Sophisticated Knowledge Representation and Reasoning Requires Philosophy

Selmer Bringsjord and Micah Clark and Joshua Taylor
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
{selmer, clarkm5, tayloj}@rpi.edu

April 15, 2009

Abstract

Knowledge Representation and Reasoning (KR&R) is based on the idea that propositional content can be rigorously represented in formal languages long the province of logic, in such a way that these representations can be productively reasoned over by humans and machines; and that this reasoning can be used to produce knowledge-based systems (KBSs). As such, KR&R is a discipline conventionally regarded to range across parts of artificial intelligence (AI), computer science, and especially logic. This standard view of KR&R's participating fields is correct — but dangerously incomplete. The view is incomplete because, as we explain herein, sophisticated KR&R must rely heavily upon *philosophy*. Encapsulated, the reason is actually quite straightforward: Sophisticated KR&R must include the representation of not only simple properties, but also concepts that are routine in the formal sciences (theoretical computer science, mathematics, logic, game theory, etc.), and everyday socio-cognitive concepts like mendacity, deception, betrayal, and evil. Because in KR&R the representation of such concepts must be rigorous in order to enable machine reasoning (e.g., machine-generated and machine-checked proofs that a is lying to b) over them, philosophy, devoted as it is in no small part to supplying analyses of such concepts, is a crucial partner in the overall enterprise. To put the point another way: When the knowledge to be represented is such as to require lengthy formulas in expressive formal languages for that representation, philosophy must be involