The refrain that interaction between logicist and non-logicist approaches is crucial for AI's success, and that breaking the two camps into wholly separate fields would be counter-productive, now rings rather hollow. The time for declaring logicist independence has thus arrived. Let's all face up to the reality that the logic-based approach (McCarthy 1959, McCarthy & Hayes 1969, Bringsjord & Ferrucci 1998*a*, Bringsjord & Ferrucci 1998*b*, Genesereth & Nilsson 1987, Turner 1984, Nilsson 1991, Bringsjord & Ferrucci 2000, Shapiro 2000)[1] to building artificial counterparts to you and me is self-sustaining, self-contained, and too-long trammeled by unstructured, sub-symbolic approaches that now need to simply be left to fend for themselves.

The plan for the sequel is straightforward. Subsequent content is partitioned into three main sections: background material (§2), some of the countless reasons for inaugurating logic-based AI's independence (§3), and objections (followed in each case by a rebuttal; §4). The paper ends with brief concluding remarks on: whether achieving independence, at this particular time, can really be achieved; two daunting challenges facing LAI; and how to in general move forward on the basis of the manifesto here proclaimed.

2 Background

The next section (2.1) explains, in broad strokes, what logic-based AI (LAI) is. Then (§2.2) the reader is reminded of the fundamental difference between "strong" versus "weak" AI, followed by a brief explanation as to why, in the present paper, the distinction is put aside. After that comes section 2.3, in which, leveraging the presentation to this point, some small slices of the life of a LAI-based agent operating at the level of a person are partially formalized.

2.1 Logic-Based AI Encapsulated

LAI is distinguished by three tightly interconnected high-level hallmarks, to wit:

Ambitious LAI is an ambitious enterprise: it aims at building artificial persons.

Logical Systems LAI is specifically based on the formalization of one of the distinguishing features of persons, namely that they are bearers of propositional attitudes (such as *knows*, *believes*, *intends*, etc.), and that persons reason over such attitudes (which are often directed toward the propositional attitudes of other agents). This formalization is achieved via *logical systems*.

Top-Down LAI is a top-down enterprise: it starts by immediately tackling that which is distinctive of persons (e.g., propositional attitudes), without wasting dwelling on the adventitious embodiment of

[1] This paper isn't, and isn't intended to be, a comprehensive survey of AI work that uses logic in some manner. Rather, this paper is a sustained argument for the founding of a self-contained, independent discipline: logic-based AI (as characterized in section §2). An immediate corollary is that this paper isn't straight AI or computer science, but rather *philosophy of* AI and computer science. Accordingly, references to activity in logic-based AI have been selected to serve this argument, and fit the philosophical nature of the present paper. It would take a book-length essay to just briefly discuss and cite the myriad examples of first-rate logicist AI carried out by AI researchers and engineers since the early 1950's. Even in cases where I have been specifically charged by publishers and editors to leave aside communicating to my readers anything fundamentally new, and to simply provide a comprehensive survey of logicist AI (e.g., in the relevant part of my AI entry in the Stanford Encyclopedia of Philosophy), I've frankly found it difficult to cover even the majority of the first-rate minds subscribing to this approach.

cognition in particular physical stuff, or (what will later be called *stage-one*) transduction between the external physical environment and high-level cognitive processes.

These three attributes can be traced back to the very dawn of LAI. For example, they can be seen clear as day in perhaps the very first specimen of LAI: the *Advice Taker* (AT) program from McCarthy and Minsky (McCarthy 1959).[2] For example, with respect to **Ambitious**, McCarthy (1959, p. 3) wrote about AT and the like: "In our opinion, a system which is to evolve intelligence *of human order* should have at least the following features ..." (emphasis mine). As to **Logical Systems** and the centrality of propositional attitudes, McCarthy writes: "We shall therefore say that a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already *knows*" (emphasis mine; McCarthy repeatedly speaks of what AT knows and believes). Thirdly, concerning a **Top-Down** methodology, McCarthy's approach is without question a thoroughgoing top-down one wholly focused upon high-level symbolic representations in propositional form, to the utter exclusion of lower-level perceptual, dynamic, or neurocomputational processes and formalisms.[3]

While it's true that at the dawn of AI in the 1950's, logic was used by many of the seminal figures in the field, at that point 'logic' was identified with only a tiny speck in the space of a field that we now know to be provably larger than classical mathematics;[4] that speck was first-order extensional logic (= part of the predicate calculus) where inference is exclusively deductive. In the AT work (and, for that matter, in McCarthyian work from that point on), the field of logic isn't leveraged in any at once broad and formidable way. To see this, one can simply note that the machine intelligence achieved, or even aimed at, in AT is restricted to only one mode of inference: deduction.[5] But logic is the science of rigorous reasoning, and that reasoning comes in the following additional modes, at the very least: inductive, abductive, defeasible, analogical, and visual. AI work that leverages or at least takes profitable account of all these and other modes is most appropriate for supporting my manifesto. In addition, while McCarthy and likeminded researchers are great friends of logic-based AI, and have made historic contributions, the fact is that mathematical logic is itself only a small part of logic: the part based on the attempt, inaugurated by Frege (as elegantly chronicled in Glymour 1992), to formalize mathematics. Logic as a whole includes systems much more expressive than those used to formalize mathematics. The space of intensional logical systems, for example, grows every month, but none of these systems is designed to model mathematics. Logic-based AI, as I define and defend it herein, takes account of *all* of logic.

We turn now to partially explicating each of the three attributes in turn.

[2] What I say here about the AT paper applies, *mutatis mutandis*, to other seminal papers at the dawn of logicist AI. E.g., my comments would apply to (McCarthy & Hayes 1969). They would also seem to apply to the two seminal logicist papers that singlehandedly brought me into AI: (Hayes 1978, Hayes 1985).

[3] Some might complain that McCarthy and Minsky had no choice, because (so it's here claimed) such formalisms were simply not available. However, such a complaint would be historically inaccurate. E.g., see (Minsky 1967), the neurologically inspired material in which has its roots in Minsky's thinking from the the 50's.

[4] The systematization of mathematics into logic was provided by many decades of formal exposition in books authored by "Bourbaki," a group allonym for the mathematicians who authored a collection of eight painstakingly rigorous, detailed books showing that all the publishable results of classical mathematics can in fact be expressed as derivations from axiomatic set theory using the logical system known as first-order logic, which is $\mathcal{L}_I$ in the family $\mathcal{F}$ of systems referred to later in the present paper. The starting place in the Bourbaki oeuvre is (Bourbaki 2004) — exposition that shows, formally speaking, that discovery and confirmation in mathematics consists, fundamentally, in the derivation and use of theorems all extractable from a small set of axioms (e.g., the Zermelo-Fraenkel axioms for set theory).

[5] Read again the definition of programs with common sense given by McCarthy and Minsky: "We shall therefore say that a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows" (McCarthy 1959, p. 2). Emphasis mine.

2.1.1 LAI is Ambitious

The first paragraph of the present paper, recall, carries the claim that LAI is the superior way to build "artificial counterparts to you and me." But what are we? We are human persons.[6] The humble goal of building mere animals can no doubt be pursued by ignoring both the cognitive powers distinctive of persons, and the only paradigm — logic — suitable for capturing these powers in computation. Telling here is the more than two-decade-old objective announced by Charniak and McDermott (1985) in a text that presented AI as a field based entirely on (first-order) logic:

> The ultimate goal of AI, which we are very far from achieving, is to build a person, or, more humbly, an animal. (Charniak & McDermott 1985, p. 7)

A chimp, despite the efforts of scientists who refused to heed Chomsky's quip that one can no more teach a chimp to talk than teach it to fly, cannot communicate *qua* person, and fails to have a number of additional properties constitutive of persons.[7] Needless to say, everything below a chimp is even dimmer, and so artificial counterparts thereof are not what logicists ultimately aiming to mechanize. Perhaps non-logicists can build an artificial animal somewhere in the spectrum from chimp to earthworm,[8] but they will never secure the loftier goal Charniak and McDermott imagined: intelligence at the level of human persons. We turn now to a more careful account of personhood.

One generic account of human personhood has been proposed, defended, and employed by Bringsjord (1997, 2000). This account is a fairly standard one; for example, it generally coincides with one given by Dennett (1978), and by others as well, for example Chisholm (1978). In addition, this account is in line with the capacities covered, chapter by chapter and topic by topic, in surveys of cognitive psychology (e.g., see Goldstein 2005, Ashcraft 1994). The account in question holds that x is a person provided that x has the *capacity*

1. to "will," to make choices and decisions, set plans and projects — autonomously;
2. for subjective consciousness: for experiencing pain and sorrow and happiness, and a thousand other emotions — love, passion, gratitude, and so on;
3. for *self*-consciousness, for being aware of his/her states of mind, inclinations, preferences, etc., and for grasping the concept of him/herself;
4. to communicate through a language;
5. to know things and believe things, and to believe things about what others believe (second-order beliefs), and to believe things about what others believe about one's beliefs (third-order beliefs), and so on;
6. to desire not only particular objects and events, but also changes in his or her character;
7. to reason (for example, in the fashion exhibited in the writing and reading/studying of this very paper).

Given this list, LAI is seen to be the field devoted to capturing these seven capacities simultaneously in a computationally implemented logical system. This position on the ultimate objective of LAI meshes seamlessly with a recent account of what human-level computational cognitive science is shooting for given by Anderson & Lebiere (2003), who, instead of defining personhood, give an operational equivalent of this definition by describing "Newell's Program," an attempt to build

[6] Please note that *human person* isn't a redundant phrase. This is so because those attributes that jointly define what it is to be a person (e.g., see the list of attributes given below) may well apply to creatures who aren't members of the species *homo sapiens sapiens*. Our entertainments, particularly in the science fiction genre, illustrate this routinely and repeatedly.

[7] And so apparently by Turing's (1950) empiricist metric, a chimp would not be classified as a thinking thing.

[8] A spectrum described in (Bringsjord et al 2000).

computational simulations of human-level intelligence, where that intelligence is cashed out in the form of a list of abilities that correspond to those on the list just given. For example, part of Newell's Program is to build a computational simulation of natural-language communication at the normal, adult level. This is attribute 4 on the list above. In the present paper, as the reader by now realizes, the emphasis is on attributes 5 and 7.

As Aristotle noted rather long ago, we are rational. Put in terms of the definition of personhood given above, we know things, we believe things, and we reason over this content, often after we have formalized it in the form of a logical system, the first of which — the theory of the syllogism — he himself introduced 300 years BC, long, long before such schemes as artificial neural networks arrived on the scene. We turn now to a closer look at logical systems.

2.1.2 LAI is Based on Logical Systems

As noted, the cognitive level distinguishes persons. And as also noted, at the center of cognition stand propositional attitudes: knowing, believing, intending, and so on. In light of the fact that if an agent knows (believes, intends to bring about, etc.) ϕ, ϕ must be a proposition (a declarative statement having a semantic value), the basic units of LAI are formal objects associated with those particular sentences or expressions in natural languages (like English, German, Chinese) that are declarative statements (as opposed to expressions in the imperative or inquisitive mode) conveying propositional content, and taking values such as TRUE, FALSE, UNKNOWN, PROBABLE (sometimes to particular numerical degrees), and so on. The basic process over such units is inference, which may, as noted above, be deductive, inductive, abductive, analogical, etc. Because the basic units of LAI are propositions, and the basic processes over these units are forms of reasoning, the foundation of LAI is the infinite class of what can be called *logical systems*.[9]

Logical systems are used to both model human cognitive powers (which gives rise to declarative computational cognitive modeling, described in Bringsjord 2008), and to enable a computing machine to acquire some of that power (which gives rise to LAI). Put very simply, a logical system $\mathcal{L}$ is composed of six parameterized elements, as follows.

1. An alphabet A, partitioned into those symbols that are invariant across the use of $\mathcal{L}$ for any application area (e.g., mathematical symbols), and those that are included by the human for formalize particular domains (e.g., `Loves` as a predicate symbol standing for the dyadic property of one thing loving another).
2. A grammar $\mathcal{G}$ that yields well-formed expressions (*formulas*) from the alphabet A.
3. An argument theory $\vdash_X^M$ (called a *proof* theory when the reasoning in question is deductive in nature) that specifies correct (relative to the system $\mathcal{L}$) inference from one or more expressions to one or more expressions. The superscript is a placeholder for the *mode* of inference: deductive, abductive, inductive, probabilistic, analogical, etc. The subscript is a placeholder for *particular* inferential mechanisms.
4. An *argument semantics* that specifies the meaning of inferences allowed by $\vdash_X^M$, which makes possible a mechanical verification of the correctness of arguments.
5. A *formula semantics* that assigns a meaning of formulas given announcements about what the application symbols are. The values traditionally include such things as TRUE, FALSE, INDETERMINATE, PROBABLE, numbers in some continuum (e.g., 0 to 1, as in the case of probability theory), and so on.
6. A metatheory that defines meta-mathematical attributes over the previous five components, and includes proofs that the attributes are or are not possessed. Examples of such attributes include soundness, completeness, decidability, etc.

[9] Some refer simply to 'logics,' but that is inaccurate for reasons that needn't detain us.

The family $\mathcal{F}$ of logical systems populated by the setting of parameters in the sextet just given is (of course) infinite, and includes zero-, first-, and higher-order extensional logics (in Hilbert style, or sequent style, or natural deduction Fitch style, etc.); modal logics (including temporal, epistemic, deontic logics, etc.); propositional dynamic logics; Hoare-Floyd logics for reasoning about imperative programs; inductive logics that subsume probability theory in the Bayesian tradition; abductive logics; strength-factor-based and probabilistic logics; non-monotonic logics, and many, many others. All of classical mathematics is derivable from just a few axioms (e.g., the Zermelo-Fraenkel axioms) in a simple speck (viz., standard first-order extensional logic) within $\mathcal{F}$ (Ebbinghaus et al. 1994). For a sustained exposition of $\mathcal{F}$, along with presentation of some examples within it given in connection with person-level performance on certain problems, see (Bringsjord 2008). When, later (§2.3), I consider a slice in the life of a LAI agent, I present the rudiments of some specific (elementary) logical systems.

2.1.3 LAI is a Top-Down Enterprise

LAI is a top-down, rather than bottom-up, approach. As Brachman & Levesque (2004) put it:

> [LAI] is at the very core of a radical idea about how to understand intelligence: instead of trying to understand or build brains from the *bottom up*, we try to understand or build intelligent behavior from the *top down*. In particular, we ask what an agent would need to know in order to behave intelligently, and what computational mechanisms could allow this knowledge to be made available to the agent as required. (Brachman & Levesque 2004, p. iv)

As reflected in relevant formalisms commonly associated with bottom-up approaches (e.g., artificial neural networks), the basic units in bottom-up processing are numerical, not declarative. If you look back to the list of attributes taken to be constitutive of personhood, you can see that that list is indeed cognitively, not physiologically, oriented, and that numbers don't seem particularly relevant.

Clearly, given the goal of mechanizing personhood, no progress is made by merely noting the particular DNA structure of humans. When it is said that x is human just in case x has a particular genetic code, the perspective is not that of LAI. Our minds are not modeled by charting the physiology of our brains. (After all, AI is committed to the dogma that implementation be produced in silicon-based substrates, not carbon-based ones.) Rather, logicists are asking what it means to be a human being from the *cognitive*, perspective. That is, the question is: What does it mean to be a human *person*? This driving question clearly reflects a top-down perspective.

2.2 Ignoring the "Strong" vs. "Weak" Distinction

"Weak" AI can be conveniently defined as the field devoted to engineering intelligent agents able to pass the Turing Test and various other tests for overt, outwardly observable behavior (for a sequence of such tests, starting with the Turing Test, see Bringsjord 1995*a*). Put differently, weak AI aims at building machines that *act* intelligently, without taking a position on whether or not the machines actually *are* intelligent. "Strong" AI is an ambitious form of the field aptly summed up by Haugeland:

> The fundamental goal [of AI research] is not merely to mimic intelligence or produce some clever fake. Not at all. AI wants only the genuine article: machines with minds, in the full and literal sense. This is not science fiction, but real science, based on a theoretical conception as deep as it is daring: namely, we are, at root, computers ourselves. (Haugeland 1985, p. 2)

While there are a number of prominent thinkers who have argued forcefully that minds can be computers (e.g., see Rapaport 1998, Dietrich 1990, Shapiro 1995), this paper leaves aside the distinction between "strong" and "weak" AI, and hence leaves aside such arguments.[10] To ease exposition, and focus on the main issue at hand (viz., whether or not logic-based AI should come out from under the ecumenical umbrella that currently covers the field and become a discipline unto itself), the distinction between an information-processing machine that *is* is a person, versus one that merely *simulates* people is ignored as beside the point at issue. For a sustained treatment of the strong versus weak issue, and related matters, see (Bringsjord & Arkoudas forthcoming).

2.3 A Slice in the Day of a Life of a LAI Agent

A LAI agent is first and foremost a system that, through time, adopts and manages certain attitudes toward propositions, and reasons over these propositions, in order to perform the actions that will secure certain desired ends. The most important attitudes, by far, are ones we've already mentioned: *believes that* and *knows that*; our focus in this brief look at a LAI agent will be on the latter. (Other propositional attitudes, recall, include *wants that* and *hopes that*. A propositional attitude is simply a relationship holding between an agent, and one or more propositions, where propositions are declarative statements.) A LAI agent can thus here be viewed as a system whose knowledge changes in tune with what is learned from the environment, and from reasoning over this knowledge, where this reasoning must be *surveyable*. Reasoning is surveyable when it is laid out explicitly in the form of either a proof or argument, with all inferences unambiguously presented.

That the reasoning in the case of logicist cognitive systems be surveyable is a crucial condition, one worth expanding upon: While logic, from the dawn of systematic human thought, has been regarded the science of reasoning, some today use the term 'reasoning' in heterodox fashion, using such phrases as "Bayesian reasoning" to refer to probabilistic *calculation*. Calculation is not surveyable reasoning. The only proofs allowable in mathematics are surveyable proofs: step-by-step chains of inference, each and every link formed in conformity with rules of logic having nothing to do with numerical calculation (Arkoudas & Bringsjord 2007). (We shall employ some of these rules in two illustrative proofs soon to be crafted.) The reasoning in such proofs is one of the cornerstones of logicist agents.

In addition, a LAI agent aims at certain goals by changing its environment. In sum, then, we can conceive of such an agent's life as a cycle of sensing, reasoning, acting; sensing, reasoning, acting; .... An immortal LAI agent would be one in which this cycle repeats *ad infinitum*, presumably with goal after goal achieved (and known to be achieved) along the way. The propositions at the heart of this cycle are represented as formulas in one or more logical systems, and the reasoning in question is also regimented by the relevant parameter in a logical system (recall the account of such systems in §2.1.2).

To cultivate a bit of a first-hand feel for LAI agents, suppose that you are one, and that you learn from the environment (somehow; the details needn't detain us) that

- Alvin loves Bill; and that
- Everyone loves anyone who loves someone.[11]

You have now acquired knowledge from the environment. Your goal, let us assume, is to determine whether or not everyone loves Bill, and whether or not Katherine loves Dave. However the answers

[10] In at least the case of Shapiro and Rapaport, logic has played an absolutely pivotal role in building systems able to display human-level cognitive behavior. E.g., see their long-established work on the basis of the SNePS system, which is linked with relevance logic (Shapiro 2000).

[11] Bringsjord is indebted to Phil Johnson-Laird for this challenge.

turn out, the reasoning in support of your new knowledge must be provided in the form of an explicit series of inferences (which serves to guarantee that the reasoning in question is surveyable). Since you are yourself a LAI agent, you should be able to reach these goals. Can you? The best way to truly understand what a LAI agent is is to try to answer this question, and ones like it. You are encouraged to carry out the necessary reasoning now. We shall return to the question a bit later.

I now give a more detailed account of the sense-reasoning-act cycle of a LAI agent, described in broad strokes earlier.

At any time t during its existence, the cognitive state of a LAI agent S consists in what the agent knows at that time, denoted by Φ_S^t. (To ease exposition, we leave aside the distinction between what S knows, versus what it merely believes. You believe a lot more than you know, for the simple reason that you believe some propositions that are false, and one cannot possibly know a false proposition.) We assume that as S moves through time, what it knows at any moment is determined, in general, by two sources: information coming directly from the external environment in which S lives, through the transducers in S's sensors that turn raw sense data into propositional content; and from reasoning carried out by S over its knowledge. For an example related to the first source, recall the example we introduced earlier in the paper. Given what we said there, your knowledge (or as it is often said, your *knowledge base*) includes that Alvin loves Bill. (It also includes 'Everyone loves anyone who loves someone'. This will soon become a crucial piece of knowledge.) You know this because information impinging upon your sensors has been transduced into propositional content added to your knowledge base. (In the case at hand, you read text appearing earlier in this paper.) Suppose that at t_n, some moment before you read the start of this paper, you didn't have this knowledge. We can summarize the situation at this point as follows. (You will note that we have represented 'Alvin loves Bill' in a certain way. As will be explained momentarily, this representation is a formula in first-order logic.)

$$\Phi_S^{t_{n+1}} = \Phi_S^{t_n} \cup \{\texttt{Loves(alvin,bill)}\}$$

Generalizing, we can define a binary function *env* from timepoint-indexed knowledge bases, and formulas generated by *trans* applied to raw information hitting sensors, to a new, augmented knowledge base at the next timepoint. So we have:

$$\Phi_S^{t_{n+1}} = env(\Phi_S^{t_n}, trans(raw))$$

where $trans(raw) = \texttt{Loves(alvin,bill)}$.

Now let's consider the second source of new knowledge, in connection with the simple example we have introduced: viz., reasoning. You know many other propositions on the basis of reasoning over the proposition that Alvin loves Bill: you know that someone loves Bill, that someone loves someone, that someone whose name starts with 'A' loves Bill, and so on. These additional propositions can be directly deduced from the single one about Alvin and Bill; each of them can be safely added to your knowledge base.

Let $\mathcal{R}[\Phi]$ denote a modification of Φ via some mode of reasoning $\mathcal{R}$. Then your knowledge at the next timepoint, t_{n+2}, is given by

$$\Phi_S^{t_{n+2}} = \mathcal{R}[env(\Phi_S^{t_n}, trans(raw))]$$

As time flows on, the environment's updating, followed by reasoning, followed by changes the cognitive system makes to the environment (the system's actions), define the cognitive life of S. This is the three-part cycle we introduced at the outset of the paper.

But what is $\mathcal{R}$, and what is the structure of propositions returned by *trans* and composing the knowledge base? This is where logic enters the stage. Knowledge, that is, propositions, are represented by formulas in one or more logical systems, and these systems provide precise machinery for carrying out reasoning. The simplest logical systems that have provided sufficient expressivity to allow engineers to build LAI agents that are at least somewhat impressive are the propositional calculus, and the predicate calculus (or what we have called above first-order logic, or just FOL); together, this pair comprises what is generally called elementary logic. I proceed now to give a very short review of how knowledge is represented and reasoned over in these systems, after which we can return to trying to meet the goals in the challenge presented earlier (viz., determine whether everyone loves Bill, and whether Katherine loves Dave).

2.3.1 Knowledge Representation in Elementary Logic

Every introductory AI textbook provides an introduction to these logical systems, and makes it clear how they are used to engineer intelligent systems (e.g., see Russell & Norvig 2002). In the case of both of these systems, and indeed in general when it comes to any logical system, three overarching components (which are selected from the six parameters set out in section 2.1.2) are required: one is purely syntactic, one is semantic, and one is metatheoretical in nature. The syntactic component includes specification of the alphabet of a given logical system, the grammar for building well-formed formulas (wffs) from this alphabet, and, more importantly, an argument theory that precisely describes how and when one formula can be inferred from a set of formulas. (These are the first three parameters in §2.1.2.) The semantic component includes a precise account of the conditions under which a formula in a given system is true or false. The metatheoretical component includes theorems, conjectures, and hypotheses concerning the syntactic component, the semantic component, and connections between them. In the following, I focus on the syntactic side of things. Thorough but refreshingly economical coverage of the formal semantics and metatheory of elementary logic can be found in (Ebbinghaus, Flum & Thomas 1994).

As to the alphabet for propositional logic, it's simply an infinite list

$$p_1, p_2, \ldots, p_n, p_{n+1}, \ldots$$

of propositional variables (according to tradition p_1 is p, p_2 is q, and p_3 is r), and the five familiar truth-functional connectives $\neg, \rightarrow, \leftrightarrow, \wedge, \vee$. The connectives can at least provisionally be read, respectively, as 'not,' 'implies' (or 'if then '), 'if and only if,' 'and,' and 'or.' Given this alphabet, we can construct formulas that carry a considerable amount of information. For example, to say that 'if Alvin loves Bill, then Bill loves Alvin, and so does Katherine' we could write

$$a_l \rightarrow (b_l \wedge k_l)$$

where the propositional variables, as you can see, are each used to represent declarative statements.

We move up to first-order logic when we allow the quantifiers $\exists x$ ('there exists at least one thing x such that ...') and $\forall x$ ('for all x ...'); the first is known as the *existential* quantifier, and the second as the *universal*. We also allow a supply of variables, constants, relations, and function symbols. Using this machinery, the proposition that 'Everyone loves anyone who loves someone' is represented as

$$\forall x \forall y (\exists z Loves(y, z) \rightarrow Loves(x, y))$$

2.3.2 Deductive Reasoning

Most readers will be familiar with the concept of deductive reasoning, the hallmark of which is that if the premises are true, then that which is deduced from them must be true as well. If in fact it's true that Alvin loves Bill (and we are given that it is), nothing is more certain than that someone loves Bill. In logic, deduction is formalized in what is called a *proof theory.*

A number of proof theories are possible, for either of the propositional or predicate calculi. When reasoning is to be understood by humans (whether that reasoning is carried out by a human or a machine), it is almost universally agreed that the proof theory of choice is *natural deduction*, not resolution. (This is not to say that logicists shouldn't use resolution-based systems. The nature of the engineering challenge at hand must be allowed to dictate selection of one's logical system.) The latter approach to reasoning (whose one and only rule of inference, in the end, is that from $\phi \vee \psi$ and $\neg\phi$ one can infer ψ), while used by a number of automated theorem provers (e.g., Otter, which, along with resolution, is presented in Wos et al. 1992), is generally impenetrable to human beings (save for those few who, by profession, generate and inspect resolution-based proofs). On the other hand, professional human reasoners (mathematicians, logicians, technical philosophers, etc.) invariably reason in no small part by making suppositions, and by discharging these suppositions when the appropriate time comes. Suppositional reasoning is at the heart of natural deduction. For example, one such common suppositional technique is to assume the opposite of what one wishes to establish, to show that from this assumption some contradiction (i.e., an absurdity) follows, and to then conclude that the assumption must be false. The technique in question is known as *reductio ad absurdum*, or indirect proof, or proof by contradiction. Another natural rule is that to establish that some conditional of the form $\phi \rightarrow \psi$ (where ϕ and ψ are any formulas in a logic L), it suffices to suppose ϕ and derive ψ based on this supposition. With this derivation accomplished, the supposition can be discharged, and the conditional $\phi \rightarrow \psi$ established. For an introduction to natural deduction, replete with proof-construction and proof-checking software, see (Barwise & Etchemendy 1999)

What follows is a natural deduction-style proof (using the two rules just described) written in the proof construction environment known as NDL, used at my university for teaching formal logic *qua* programming language.[12] It is a very simple proof of a theorem in the propositional calculus — a theorem that Newell and Simon's Logic Theorist, to great fanfare, was able to muster at the dawn of AI in 1956, at the original Dartmouth AI conference. Readers will note its natural structure. The rule *modus ponens*, the most fundamental rule in the formal sciences, allows one to infer ψ when one knows two things: that if ϕ then ψ, and ϕ.

```
// Logic Theorist's claim to fame (reductio):
// (p ==> q) ==> (~q ==> ~p)

Relations p:0, q:0. // this is the signature in this case;
                    // propositional variables are 0-ary
                    // relations

assume p ==> q
   assume ~q
      suppose-absurd p
          begin
            modus-ponens p ==> q, p;
            absurd q, ~q
          end
```

This style of discovering and confirming a proof parallels what happens in computer programming.

[12] NDL was invented by Konstantine Arkoudas. To acquire the system for your own local use, got to http://www.cogsci.rpi.edu/courses/intrologic/materials.php

You can view the proof immediately above as a program. If, upon evaluation, the desired theorem is produced, we have succeeded. In the present case, sure enough, we receive this back from NDL when the code is executed:[13]

```
    Theorem: (p ==> q) ==> (~q ==> ~p)
```

The example just given, note, relies on human ingenuity: the machine in this case doesn't find for itself that the theorem holds. But a LAI agent that does that is easy to build, and there is no need to insist that such an agent use natural deduction. For example, a few minutes ago I built such an agent by using the Otter (Wos, Overbeek, e. Lusk & Boyle 1992) automated theorem prover; the agent responds to a query as to whether the formula in question is a theorem by finding and yielding as output this proof:

```
---------------- PROOF ----------------
1 [] -p|q.
2 [] -q.
3 [] p.
4 [hyper,3,1] q.
5 [binary,4.1,2.1] $F.
------------ end of proof -------------
```

Notice that this proof is surveyable. Lines 1–3 are the clausification of the negated conditional in question; line 4 is the result of applying the rule of hyper-resolution to lines 3 and 1 (which in a nutshell is to allow `p` and `-p` to "cancel" each other out, leaving just `q` alone. Line 5, the final step, can be seen to result from applying binary resolution to lines 4 and 2: That is, `q` in line 4 and `-q` in line 2 cancel each other out, leaving nothing, that is, leaving the empty clause (`$F`), which indicates the sought-for contradiction has been found. The upshot is that surveyability is an attribute that proofs have or lack independent of the type of proof calculus (e.g., natural versus resolution-based) used. The essential thing for surveyability is that some proof calculus is employed.

The previous reasoning is in the propositional calculus. What about the predicate calculus, that is, first-order logic? Recall that you were challenged to determine whether or not everyone loves Bill, given that Alvin loves Bill, and that everyone loves anyone who loves someone. How did you fare on that challenge? Well, in fact, it can be rather easily proved that everyone loves Bill. The following program/proof shows this (and along the way shows that Bill loves Alvin, a rather crucial intermediate conclusion). The comments (text coming after '`//`') should provide all the guidance you might need to follow the reasoning.

```
Constants alvin, bill.  // We declare two constants.

Relations Loves:2. // This concludes our simple signature, which
                   // declares Loves to be a two-place relation.
```

[13] Note that the proof here is unquestionably surveyable. This is after all precisely why the computer is able to verify the proof, and add the theorem to the knowledge base. Each inference can be independently inspected and certified; and since each inference rule in NDL yields some output, there is the composite output, viz., the theorem. By contrast, consider a computer program P which, upon receiving some formula ϕ in some logical system, simply returns YES, and nothing more. In this case there is no surveyable proof provided: there is no explicit chain of deductively valid inferences that can be inspected, and certified. More realistically, suppose that ϕ is of the form $\forall x Rx$, where the domain of quantification is the natural numbers $\mathcal{N}$. And suppose that P returns as justification for its affirmative response that $R1, R2, R3, \ldots, Rn$, for some $n \in \mathcal{N}$. In this case, once again, there is no surveyable proof provided, because there is no rule of inference that sanctions the inference to $\forall x Rx$ from $R1 \wedge R2 \wedge \ldots \wedge Rn$.

```
// We add 'Alvin loves Bill' to the knowledge base:
assert Loves(alvin, bill).

// We add 'Everyone loves anyone who loves someone' to the knowledge base:
assert (forall x (forall y ((exists z (Loves(y, z))) ==> (Loves(x, y))))).

// Now we write the program in earnest:

// We begin by adding to the knowledge based that someone loves Alvin:
ex-generalize (exists z (Loves(alvin,z))) from bill;

// Next, we substitute 'bill' for x, and put the result into the knowledge base:
specialize
    (forall x (forall y ((exists z (Loves(y, z))) ==> (Loves(x, y))))) with bill;

// And now we substitute 'alvin' for y, and put the result into the knowledge base:
specialize (forall y ((exists z (Loves(y, z))) ==> (Loves(bill, y)))) with alvin;
        // We now have (exists z (Loves(alvin, z))) ==> Loves(bill, alvin)
        // in the knowledge base.

modus-ponens (exists z (Loves(alvin,z))) ==> Loves(bill,alvin),
    (exists z (Loves(alvin,z)))
        // At this point we have 'Bill loves Alvin' in the knowledge base.
        // But we know that everyone loves anyone who loves someone -- and
        // Bill loves someone!  Let's finish the proof, showing that everyone
        // loves Bill:

// At this point we announce that x is a fresh, arbitrary variable, and proceed to
// prove that everyone loves Bill.
pick-any x
  begin
    ex-generalize (exists z (Loves(bill, z))) from alvin;
    specialize (forall x (forall y ((exists z (Loves(y, z))) ==> (Loves(x, y))))) with x;
    specialize (forall y ((exists z (Loves(y, z))) ==> (Loves(x, y)))) with bill;
    modus-ponens ((exists z (Loves(bill, z))) ==> (Loves(x, bill))),
      (exists z (Loves(bill, z)))
  end
```

When this program is run, we receive back precisely what we desire: two theorems are announced as having been now added to the knowledge base:

```
Theorem: Loves(bill,alvin)

Theorem: forall x Loves(x,bill)
```

The second of these, of course, says in FOL that everyone loves Bill. Once again, please note that LAI agents able to make such discoveries on their own are easy enough to engineer. Here, for example, is a proof divined by such an agent armed not with Otter, but rather with SNARK, a resolution-based ATP written in Common Lisp by Mark Stickel.[14]

```
(Refutation
(Row 1
   (loves alvin bill)
   assertion)
(Row 2
   (or (not (loves ?x ?y)) (loves ?z ?x))
   assertion)
(Row 3
```

[14] For SNARK, and information about the system, please go to http://www.ai.sri.com/∼stickel/snark.html. The SNARK proof shown here was obtained via programming carried out by Joshua Taylor.

```
   (not (loves skolembrjz1 bill))
   negated_conjecture)
(Row 4
   (loves ?x alvin)
   (resolve 2 1))
(Row 5
   (not (loves bill ?x))
   (resolve 3 2))
(Row 7
   false
   (resolve 5 4)))
```

You were also challenged to determine whether Katherine loves Dave, given the two starting pieces of knowledge (again: Alvin loves Bill, and everyone loves anyone who loves someone). In this case, I don't provide the answer. You now have quite enough machinery to settle the issue. Feel free to email the author if you want your answer assessed.

2.3.3 A Note on Nonmonotonic Reasoning

Deductive reasoning is monotonic. That is to say, if ϕ can be deduced from some knowledge base Φ of formulas (written, using notation introduced earlier, $\Phi \vdash^{Deduction}_{ND} \phi$, where '$ND$' indicates that the proof theory is of the natural deduction variety), then for any formula $\psi \notin \Phi$, it remains true that $\Phi \cup \{\psi\} \vdash^{Deduction}_{ND} \phi$. In other words, when $\mathcal{R}$ is deductive in nature, new knowledge never invalidates prior reasoning. This is not how real life works, at least when it comes to humans; this is easy to see. At present, I (= Bringsjord) know that my house is still standing (as I'm sitting in it, typing this sentence). But if, later in the day, while away from my home and working at RPI, I learn that a vicious tornado passed over the Hudson River, over RPI, moved a bit further east toward the Taconic Mountains, and touched down in the town of Brunswick, where my house is located, I have new information that probably leads me to at least suspend judgment as to whether or not my house still stands. Or to take a slight variant of the much-used example from AI, if I know that Pete is a bird, I will probably deduce that Pete can fly, on the strength of a general principle saying that birds can fly. But if I learn that Pete is a penguin, the situation must be revised: that Pete can fly should now not be in my knowledge base. Nonmonotonic reasoning is the form of reasoning designed to model, formally, this kind of — as we can say — *defeasible* inference.

There are many different logical systems that have been designed to model defeasible reasoning — default logic, circumscription, argument-based defeasible reasoning, and so on. (The *locus classicus* of a survey can be found in (Genesereth & Nilsson 1987). An excellent survey is also provided in the Stanford Encyclopedia of Philosophy.[15]) In the limited space available to me in the present paper, perhaps the wisest course is to briefly explain one of these approaches. I select argument-based defeasible reasoning, because it seems to accord best with what humans actually do as they adjust their knowledge through time.

Let us return to the tornado example. What is the argument that Bringsjord might give to support his belief that his house still stands, while he sits within it, typing? There are many possibilities, one respectable one is what I call 'Argument 1':

(1) I perceive that my house is still standing.
(2) If I perceive ϕ, ϕ holds.
∴ (3) My house is still standing.

[15] At
http://plato.stanford.edu/entries/logic-ai

The second premise is a principle that seems a bit risky, perhaps. No doubt there should be some caveats included within it: that when the perception in question occurs, I'm not under the influence of drugs, not insane, and so on. But to ease exposition, let's leave aside such clauses. So, on the strength of this argument, we assume that my knowledge base includes (3), at time t_1.

Later on, as we have said, I find myself working in my office, at RPI, overlooking the Hudson River. A tornado passes over my building. I quickly query my browser once the roar and rumble dies down, and learn from the National Weather Service this very same tornado has touched down due east of RPI, somewhere in Brunswick, and devastating damage to some homes has come to pass. At this point (t_2, assume), if I were pressed to articulate my current position on (3), and my reasoning for that position, and I had sufficient time and patience to comply, I might offer something like this (Argument 2):

(4) A tornado has just (i.e., at some time between t_1 and t_2) touched down in Brunswick, and destroyed some houses there.

(5) My house is located in Brunswick.

(6) I have no evidence that my house was *not* struck to smithereens by a tornado that recently passed through the town in which my house is located.

(7) If a tornado has just destroyed some houses in town T, and house h is located in T, and one has no evidence that h is not among the houses destroyed by the tornado, then one ought not to believe that h wasn't destroyed.

$\therefore$ (8) I ought not to believe that my house is still standing. (I.e., I ought not to believe (3).)

Assuming that I meet all of my "epistemic obligations" (in other words, assuming that I'm rational), I will not believe (3) at t_2. Therefore, at this time, (3) will not be in my knowledge base. (If a LAI agent s doesn't believe ϕ, it follows immediately that s doesn't know ϕ.)

The challenge is to devise formalisms and mechanisms that model this kind of mental activity through time. The argument-based approach to nonmonotonic reasoning does this. While the details of the approach must be left to outside reading (see Pollock 1992), it should be easy enough to see that the main point is to allow one argument to shoot down another (and one argument to shoot down an argument that shoots down an argument, which revives the original, etc.), and to keep a running tab on which propositions should be believed at any particular time. Argument 2 above rather obviously shoots down Argument 1; this is the situation at t_2. Should I then learn that only two houses in Brunswick were leveled, and that they are both located on a street other than my own, Argument 2 would be defeated by a third argument, because this third argument would overthrow (6). With Argument 2 defeated, (3) would be reinstated, and back in my knowledge base. Notice that this ebb and flow in argument-versus-argument activity is far more than just straight deductive reasoning.

2.3.4 Beyond Elementary Logical Systems

So far, we have only discussed three logical systems from the infinite space $\mathcal{F}$. (Given this, and given that you now know it, you are prepared to correct anyone who identifies logic-based AI with research that is based just on elementary logic.) In general, you can partition them into two sub-spaces: those appropriate for formalizing purely mathematical concepts and relationships, and those appropriate for formalizing concepts that have psychological dimensions (such as the propositional attitudes that have been central to our discussion). The former category is referred to as mathematical logic; the latter as philosophical logic. Philosophical logic has proved to be especially useful in AI. Mathematical logic can be viewed as the foundation and circulatory system for computer science (Halpern, Harper, Immerman, Kolaitis, Vardi & Vianu 2001), a point I return to below.

The Alvin-Bill problem given above requires only elementary logical systems be active in the "mind" of a LAI agent. Allow me to quickly present a problem that requires moving to more advanced systems. I present the so-called Wise Man Puzzle (WMP):

> Suppose there are three blindfolded wise men who are told by their king that he is going to select a fez for each of them, and place it on their heads. Each fez will be selected, he informs them, from a collection of five fezes, three of which are white, and two of which are black. The king puts a white fez atop each of the three heads, and then he removes the blindfolds. We assume that each wise man can see the others' hats but not his own, and thus each knows whether the others have white fezes. Suppose we are told that the first wise man says, "I do not know whether I have a white fez," and that the second wise man then says, "I also do not know whether I have a white fez." Now we would like to ask you to attempt to answer the following questions:
>
> **(1)** Does the third wise man now know whether or not he has a white fez?
>
> **(2)** If so, what does he know, that he has one or doesn't have one?
>
> **(3)** And, if so, that is, if the third wise man does know one way or the other, provide a detailed account (showing all work, all notes, etc.; use scrap paper as necessary) of the reasoning that produces his knowledge.

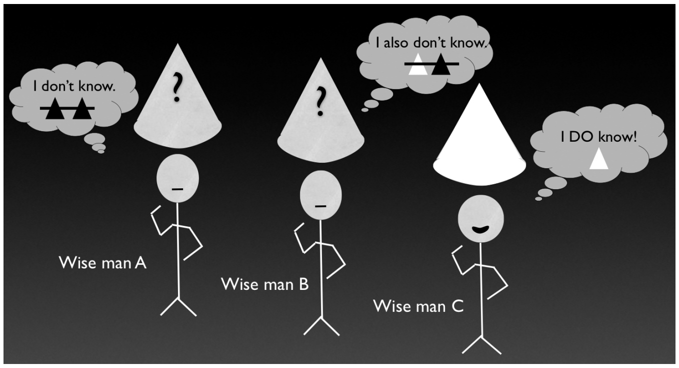


Figure 1: Graphical Summary of Solution to WMP$_3$

The logic that allows a LAI agent to answer these questions is a (modal) propositional epistemic logic; we refer to it simply as $\mathcal{L}_{KT}$. This logic is produced by adding to the propositional calculus the modal operators **B** (for *believes*) and **K** (for *knows*). To see how this can be all be implemented so as to produce a logicist cognitive system, running in real time, that solves the problem in question, see (Arkoudas & Bringsjord 2005). We give here only the intuitive idea behind the relevant proof (expressed as if being uttered by Wise Man C to the king), which is encapsulated in Figure 1:

> "Dear King, I know that I have a white fez! I know this on the strength of the following deduction. First, once Wise Man A spoke, and confessed his ignorance, I deduced immediately that the two heads he was looking at couldn't possibly both have black fezes — for had this been the case, A would have immediately declared that he knew he had a white fez. (There weren't enough black fezes to go around, remember.) Second, once Wise Man B spoke, and admitted *his* ignorance, I was immediately able to rule out the possibility of a white fez atop his head, combined with a black one atop my own. After all, had he seen a black one atop

> mine, he would have immediately proclaimed that he had a white one atop his own, since the mine-black-and-his-black option had already been ruled out. This leaves only a situation where I have a white fez, and that, O king, is what I have!"

Alert readers will have realized that this kind of reasoning works in the general case, that is, it holds for any n wise men and corresponding $n+2$ (suitably color-partitioned) fezes. However, proving that the result holds in the general case requires an even more sophisticated logical system; see (Arkoudas & Bringsjord 2005) for details.

2.4 Examples of Logic-Based Cognitive Systems

There is of course insufficient space to put on display a LAI agent of considerable size. But I have put on display a perfectly respectable, non-trivial example of such an agent in the foregoing (2.3): namely, you — if you have studied and followed the discussion. Specifically, recall that we left off noting that your knowledge at t_{n+2} is

$$\Phi_S^{t_{n+2}} = \mathcal{R}[env(\Phi_S^{t_n}, trans(raw))]$$

Knowing what we now know after further investigation, we can let $\mathcal{R}$ be deductive reasoning, and we can draw upon what the proof above disclosed. This let's us set

$$\Phi_S^{t_{n+2}} = \mathcal{R}[env(\Phi_S^{t_n}, trans(raw))] =$$

$$\Phi_S^{t_{n+1}} \cup \{\texttt{Loves(bill,alvin), (forall x Loves(x,bill)}\}$$

as a summary of your progress on the problem in question as a LAI agent. You will also remember that we asked you to consider whether Katherine loves Dave, given your knowledge base to this point. Were we to inform you that the answer is "Yes," then it would be information directly from the environment that would settle the issue for you. However, we have left it up to you; this means that it's *reasoning* that will settle the issue, not information coming directly to you from the external environment.

3 Factors Supporting Logicist AI as an Independent Field

There are myriad reasons for the independence I recommend. Here are six of them.

3.1 History Supports the Divorce

As is well-known, in 1956, Logic Theorist was demonstrated at the original Dartmouth AI conference to much fanfare. Yes, the theorems were simple. Today's automated theorem provers (ATPs) laugh at the challenge of deriving $\neg q \rightarrow \neg p$ from $p \rightarrow q$, which LT famously met.[16] But the top-down orientation five decades back was wise, and fortunately it is sustained to this day. This orientation was established long before Newell and Simon's success half a century ago. It began with Euclid's archival reasoning, and puzzlement as to why this reasoning is indestructibly compelling. This Euclid-started history is not the history of the mish-mash that AI has become, the grab-bag of techniques and formalisms that students are unfortunately taught in the mad, heterogeneous scurry

[16] This is true whether the ATP in question is resolution-based (by far the more frequent case), or natural deduction-based. E.g., resolution-based Vampire (Voronkov 1995) can find this proof in an instant, and so can natural deduction-based Oscar (Pollock 1995). (Examples much more challenging than proving transposition are available for study in the two publications just cite, and in many others.)

that introductory AI courses have now become. In the beginning there was something shining like nothing else in the intellectual landscape: a proof. Euclid was able to establish certain declarative sentences as indubitable, and the question of what made his reasoning indubitable was answered, at least in large part, when, standing on the shoulders of Boole and his precursor to the modern-day propositional calculus, Frege gave us (albeit in a bizarre notation) first-order logic. Now, Frege's logic is but a dot in the infinite space $\mathcal{F}$ of logical systems, since, as noted earlier, they cover not only the extensional realm (which we now know to include n-order logic, infinitary logic, etc.), but also the modal, epistemic, deontic, paraconsistent, trivalent (indeed, n-valent) ... realms. The history is one in which reasoning gave birth to computation, and has ascended to the point where there is no information processing-based aspect of personhood beyond the reach of logic to simulate, and hence no need of non-logic to capture intelligence.

3.2 The Advent of the Web

There can be little question that the World Wide Web (= W^3, or simply "the Web") has galvanized not only AI, but indeed the entire world of computing. But more importantly, what the Web is *becoming* underscores the importance of logic. Specifically, because knowledge on the Web is increasingly represented in logical systems, so that automated reasoning can take place over this knowledge, the usefulness and power of the Web will skyrocket. In short, when the Web is supplanted by the *Semantic* Web, LAI agents will usher in a new epoch in the information age. This is nicely explained by Berners-Lee, Hendler & Lassila (2001).

It's already being realized that the true power of the Web can be harnessed only if logical systems occupy a position front and center. Oracle's Semantic Technologies Center[17] makes this point immediately. There, you will find that web-based computing, and in fact relational computing, has now explicitly embraced simple logical systems (viz., description logics; for an elegant, comprehensive presentation see Baader, Calvanese & McGuinness 2007) for representing knowledge, and reasoning over it. This is a trend that, it's safe to say, will continue onward and upward. Part of the impetus for this trend is the mathematical fact that relational databases and logic-based knowledge bases can interoperate on the strength of logical systems that have been shown to be unifiers (e.g., see Taylor, Shilliday & Bringsjord 2007).[18]

3.3 The Remarkable Effectiveness of Logic

We now know beyond a shadow of a doubt that logicism in the purely mathematical realm is true. The systematization of mathematics has been provided by many decades of formal exposition in books authored by Bourbaki[19] — exposition that shows, formally speaking, that discovery and confirmation in mathematics consists, fundamentally, in the derivation and use of theorems all extractable from a small set of axioms (e.g., the Zermelo-Fraenkel axioms for set theory) by the use of formal logic.

Likewise, there is simply no denying the power and centrality of logic in computer science (Halpern et al. 2001). For example, formal logic (when used to build the Arithmetical Hierarchy; see Davis, Sigal & Weyuker 1994) is exactly the framework used to specify relative computability.

[17] The Center can be accessed at

http://www.oracle.com/technology/tech/semantic_technologies/index.html

[18] The unifying power of multi-sorted logic (MSL) goes back to the 1950's, as explained in (Manzano 1996), which also offers proofs confirming the unifying power of MSL.

[19] See note 4.

When we specifically turn to intelligence and cognition, and try to be as rigorous about these concepts as logic has allowed us to be about mathematics and computer science, logic quickly becomes the only game in town. For example, consider the problem of determining how impressive is the human ability to take as input the specification of a function, in the context of the attempt to build a computing machine able to itself write computer programs; that is, in the context of the challenge of achieving a program that can automatically write computer programs. It is logic and logic alone that informs us how difficult this engineering challenge is.[20] It's high time that the remarkable effectiveness of logic in mathematics and computer science be naturally extended to cover AI, which is after all in large part a part of computer science.[21]

The present point can be tied specifically to AI itself. That we possess such powers is why we're smart enough, in the first place, to do AI, which is itself, as any decent textbook reveals (e.g., Russell & Norvig 2002), a deliberative, logic-based enterprise. One of the great ironies of non-logicist AI is that following it alone precludes doing AI itself. In particular, we wouldn't have a single theorem undergirding any part of AI were it not for cognition based on explicit declarative information, and logical reasoning over that information. And the point can be generalized: We understand computation only because logic has been brought to bear in the attempt to answer such questions as what, precisely, a computer is, and what are its limits.

3.4 Logic Top to Bottom Now Possible

In the past, there has been a general consensus that while logic-based approaches might be well-suited for deliberative processes that aren't time-sensitive and dynamic, they can't be extended to cover lower-level aspects of human intelligence. In short, the view has been that if one takes the top-down approach described above by Brachman and Levesque, one just doesn't get that far down, and moreover, one certainly can't use logic to work bottom-up, the direction that is part and parcel, and the supposedly the chief virtue, of approaches based directly modeling neural information processing. For example, so the story went: "How could you ever use logic for capturing the real-time interaction between a robot and the physical environment?"

But times, thankfully, are changing. We are now beginning to see that logic can be all-encompassing. Even dynamic perception and action can be systematically logic-based.

Of course, this is simply an assertion, and the proof, I admit, is in the pudding. The fact is, we still don't have ubiquitous logicist robots. But they will come. We know this because, in certain challenging environments, logicist robots are *already* here (e.g., see Bernard, Dorais, Fry, Jr., Kanefsky, Kurien, Millar, Muscettola, Nayak, Pell, Rajan, Rouquette, Smith & Willams 1998, Bernard, Dorais, Jr., Kanefsky, Kurien, Man, Millar, Muscettola, Nayak, Rajan, Rouquette, Smith, Taylor & Tung 1999, Ingham, Clark, Williams, Lockhart, Oyake & Aljabri 2001, Williams, Ingham, Chung & Elliott 2003). While logic has been criticized as too slow for real-time perception-and-action-heavy computation, as you might see in the computational modeling of a human playing a

[20] We know, for example, that in order to write computer programs that compute given functions, it's necessary to be able to decide whether two computer programs compute the same function. Logic informs us that such a decision, at bottom, amounts to deciding whether two Turing machines, n and m, upon being given input u, halt and produce output v. Specifically, the decision revolves around whether the Pi_2 formula

$$\forall u \forall v [\exists k H(n,k,u,v) \leftrightarrow \exists k' H(m,k',u,v)]$$

is true.

[21] It's worth nothing at this juncture that LAI is intimately connected to what Eden (n.d.) calls the *rationalist paradigm* in computer science, in his excellent analysis of the nature of this field. Somewhat predictably, I happen to view computer science as a field falling under this paradigm — but this isn't the place to defend this view, and my affirmation of it is separate from the view set out and defended in the present paper.

first-person shooter game (as opposed to a strategy game, which for obvious reasons fits comfortably under the paradigm of logicism), it has been shown that computational logic, on the strength of ATPs like Vampire (Voronkov 1995), is so fast that it can enable the real-time behavior of a mobile robot simulating human behavior in a robust environment. This has been shown in my lab by having a logic-based mobile robot successfully navigate the wumpus world game (Bringsjord, Khemlani, Arkoudas, McEvoy, Destefano & Daigle 2005), a staple in AI, and a game that humans have long played. (See Figures 2 and 3.) In addition, recall the Wise Man Puzzle discussed above (§2.3.4). LAI engineering carried out in my lab has shown that robots operating in real time can solve this puzzle in real time, when they compute functions based on a logical analysis of the puzzle provided in (Arkoudas & Bringsjord 2005); see Figure 4.[22] This work augments work done in John McCarthy's (logicist) AI Lab that has shown it to be possible to control a real robot, operating in a realistic office environment in real time (Amir & Maynard-Reid 2001, Amir & Maynard-Reid 2000, Amir & Maynard-Reid 1999).[23] In this approach, a logicist calculus is used to represent time and change. Usually the calculus is the situation calculus, but the event calculus can also be used; both are summarized in (Russell & Norvig 2002), and a recent book dedicated to the latter is (Mueller 2006). It's important to know that such work is far from peripheral and tentative: Logic-based AI is starting to reveal that even in the area of perception and action, the speed demands can be met via well-established techniques that are fast becoming part of the standard toolkit for the field, as seen by such textbooks as (Reiter 2001).

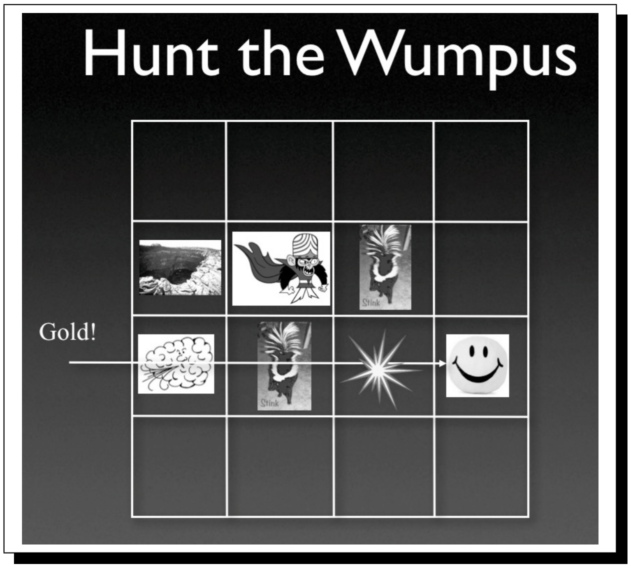


Figure 2: The Wumpus World Game. *In the wumpus world, a robot must navigate a work in matrix form, where cells in the grid may contain pits or a monster (the Wumpus). The robot must shoot and kill the Wumpus, and retrieve the gold.*

When perception and action yield to logic, non-logic will be reserved for stage-one transduction. Transduction is discussed below in section 4.4, and there I distinguish between stage-one and stage-two transduction.

[22] The robots in my lab are autonomous. Humans do not find the proofs in question. Rather, the robots do.

[23] This research can be found online at: http://www-formal.stanford.edu/eyal/lsa.

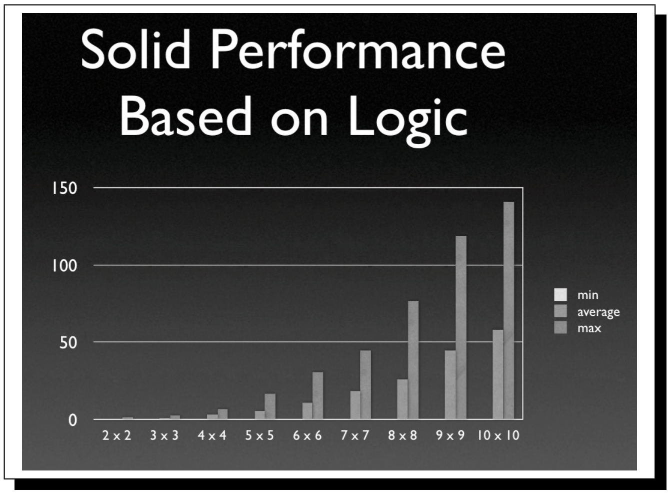

Solid Performance Based on Logic

Figure 3: Performance of a Logic-Powered Robot in the Wumpus World. *This graph shows the time (in secs) it takes the logic-powered robot to succeed in the wumpus world, as a function of the size of the world (i.e., the size of the grid). The speed is really quite remarkable. Fine-tuning the use of the SNARK theorem prover was carried out by Matt Daigle.*

3.5 Learning and Denial

What has been called 'learning' in AI simply isn't. The field is in denial about this, and logicists need to call a spade a spade, and then leave to solve the problem on their own.

You have learned that Selmer is inclined to issue some rather strong recommendations regarding AI — and indeed you've learned hundreds of other things since reading from the title of this piece, to this phrase in this sentence. In general, the bulk of your learning, since you started school, has come by way of reading and reasoning. And yet in AI we operate with the bizarre concept that "machine learning" is all and only about divining a function based on multiple trials. In this paradigm, the sample input-output pairs from which the function is to be induced are far, far removed from the reading-and-reasoning-based learning that has put you in position to understand the present paper. The same absurdity is seen when we look at science and engineering itself: We add to our stock of knowledge about the cosmos by deploying painstaking logic-based formalisms, and by careful reasoning over representations in them. We build theories by abduction; we look to accumulate evidence for them by experiment and induction; and we look to refute them by deduction (when what they predict doesn't obtain). It's all declarative reasoning, with a logical system available to formalize every inch. Again, we are persons, so we can do science and engineering to learn about the universe, and the learning is of a type that has nothing to do with multi-trial runs in an impenetrable neural network. Logic and logic alone is the paradigm available for giving the power of learning by reading, and related techniques, to a machine. (For a look at logicist machine reading research, see (Bringsjord, Arkoudas, Clark, Shilliday, Taylor, Schimanski & Yang 2007).)

3.6 Logic is an Antidote to "Cheating" in AI

Logic prevents cheating. I'm not talking about turpitude; I'm talking about a form of involuntary cheating in AI that is all too easy for researchers to slide into when the formalisms upon which their paradigms are based don't generate argument-based justifications, including justifications that are full-blown proofs. Here's one way to explain the situation.

Suppose one is striving to engineer an intelligent computational agent capable of excelling

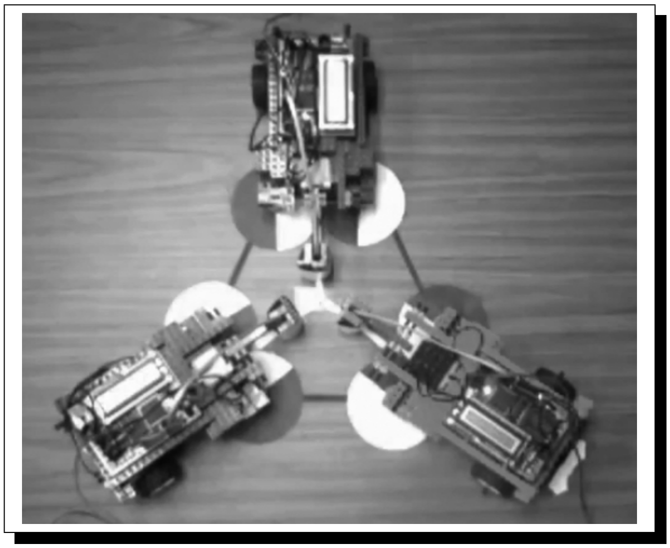


Figure 4: Three Robots In the Wise Man Puzzle; Far-left One Solving the Puzzle. *Implementation carried out by Evan Gilbert.*

on tests of intelligence.[24] Let's assume, specifically (but without loss of generality), that this agent receives a series of short multiple choice questions designed to test for context-independent reasoning. For example, consider the following simple (from the logicist standpoint) problem, a slight variant[25] of a puzzle introduced by Johnson-Laird (1997).

> Assume that the following is true:
>
> 'If there is a king in the hand, then there is an ace in the hand,' or 'If there is not a king in the hand, then there is an ace in the hand,' — but not both of these if-thens are true.
>
> What can you infer from this assumption?

Now suppose that some AI agent answers in surprising but correct (!) fashion by printing out or uttering

"That there isn't an ace in the hand."

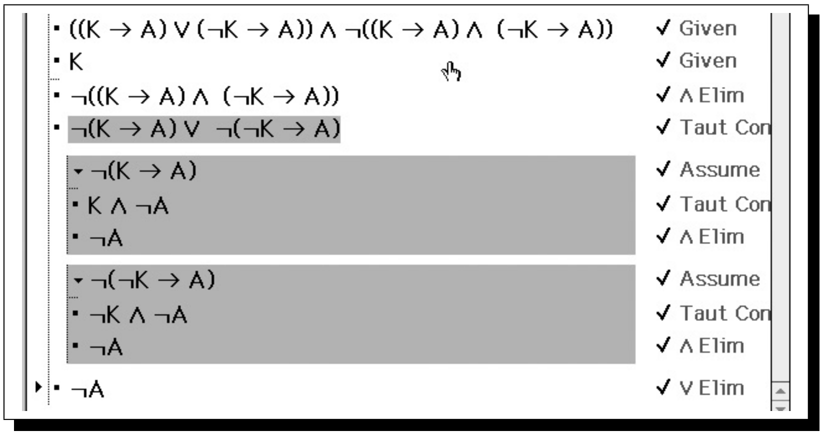
but that it cannot provide a proof that this response is correct. In this case, how do we know that the agent didn't cheat? How do we know, for example, that the agent didn't produce this response because it blindly computed a function from the number of characters used in the question, to that-clause strings built randomly from the same question — a function having nothing to do with the fact that the correct answer is correct precisely because it can be deduced from the given information? The fact is, we don't. (For a proof that the right answer is indeed that there isn't an ace in the hand, see Figure 5.) This is why the first stage of Project Halo, in which AI systems able to answer questions on the Advanced Placement (AP) Chemistry Exam[26] were engineered, required of the systems involved that they provide declarative justifications (Friedland et al. 2004).

[24] One might seek such an agent if one subscribes to the view that the concept of intelligence should be operationalized in a way paralleling Turing's test-based operationalization of thinking: viz., by tests used in psychometrics, the field devoted to measuring intelligence and other mental abilities. Given this view, an *artificial* intelligence is an artifact able to successfully take, and do well, if not excel, on tests of mental ability. For an account of this kind of AI, called "Psychometric AI," see (Bringsjord & Schimanski 2003).

[25] The variation arises from disambiguating Johnson-Laird's 's or else s'' as 'either s or s', but not both.'

[26] Published and administered by ETS. Information is available on the Web at http://www.ets.org.

And in general, this is why logicist AI prevents cheating. As AI unfolds into the future, only logic will prevent researchers from resting content with agents that merely *appear* to have meaningful, compelling reasons for doing what they do.



$((K \rightarrow A) \lor (\neg K \rightarrow A)) \land \neg((K \rightarrow A) \land (\neg K \rightarrow A))$ ✓ Given

K ✓ Given

$\neg((K \rightarrow A) \land (\neg K \rightarrow A))$ ✓ ∧ Elim

$\neg(K \rightarrow A) \lor \neg(\neg K \rightarrow A)$ ✓ Taut Con

$\neg(K \rightarrow A)$ ✓ Assume

$K \land \neg A$ ✓ Taut Con

$\neg A$ ✓ ∧ Elim

$\neg(\neg K \rightarrow A)$ ✓ Assume

$\neg K \land \neg A$ ✓ Taut Con

$\neg A$ ✓ ∧ Elim

$\neg A$ ✓ ∨ Elim

Figure 5: A Proof That There is No Ace in the Hand in $\mathcal{F}$

3.7 Logic Our Only Hope Against the Dark AI Future

As is well-known, Joy (2000) has famously predicted that the future will bring our demise, in no small part because of advances in AI and robotics. Unless we build robots on the basis of logic, we will indeed be overrun by malicious robots, and what is now fiction from Asimov, Kubrick, Spielberg and others will become reality. The antidote is to ensure that robots are reasoning in correct fashion with the ethical codes we supply. A bit more precisely, we need to put two constraints into play:

1. Robots only take permissible actions.
2. All relevant actions that are obligatory for robots are actually performed by them, subject to ties and conflicts among available actions.

But here is the interesting thing: Ethics is itself an immutably logic-based field, and we have no hope of sorting out how these two conditions are to be spelled out and applied unless we bring ethics to bear. Ethicists work by rendering ethical theories and dilemmas in declarative form, and reasoning over this information using informal and/or formal logic. This can be verified by picking up any bioethics textbook (e.g., see Kuhse & Singer 2001). Ethicists never search for ways of reducing ethical concepts, theories, principles to sub-symbolic form, say in some numerical format, let alone in some set of formalisms used for dynamical systems. They may do numerical calculation in *part*, of course. Utilitarianism does ultimately need to attach value to states of affairs, and that value may well be formalized using numerical constructs. But what one ought to do, what is permissible to do, and what is forbidden — this is by definition couched in declarative fashion, and a defense of such claims is invariably and unavoidably mounted on the shoulders of logic. This applies to ethicists from Aristotle to Kant to G.E. Moore to J.S. Mill to contemporary thinkers. If we want our robots to be ethically regulated so as not to behave as Joy tells us they will, we are going to need to figure out how the mechanization of ethical reasoning can be applied to the control of robots.[27] What is the alternative?[28]

[27] Along with others (e.g., Arkin 2008), I've partially figured this out: (Bringsjord, Arkoudas & Bello 2006).

[28] Of course, *in general*, without logic, we don't have the foggiest idea whether our software will behave the way we want it to. It's no coincidence that programming languages are formalized and verified using logic. Again, what is

4 Objections; Rebuttals

4.1 "But you are trapped in a fundamental dilemma: your position is either redundant, or false."

"Unfortunately, you stand between the horns of a dilemma, and either way you lose. On one interpretation of the 'independence' you're advancing, your recommendation is already functionally in place, which makes your paper superfluous. After all, in the day-to-day life of AI, all the subfields are already essentially 'divorced.' Just like cell biologists work independently of whole-animal biologists, and topologists work independently of number theorists (usually, anyway), in AI (and, for that matter, Cog Sci) logicist researchers work independently of, for example, connectionists. Surely you must concede that, as a rule, fields are composed of sub-fields, and those sub-fields often are driven by radically different methodologies and formalisms. What unifies the sub-fields is a shared objective. For example, psychology seeks to understand the mind/brain, and has under its umbrella behaviorists standing alongside cognitive psychologists, and both groups stand alongside neuropsychologists. But on the other hand, perhaps you're recommending something stronger, namely, logicist AI *uber alles*. Here arises the other horn of the dilemma you face. For the fact is, LAI *uber alles* is highly contentious, and worse, it's probably false (e.g., see Michael Spivey's *The Continuity of Mind*)."[29]

Actually, the fact of the matter is that I'm advancing *both* propositions; that is, that (i) LAI, at the operational level (if you will), should be independent in concrete practice, *and* that (ii) LAI *uber alles* is correct. As to (ii), my critic is without question right that logicist independence is contentious — but that, of course, is no small part of why I have crafted a sustained argument in support of such independence. Were there no controversy, were there no unsettled questions about the best relationship between the logic-inclined and those focused on continuous systems, there would be no point in articulating and defending this present manifesto. (Now the critic also claims that LAI is overthrown, in light of specific arguments in favor of the continuity of mind. I rebut this part of the objection in a separate dialectic below (§4.3).) Regarding the alleged superfluity of my case for (i), the objection simply goes wrong, for many reasons. Let me explain.

First, mathematics is actually not at all a field whose compartmentalization supports the objection here given against my call for logicist independence. This is easy to see once one studies sufficiently robust theorems. A recent example would be Wiles' proof of Fermat's Last Theorem (Wiles 1995, Wiles & Taylor 1995). Wiles' reasoning shows, without question, that subfields of mathematics (in this case, number theory, geometry, and algebra) are actually connected at a deep level — a level that some brilliant humans can access, and exploit. In fact, in general, it is now known that *every* subfield is related to every other subfield, for at the deepest level (axiomatic set theory expressed in first-order logic) all subfields are about the very same structures (see note 4). Nothing like this unites logicist AI with (say) connectionist AI: Advances in the former (latter) are not made by linking to advances in the latter (former).[30]

the alternative?

[29] Some of those inclined to press the objection just given against me might also claim that logic *uber alles* runs afoul of the brute fact that no one knows how to build a computing machine with human-level intelligence. But my manifesto in no way presupposes any such thing as that logicists have solved the AI problem. Instead, the idea is that to assemble the best possible attack on this problem, a new, standalone field (logic-based AI) needs to be founded. In short, the idea is that the best bet for reaching AI's ultimate goal of building an artificial person is to establish logic-based AI as a field unto itself.

[30] Of course, it nonetheless remains true that any classical mathematics used in any subfield of AI, from dynamical systems to artificial neural networks to Bayesian networks to quantified modal logic to ..., is all based on formal logic. This is simply an indisputable aspect of the formal sciences, and cuts through any and all controversy.

For the next step in my rebuttal, consider that philosophy, along with psychology, has the objective of understanding the mind/brain — indeed this goal is what drives philosophy of mind. But psychology, as is well-known, split off from philosophy nonetheless. The divorce occurred because psychology wanted to discover things via the *a posteriori* route, not the *a priori* one. Psychologists put controlled, empirical experiments at the heart of their *modus operandi*. Philosophers, though sometimes empirically inclined, continue to place armchair reasoning and reflection at the heart of what they do. So the split did take place because of a difference in methodologies and formalisms, and having an objective in common failed to keep the marriage intact. Were the objection here allowed to stand, the divorce between philosophy and psychology would have made, and indeed would *now* make, little sense. Parallel points could be made for the other major splits; for example, for philosophy and physics, philosophy and linguistics, and so on. Methodologically speaking, the difference between logicist AI and sub-symbolic AI is a vast canyon, and it's high time that a split take place in AI that parallels those that have been salubriously achieved in the case of philosophy and other such examples.

The objection also cites the compartmentalization that exists in AI, and in biology (cell- v. animal-based). (I have already addressed the example cited in mathematics (topology v. number theory).) The idea is that such compartmentalization makes my call for independence redundant. The fatal problem with this part of the objection is twofold: One, compartmentalization of areas/approaches $A_1, \ldots, A_n$ within field F is radically different than fields $F_1, \ldots, F_n$ overlapping to a degree by virtue of their long-term objective. As already pointed out, if these scenarios were the same, or if they could be collapsed, then philosophy and psychology (and any number of such divorced, now-self-contained, and separate fields) would themselves be the same and collapsible — which most assuredly they are not. What is the difference between the two scenarios? What's the difference between cell biology versus animal biology on the one hand, and the separate fields of psychology and philosophy, or linguistics and philosophy, or physics and philosophy, on the other? The answer is surprisingly easy to find, by attending to the fact that whereas the field of biology offers students of this field *single* introductory textbooks that cover both the cell and animal levels, introductory psychology textbooks, which include coverage of all the many compartmentalized subfields in psychology, are quite disjoint from introductory philosophy textbooks, which in turn have no overlap with introductory physics textbooks, and so on. The textbook situation is a tell-tale sign: A single textbook for a field F is effective when all those working in A_i under F, to be maximally effective, should both know of particular structures, tools, techniques, and formalisms relevant to each A_i, and should also know of each A_i itself. This situation obtains in the case of biology and mathematics, fields invariably approached by students exposed to comprehensive textbooks. But the situation in AI is actually the reverse: Not only is it false that all those working in AI, to be maximally effective, should know of the particular aspects of both (say) connectionism and LAI, but in point of fact a requirement that students have such knowledge is an obstruction. And, it's false as well that to be effective as a connectionist, one should be versed in formal logic (with the *vice versa* false as well). In fact, a concrete consequence of my manifesto would be the appearance of new textbooks dedicated to only one of the approaches to AI that are now typically mashed nonsensically together is gargantuan introductory textbooks (such as, e.g., the encyclopedic Russell & Norvig 2002).

4.2 "But you're neglecting probabilistic AI."

"You have been speaking as if the choice is between top-down logic and bottom-up neurocomputational approaches. But that is an illusory dichotomy. The real choice is between a symbolic approach that is probabilistic in nature, armed now with Bayesian networks, versus the non-symbolic bottom-up neurocomputational approach."

Of course, I'm well aware of the resurgence of probabilistic techniques in AI, the start of which, in my opinion, coincided with the arrival of (Pearl 1988). These techniques are indeed symbolic in nature,[31] and should be dutifully included in any purportedly complete overview of AI — but of course the present paper is no such overview. Furthermore, yes, Bayesian nets provide a representation framework that allows for the calculation of posterior probabilities in a manner much more efficient than standard calculation over a full probability distribution. But what, pray tell, is the relationship between these facts and the call for logicist independence? I very much hope that no one is silly enough to propose any such idea as that because probabilistic techniques are so powerful, logical systems can be defenestrated. The reason this is silly is because even the simplest of logic problems, in fact ones given to students in *Logic 101*, cannot be answered by intelligent agents powered solely by probabilistic inference. We have already seen an example earlier: the King-Ace Puzzle. How would a Bayesian system solve this problem, where the solution includes a full, watertight deduction in support of the answer that there is in fact no ace in the hand? I don't even think Bayesian systems can possibly solve logic problems that involve probability. For example, here is another problem involving cards, kings, and aces that Johnson-Laird challenged me with quite a while back:

> If one of the following assertions is true then so is the other:
>
> 1. There is a king in the hand if and only if there is an ace in the hand.
> 2. There is a king in the hand.
>
> Which is more likely to be in the hand, if either: the king or the ace? Prove that you are correct.

I would very much like to see a Bayesian system take this declarative information in as input, and yield the correct answer, and a proof that that *is* the answer. I assure you that I will not hold my breath. For a human or machine using just elementary techniques in logicist AI, this is a trivial problem.[32] We will return to this problem in the section immediately below, in our discussion of the dynamicist's rejection of LAI.

To sum up, Bayesian approaches are parasitical on, and in fact derived from, a smallish amount of knowledge expressed in relatively simple logical systems. To the extent that Bayesian approaches are desirable, logicism is desirable. However, as illustrated by the impotence of Bayesianism to solve even simple problems that persons can routinely solve using logic, Bayesianism is only a tiny part of logicist AI. And finally, to make a point there isn't space to fully consider, persons routinely handle uncertainty in ways that don't involve probabilities in the least, but center around arguments and counter-arguments, sometimes along with strength factors, instead.[33]

[31] In fact, they are ultimately logicist in nature! Kolmogorov's axioms, viz.,

1. All probabilities fall between 0 and 1. I.e., $\forall p(0 \leq P(p) \leq 1)$.
2. Valid (in the traditional logic-based sense of being true on all formal interpretations) propositions have a probability of 1; unsatisfiable (in the traditional logic-based sense) propositions have a probability of 0.
3. $P(p \vee q) = P(p) + P(q) - P(p \wedge q)$

are simple formulas from a simple logical system, and modern probability theory can be derived from them in straightforward fashion. The expressiveness of probabilistic representation and reasoning is bounded by the *logical system* in which it's expressed, and the two systems in question (the propositional calculus and first-order logic), from the standpoint of the infinite space of logical systems available in LAI, are two particles of sand on a beach reaching from here to Mars.

[32] A full solution generated by these techniques will be supplied upon request. The proof is not difficult.

[33] For logicist AI based on this way of formalizing and mechanizing uncertainty, see (Pollock 2001, Pollock 1992).

4.3 "But we now know that the mind, *contra* logicists, is continuous, and hence *dynamical*, not logical, systems are superior."

"You argue for logic *uber alles*, as you yourself say. Unfortunately, we now know that the mind is a continuous thing, one best understood and modeled by bringing together dynamical systems theory, cognitive and computational neuroscience, connectionism, and ecological psychology to provide an understanding of the mind (Spivey 2006). Logic-based AI is passé."

Spivey's book provides an elegant review of work designed to understand aspects of the brain and cognition in terms of his preferred collection of subfields. (To facilitate exposition, let's denote this collection — dynamical systems theory, cognitive and computational neuroscience, connectionism, and ecological psychology; the collection is characterized in Chapter 1 of (Spivey 2006) — as $\bar{\mathcal{L}}$.) Of course, everyone knows that within the confines of $\bar{\mathcal{L}}$, many, many impressive achievements have been won under the banner of AI, just as everyone knows that within the confines of logicist AI, many, many impressive achievements have *also* been won under the very same banner. In general, prize achievements within $\bar{\mathcal{L}}$ have been those that require low-level perception and action, and are in the sphere of the *sensible*; and prize achievements in LAI have been in the sphere of the *intellectual*, and involve high-level reasoning and problem-solving. Spivey (2006) hits the nail on the head in Chapter 10 by pointing out that logicists are with Kant in the following quote, while dynamicists reject the dichotomy:

> Now, man actually finds in himself a power which distinguishes him from all other things—and even from himself so far as he is affected by objects. This power is *reason* ... Because of this, a rational being must regard himself *qua intelligence* (and accordingly not on the side of his lower faculties) as belonging to the intelligible world, not the sensible one. (Kant, as trans. by Seidler 1986)

Everyone (including even the dualists Spivey lampoons) also knows that every physical object, including the brains and central nervous systems of intelligent carbon-based creatures, are objects moving through time and space in a manner that, from certain perspectives (e.g., physics), are best described using continuous formalisms. But in the context of the debate in which my manifesto figures, the issue is whether (a) the kind of intelligence at the heart of the specific arguments in favor of independence given above (§3)[34] can be captured by $\bar{\mathcal{L}}$ in light of *The Continuity of Mind*, and whether (b) in light of Spivey's book it's revealed that there are elements of AI's ultimate goal that can be reached only if pursued in the paradigm of $\bar{\mathcal{L}}$. Does Spivey make a case for either (a) or (b)? If so, is the case successful?

Actually, the fact of the matter is that Spivey doesn't really make a clear case for (a) or (b). When it comes to reasoning, when it comes to the kind of intellectual activity that makes use of logical systems (which by definition are highly abstract; in fact, often they are abstractions of abstractions, as when for example a logical system is used to formalize abstract mathematics), Spivey does *not* maintain that $\bar{\mathcal{L}}$ is up to the task. For example, we read:

> [C]ontrary to the attitude that I cop throughout most of the book, reasoning and problem solving may very well be the one area of cognition where the rule-and-symbol framework has not yet run its course, and some further useful advances may still be coming from this approach from a little longer. (Spivey 2006, p. 259)

Given that logical systems, as defined above, from the mathematical standpoint, have "rule-and-symbol frameworks" as but a tiny part; and given that, hitherto, LAI has not implemented logical systems in parallelized fashion on high-performance hardware, Spivey should no doubt be even

[34] And others that I haven't the space to cover in this paper.

more cautious, in my opinion. And indeed later in his book, he is: A more circumspect attitude is divulged at the conclusion of Spivey's chapter on reasoning, where he writes that "Perhaps ... continuous dynamical descriptions of cognitive phenomena [will] finally wash up against some firm bedrock that forms the core of highly complex mental processes like reasoning and problem solving ... And then again perhaps not" (Spivey 2006, p. 285).

Nonetheless, Spivey does make a few brave gestures in the direction of supporting (a) and (b). I don't have the space to evaluate all these gestures; I consider just one now, one given in support of (a). This gesture consists in Spivey's pointing out that some reasoning problems are cracked by the triggering and use of perceptual-motor subsystems in the human case. (Such subsystems, as is well known, are traditionally modeled rather credibly by $\bar{\mathcal{L}}$.) For example, he discusses Duncker's (1945) candle-mounting problem, in which subjects, after being given as tools a candle, a box of tacks, and a book of matches, are asked to try to figure out how to mount the candle on the wall using only these items. Only some subjects see the solution, which is to tack the box onto the wall, and place the candle on (or in, possibly) the box. Spivey, echoing Glucksberg (1964), explains that when subjects do physically touch and manipulate the items in question, they are more likely to have that "Aha!" discovery of the solution.

Unfortunately, the candle-mounting problem (and others cited by Spivey) is tailor-made to fit the $\bar{\mathcal{L}}$ paradigm. In order to substantiate (a), or even take appreciable steps toward substantiating this proposition, it would be necessary to take a problem that is firmly within the — to harken back to Kant's distinction — intellectual realm, and show that somehow the solution to the problem doesn't involve processes defined over the relevant logical system. For example, recall the probabilistic king-ace problem presented above. Here is an abstract version of the problem:

> If one of the following propositions is true then so is the other:
>
> 1. K if and only if A
> 2. K
>
> Which is more likely, if either: K or A? Prove that you are correct.

I submit that all readers, whether of the logicist or dynamicist persuasion (or for that matter *any* persuasion), can rather easily see (esp. if they manage to solve the problem in question) that solving this problem happens only on the strength of manipulating symbols, pure and simple. There is absolutely nothing in *The Continuity of Mind*, nor in the literature it ranges over in order to find results that support the dynamicist position, that provides even an iota of guidance for how such a problem can be solved without relying upon one or more logical systems. Therefore, from the standpoint of an AI researcher who wants to engineer a system able to autonomously solve such a problem (which, frankly, compared to the kind of problems that those in the formal sciences get paid to solve is laughably concrete and simple), it is a complete non-starter to forsake LAI.

4.4 "But surely human-level cognition is *partly* sub-symbolic."

"Well, okay, I concede it can't be concluded that the mind does *everything* on the strength of continuous processes, nor that a computing machine can be engineered to operate at the level of human persons only if such processes are discovered, dissected, and implemented to operate on that machine. But surely persons do *some* things on the strength of non-symbolic processing. Logicist AI as you define it, if launched, would be a field completely ignoring these things!"

It's probably true that a toddler catches a ball on the strength of a process that, though logicizable in a machine, is not logic-in-action in her body, at least at some level of description.

But there are at least three reasons why this kind of rapid, non-deliberative perception and action is no threat to my declaration of independence.

First, if the toddler seeks to be maximally proficient at such tasks as she grows, she must rely on logic. This is why great athletes have coaches: to bring them to the next level by tapping into training based on explicit reasoning and knowledge.

Second, AI isn't cognitive science (CogSci). The latter is often quite computational, but computation in this field is used to model and understand human (and animal) cognition. By contrast, AI is concerned with engineering artifacts that are intelligent, whether or not what's under the hood matches processing in the human case.

Third, it's important to realize that the first stage of transduction is *not* part of what it is to be a first-rate cognizer. Transduction is the translation of raw data from the environment into logicist form. This translation occurs in two stages. In stage one, the physical changes in sensors caused by their direct interchange with the external environment is recorded and represented in some format, for example in an array of pixels each of which has some particular value. In stage two, these values are translated into declarative representations expressed in some logical system. For example, an array of zero's and one's might lead to some configuration of objects having semantic value, in the sense that the agent in question already has some declarative knowledge about these objects. In general, stage-two transduction is not a conscious process, but at least in principle it can be. As such, it is a process that can itself be carried out by reasoning; hence, naturally enough, the reasoning can be captured in a logical system.

I'm happy to concede that stage-one transduction may best be mechanized in non-logicist ways. But as has been pointed out rather long ago, it is entirely possible, mathematically speaking, for a creature with the intelligence of a person to exist without there having been *any* physical interaction between this creature and an outside environment (Bringsjord & Zenzen 1991) at the level of stage-one transduction. In other words, the cognition constitutive of personhood doesn't include stage-one transduction (and for that matter doesn't include perception and action in interchange with an external environment). Work by non-logicists on stage-one transduction-level problems can complement work done by logicists, but only the latter effort pertains directly to personhood, which is after all the goal (recall §2.1.1).

4.5 "But what concrete benefits flow from divorce?"

"What benefits accrue from declaring independence, instead of staying under the umbrella? After all, at present, parts of AI are certainly logic-based: there are logicists doing planning, and even (e.g., through inductive logic programming) learning. Plenty of logicists seem happy and productive in the current under-one-umbrella situation, even if folks under that umbrella are arranged in cliques, and indeed cliques within cliques (e.g., logicist planning people working unto themselves)."

We've already discussed some of the thrust of This objection, but it certainly compels one to consider the general question as to why *any* field is unto itself — or better: why any field *ought* to be unto itself. It seems to me that there can only be one general rationale in favor of a field unto itself, as opposed to being a sub-field: there must be something that separatism buys that ecumenism doesn't. Well, there *is* something that an independent logic-based AI buys us that is apparently otherwise unattainable: unification of logicist activities in the service of reaching the ultimate goal: building a person. As of now, logicist AI as practiced under the all-inclusive tent is fragmented. There are logicist researchers doing NLP, planning, learning, theorem proving, robotics, and so on — but these researchers aren't working together in order to build an artificial system that calls upon all these areas *at once*. In short, the goal of building an artificial person is not one that logicists seem to be able to shoot for, in the present situation. Ironically, logicists in AI are working in the kind of stultifying (relative to reaching the grand goal of mechanizing personhood) isolation that Newell (1973) detected and damned in psychology. Recent developments indicate that a recrudescence of *human-level* AI is underway (Cassimatis 2006, Nilsson 1995, Nilsson 2005, Brooks, Breazeal, Marjanovic, Scassellati & Williamson 1999, Pollock 1989, Pollock 1995), and in my opinion this activity is sustainable only if logic leads the way, something that appears to be the case as of now, given the nature of the work of these researchers.

In addition, logicist independence will allow us to once and for all stop pretending to ignore deep sociological rifts at the heart of the logicist vs non-logicist clash — rifts I used to believe were just an immutable part of the nature of the field (Bringsjord 1991). Breaking off, I believe, will release cathartic honesty.

5 Conclusion

5.1 Is Independence Realistic?

How realistic is independence for LAI? My view is that, at this particular time, there really and truly is a chance to plant logic-based AI as a self-contained field. This belief stems not just from the fact that I affirm the reasoning given above, and expect rational readers to do so as well, but from a more concrete development in funded AI R&D. As a covering label for the phenomenon in question, 'logicist interoperability' is as good as any. The basic idea is that there is a growing realization that different logical systems can be connected in very fertile ways. One of these ways is a downward direction in which a logical system $\mathcal{L}$ is encoded, at least in part, in a less expressive logical system $\mathcal{L}'$, where the key algorithms run on knowledge in $\mathcal{L}'$ are much more efficient than their counterparts run on knowledge in $\mathcal{L}$. For example, it has been shown by Arkoudas & Bringsjord (2005) that computationally expensive epistemic logics can be "encoded down" to multi-sorted logic, known via some longstanding theorems to be a "grand unifier" (Manzano 1996). MSL provides a very efficient framework for automated reasoning. As another example, knowledge in FOL can often be encoded down to knowledge in the propositional calculus, allowing SAT solvers, which are very efficient compared to general reasoning in FOL, to be employed. This approach is taken by Mueller (2006). (For information on SAT-based work, see (Kautz & Selman 1999, Kautz & Selman 1996).) In addition, Common Logic, about now an ISO standard, is currently being used in funded R&D to serve as an *inter lingua* enabling machine translation from knowledge bases expressed in one logical system $\mathcal{L}$ to another $\mathcal{L}'$. The bulk of this work, to date, has been funded by the research arm of the U.S. intelligence community.

5.2 The Two Big Challenges

Logicists must confess that two mammoth obstacles stand in the way of full success. The first big challenge is the mechanization of natural language understanding and generation.

Turing (1950) predicted over half a century back that by now we would be able to engineer machines linguistically indistinguishable from us (i.e., machines able to pass his so-called "Turing Test"), but the fact of the matter is that, today, a bright toddler's conversational reach still exceeds that of any and all computers on our planet. No robust computational accounts of human-level communication (attribute 4 in the list of capacities constitutive of personhood, given, recall, in §2.1.1) exist. Even Anderson (2003), someone quite sanguine about the future of attempts to reduce human-level cognition to computation, concedes that natural language is currently out of reach; that in this regard "Newell's Program" has not yet succeeded. There are those (e.g., Moravec 1999) who hold that, relatively soon, person-level communication will be mechanized. Unfortunately, such writers are confident because of the continuous increase in processing speed produced by Moore's Law, but raw processing speed is not the problem (as explained in Bringsjord 2000): the challenge, from the standpoint of LAI, is to discover the logic-based structures and procedures that enable human persons to communicate in natural languages.

What are the prospects for this discovery coming to pass? At the dawn of AI in the United States, and for at least three decades thereafter, the dream was to capture natural languages like English, German, and Norwegian completely in first-order logic (= in $\mathcal{L}_I$) (e.g., see the FOL-based Charniak & McDermott 1985). Unfortunately, this specific logic-based approach has not succeeded. In fact, some originally logic-based experts in computational language processing have turned their backs on logic, in favor of purely statistical approaches. Charniak himself is an example. In 1985, his comprehensive-at-the-time *Introduction to Artificial Intelligence*, which we of course visited at the outset of the present essay, gave a strikingly unified presentation of AI, including natural language processing. This unification was achieved via $\mathcal{L}_I$, which runs throughout the book and binds things together. But Charniak abandoned logic in favor of purely statistical approaches (Charniak 1993).

To this point, despite the richness of the family $\mathcal{F}$, natural language has resisted attempts to model it in logico-computational terms (and, indeed, in *any* terms). However, it seems clear that some traction has taken hold in the attempt to model *fragments* of natural language in formal logic (e.g., see Fuchs, Schwertel & Schwitter 1999), and this direction is certain to see more investment, and, I believe, great progress. Only time will tell if this research and development will be able to scale up to all of natural language.

The second major challenge is one that currently paralyzes *any* approach to AI: subjective consciousness. While some forms of consciousness have been modeled (e.g., see Sun 1999), there are today no simulations of *subjective* consciousness (attribute 2 in the list of capacities constitutive of personhood). No one has a third-person account of what it is to (say) experience the taste of deep, dark chocolate, or what it is to *be* you (Yang & Bringsjord 2003, Bringsjord 1998, Bringsjord 2001, Bringsjord 1995*b*, Bringsjord 1999). (For a discussion of these matters in connection with robotics, see Bringsjord 2007.) Absent such an account, mechanization — indeed, taking just initial steps toward some mechanization — is rather difficult. Given the importance of consciousness in human cognition (after all, the reason humans seek to continue to live is to continue to have conscious experiences), there is little doubt that in the future LAI will increasingly be marked by a persistent attempt to capture consciousness. Breakthroughs are waiting to be made.

5.3 Forward to What Could Have Been

DARPA sponsored the original 1956 conference at Dartmouth because the proposal for that conference contained a rather attractive idea, one that still rings as strong today as it did then, namely: "It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture." The original proposal, however, was ecumenical. For example, it also called for work carried out on the basis of "neuron nets."[35] Well, enough; enough AI by multi-paradigm coalition. For the reasons expressed above, and others, it is time to simply delete 'a large part' from the quote above, and to forge ahead full steam with the intention of showing that even synthetic workalikes for those parts of the human person generally considered to be below "thought" can be engineered through logic.

[35] The original proposal is available online, e.g. at

http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

. It is entitled "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," and was authored by McCarthy, Minsky, Rochester, and Shannon.

References

Amir, E. & Maynard-Reid, P. (1999), Logic-based subsumption architecture, *in* 'Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence'.

Amir, E. & Maynard-Reid, P. (2000), Logic-based subsumption architecture: Empirical evaluation, *in* 'Proceedings of the AAAI Fall Symposium on Parallel Archittectures for Cognition'.

Amir, E. & Maynard-Reid, P. (2001), LiSA: A robot driven by logical subsumption, *in* 'Proceedings of the Fifth Symposium on the Logical Formalization of Commonsense Reasoning'.

Anderson, J. & Lebiere, C. (2003), 'The newell test for a theory of cognition', *Behavioral and Brain Sciences* **26**, 587–640.

Arkin, R. C. (2008), Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture – Part iii: Representational and architectural considerations, *in* 'Proceedings of Technology in Wartime Conference', Palo Alto, CA. This and many other papers on the topic are available at the url here given.
URL: *http://www.cc.gatech.edu/ai/robot-lab/publications.html*

Arkoudas, K. & Bringsjord, S. (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 Artificial Intelligence (LNAI)*, Springer-Verlag, New York, pp. 111–125.
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*

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

Baader, F., Calvanese, D. & McGuinness, D., eds (2007), *The Description Logic Handbook: Theory, Implementation (Second Edition)*, Cambridge University Press, Cambridge, UK.

Barwise, J. & Etchemendy, J. (1999), *Language, Proof, and Logic*, Seven Bridges, New York, NY.

Bernard, D. E., Dorais, G. A., Fry, C., Jr., E. B. G., Kanefsky, B., Kurien, J., Millar, W., Muscettola, N., Nayak, P. P., Pell, B., Rajan, K., Rouquette, N., Smith, B. & Willams, B. C. (1998), Design of the Remote Agent Experiment for Spacecraft Autonomy, *in* 'Proceedings of the IEEE Aerospace Conference', Vol. 2, pp. 259–281.

Bernard, D. E., Dorais, G. A., Jr., E. B. G., Kanefsky, B., Kurien, J., Man, G. K., Millar, W., Muscettola, N., Nayak, P. P., Rajan, K., Rouquette, N., Smith, B., Taylor, W. & Tung, Y.-W. (1999), Spacecraft Autonomy Flight Experience: The DS1 Remote Agent Experiment, *in* 'AIAA Space Technology Conference and Exposition', Albuquerque, NM.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001), 'The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities', *Scientific American* **284**(5), 34–43.

Bourbaki, N. (2004), *Elements of Mathematics: Theory of Sets*, Verlag, New York, NY. This is a recent release. The original publication date was 1939.

Brachman, R. J. & Levesque, H. J. (2004), *Knowledge Representation and Reasoning*, Morgan Kaufmann/Elsevier, San Francisco, CA.

Bringsjord, S. (1991), 'Is the connectionist-logicist clash one of AI's wonderful red herrings?', *Journal of Experimental & Theoretical AI* **3.4**, 319–349.

Bringsjord, S. (1995*a*), Could, how could we tell if, and why should–androids have inner lives?, *in* K. Ford, C. Glymour & P. Hayes, eds, 'Android Epistemology', MIT Press, Cambridge, MA, pp. 93–122.

Bringsjord, S. (1995*b*), 'In defense of impenetrable zombies', *Journal of Consciousness Studies* **2**(4), 348–351.

Bringsjord, S. (1997), *Abortion: A Dialogue*, Hackett, Indianapolis, IN.

Bringsjord, S. (1998), 'Chess is too easy', *Technology Review* **101**(2), 23–28.

Bringsjord, S. (1999), 'The zombie attack on the computational conception of mind', *Philosophy and Phenomenological Research* **59.1**, 41–69.

Bringsjord, S. (2000), 'A contrarian future for minds and machines', *Chronicle of Higher Education* p. B5. Reprinted in *The Education Digest* **66.6**: 31–33.

Bringsjord, S. (2001), 'Is it possible to build dramatically compelling interactive digital entertainment (in the form, e.g., of computer games)?', *Game Studies* **1**(1). This is the inaugural issue.
URL: *http://www.gamestudies.org*

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), 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. & Arkoudas, K. (forthcoming), The philosophical foundations of artificial intelligence, *in* K. Frankish & W. Ramsey, eds, 'The Cambridge Handbook of Artificial Intelligence', Cambridge University Press, Cambridge, UK.
URL: *http://kryten.mm.rpi.edu/sb_ka_fai_aihand.pdf*

Bringsjord, S., Arkoudas, K. & Bello, P. (2006), 'Toward a general logicist methodology for engineering ethically correct robots', *IEEE Intelligent Systems* **21**(4), 38–44.
URL: *http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf*

Bringsjord, S., Arkoudas, K., Clark, M., Shilliday, A., Taylor, J., Schimanski, B. & Yang, Y. (2007), Reporting on some logic-based machine reading research, *in* 'Proceedings of the 2007 AAAI Spring Symposium: Machine Reading (SS–07–06)', AAAI Press, Menlo Park, CA, pp. 23–28.
URL: *http://kryten.mm.rpi.edu/sb_ka_machinereading_ss07_012907.pdf*

Bringsjord, S. & Ferrucci, D. (1998*a*), 'Logic and artificial intelligence: Divorced, still married, separated...?', *Minds and Machines* **8**, 273–308.

Bringsjord, S. & Ferrucci, D. (1998*b*), 'Reply to Thayse and Glymour on logic and artificial intelligence', *Minds and Machines* **8**, 313–315.

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

Bringsjord, S., Khemlani, S., Arkoudas, K., McEvoy, C., Destefano, M. & Daigle, M. (2005), Advanced synthetic characters, evil, and E, *in* M. Al-Akaidi & A. E. Rhalibi, eds, 'Game-On 2005, 6th International Conference on Intelligent Games and Simulation', European Simulation Society, Ghent-Zwijnaarde, Belgium, pp. 31–39.
URL: *http://kryten.mm.rpi.edu/GameOnpaper.pdf*

Bringsjord, S., Noel, R. & Caporale, C. (2000), 'Animals, zombanimals, and the total Turing test: The essence of artificial intelligence', *Journal of Logic, Language, and Information* **9**, 397–418.

Bringsjord, S. & Schimanski, B. (2003), What is artificial intelligence? Psychometric AI as an answer, *in* 'Proceedings of the 18^{th} International Joint Conference on Artificial Intelligence (IJCAI–03)', Morgan Kaufmann, San Francisco, CA, pp. 887–893.

Bringsjord, S. & Zenzen, M. (1991), In defense of hyper-logicist AI, *in* 'IJCAI 91', Morgan Kaufman, Moutain View, CA, pp. 1066–1072.

Brooks, R. A., Breazeal, C., Marjanovic, M., Scassellati, B. & Williamson, M. M. (1999), 'The cog project: Building a humanoid robot', *Lecture Notes in Computer Science* **1562**, 52–87.

Cassimatis, N. (2006), 'Cognitive substrate for human-level intelligence', *AI Magazine* **27**(2), 71–82.

Charniak, E. (1993), *Statistical Language Learning*, MIT Press, Cambridge, MA.

Charniak, E. & McDermott, D. (1985), *Introduction to Artificial Intelligence*, Addison-Wesley, Reading, MA.

Chisholm, R. (1978), 'Is there a mind-body problem?', *Philosophic Exchange* **2**, 25–32.

Davis, M., Sigal, R. & Weyuker, E. (1994), *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science*, Academic Press, New York, NY.

Dennett, D. (1978), Conditions of personhood, *in* 'Brainstorms: Philosophical Essays on Mind and Psychology', Bradford Books, Montgomery, VT, pp. 267–285.

Dietrich, E. (1990), 'Computationalism', *Social Epistemology* **4**(2), 135–154.

Duncker, K. (1945), 'On problem solving', *Psychological Monographs* **58**(5 (Whole No. 270)).

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

Eden, A. (n.d.), 'Three paradigms of computer science', *Minds and Machines* **17**(2), 135–167.

Friedland, N., Allen, P., Matthews, G., Witbrock, M., Baxter, D., Curtis, J., Shepard, B., Miraglia, P., Angele, J., Staab, S., Moench, E., Oppermann, H., Wenke, D., Israel, D., Chaudhri, V., Porter, B., Barker, K., Fan, J., Chaw, S. Y., Yeh, P., Tecuci, D. & Clark, P. (2004), 'Project halo: Towards a digital aristotle', *AI Magazine* **25**(4), 29–47.

Fuchs, N. E., Schwertel, U. & Schwitter, R. (1999), Attempto Controlled English (ACE) Language Manual, Version 3.0, Technical Report 99.03, Department of Computer Science, University of Zurich, Zurich, Switzerland.

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

Glucksberg, S. (1964), 'Functional fixedness: Problem solution as a function of observing responses', *Psychonomic Science* **1**, 117–118.

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

Goldstein, E. B. (2005), *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience*, Wadsworth, Belmont, CA.

Halpern, J., Harper, R., Immerman, N., Kolaitis, P., Vardi, M. & Vianu, V. (2001), 'On the unusual effectiveness of logic in computer science', *The Bulletin of Symbolic Logic* **7**(2), 213–236.

Haugeland, J. (1985), *Artificial Intelligence: The Very Idea*, MIT Press, Cambridge, MA.

Hayes, P. (1978), The naïve physics manifesto, *in* D. Ritchie, ed., 'Expert Systems in the Microeletronics Age', Edinburgh University Press, Edinburgh, Scotland, pp. 242–270.

Hayes, P. J. (1985), The second naïve physics manifesto, *in* J. R. Hobbs & B. Moore, eds, 'Formal Theories of the Commonsense World', Ablex, pp. 1–36.

Ingham, M., Clark, M., Williams, B., Lockhart, T., Oyake, A. & Aljabri, A. (2001), Autonomous Sequencing and Model-based Fault Protection for Space Interferometry, *in* 'Proceedings of the Sixth International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS 2001)', Canadian Space Agency, Montreal, Canada. Proceedings published electronically on CD-ROM.

Johnson-Laird, P. (1997), 'Rules and illusions: A criticial study of Rips's *The Psychology of Proof*, *Minds and Machines* **7**(3), 387–407.

Joy, W. (2000), 'Why the Future Doesn't Need Us', *Wired* **8**(4).

Kautz, H. & Selman, B. (1996), Pushing the envelope: Planning, propositional logic, and stochastic search, *in* H. Shrobe & T. Senator, eds, 'Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference', AAAI Press, Menlo Park, California, pp. 1194–1201.
URL: *citeseer.ist.psu.edu/kautz96pushing.html*

Kautz, H. & Selman, B. (1999), Unifying SAT-based and graph-based planning, *in* J. Minker, ed., 'Workshop on Logic-Based Artificial Intelligence, Washington, DC, June 14– 16, 1999', Computer Science Department, University of Maryland, College Park, Maryland.
URL: *citeseer.ist.psu.edu/kautz99unifying.html*

Kuhse, H. & Singer, P., eds (2001), *Bioethics: An Anthology*, Blackwell, Oxford, UK.

Manzano, M. (1996), *Extensions of First Order Logic*, Cambridge University Press, Cambridge, UK.

McCarthy, J. (1959), Programs with common sense, *in* 'Proceedings of the Teddington Conference on the Mechanization of Thought Processes'. This is probably the first clear use of logic in the design of an AI system.
URL: *http://www-formal.stanford.edu/jmc*

McCarthy, J. & Hayes, P. J. (1969), Some philosophical problems from the standpoint of artificial intelligence, *in* B. Meltzer & D. Michie, eds, 'Machine Intelligence 4', Edinburgh University Press, pp. 463–502.

Minsky, M. (1967), *Computation: Finite and Infinite Machines*, Prentice-Hall, Englewood Cliffs, NJ.

Moravec, H. (1999), *Robot: Mere Machine to Transcendant Mind*, Oxford University Press, Oxford, UK.

Mueller, E. (2006), *Commonsense Reasoning*, Morgan Kaufmann, San Francisco, CA.

Newell, A. (1973), You can't play 20 questions with nature and win: Projective comments on the papers of this symposium, *in* W. Chase, ed., 'Visual Information Processing', New York: Academic Press, pp. 283–308.

Nilsson, N. (1991), 'Logic and Artificial Intelligence', *Artificial Intelligence* **47**, 31–56.

Nilsson, N. (1995), 'Eye on the prize', *AI Magazine* **16**(2), 9–16.

Nilsson, N. (2005), 'Human-level artificial intelligence? Be serious!', *AI Magazine* **26**(4), 68–75.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA.

Pollock, J. (1989), *How to Build a Person: A Prolegomenon*, MIT Press, Cambridge, MA.

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

Pollock, J. (2001), 'Defasible reasoning with variable degrees of justification', *Artificial Intelligence* **133**, 233–282.

Pollock, J. L. (1992), 'How to reason defeasibly', *Artificial Intelligence* **57**(1), 1–42.
URL: *citeseer.ist.psu.edu/pollock92how.html*

Rapaport, W. (1998), 'How minds can be computational systems', *Journal of Experimental and Theoretical Artificial Intelligence* **10**, 403–419.

Reiter, R. (2001), *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*, MIT Press, Cambridge, MA.

Russell, S. & Norvig, P. (2002), *Artificial Intelligence: A Modern Approach*, Prentice Hall, Upper Saddle River, NJ.

Seidler, V. (1986), *Kant, Respect, and Injustice: The Limits of Liberal Moral Theory*, Routledge and Kegan Paul, London, England.

Shapiro, S. (1995), 'Computationalism', *Minds and Machines* **5**(4), 517–524.

Shapiro, S. (2000), SNePS: A logic for natural language understanding and commonsense reasoning, *in* L. Iwanska & S. Shapiro, eds, 'Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language', AAAI Press/MIT Press, Menlo Park, CA, pp. 175–195.

Spivey, M. (2006), *The Continuity of Mind*, Oxford University Press, Oxford, UK.

Sun, R. (1999), 'Accounting for the computational basis of consciousness: A connectionist approach', *Consciousness and Cognition* **8**, 529–565.

Taylor, J., Shilliday, A. & Bringsjord, S. (2007), Provability-based semantic interoperability via translation graphs, *in* 'Advances in Conceptual Modeling; Lecture Notes in Computer Science', Vol. 4802, Springer, New York, NY, pp. 180–189.
URL: *http://kryten.mm.rpi.edu/jt_as_sb_PBSITG_crc.pdf*

Turing, A. (1950), 'Computing machinery and intelligence', *Mind* **LIX (59)**(236), 433–460.

Turner, R. (1984), *Logics for Artificial Intelligence*, Ellis Horwood, West Sussex, England.

Voronkov, A. (1995), 'The anatomy of vampire: Implementing bottom-up procedures with code trees', *Journal of Automated Reasoning* **15**(2).

Wiles, A. (1995), 'Modular elliptic curves and fermat's last theorem', *Annals of Mathematics* **141**(3), 443–551.

Wiles, A. & Taylor, R. (1995), 'Ring-theoretic properties of certain Hecke algebras', *Annals of Mathematics* **141**(3), 553–572.

Williams, B. C., Ingham, M. E., Chung, S. H. & Elliott, P. H. (2003), Model-Based Programming of Intelligent Embedded Systems and Robotic Space Explorers, *in* 'Proceeding of the IEEE: Special Issue on Modeling and Design of Embedded Software', Vol. 91, pp. 212–237.

Wos, L., Overbeek, R., e. Lusk & Boyle, J. (1992), *Automated Reasoning: Introduction and Applications*, McGraw Hill, New York, NY.

Yang, Y. & Bringsjord, S. (2003), 'Newell's program, like Hilbert's, is dead; let's move on', *Behavioral and Brain Sciences* **26**(5), 627.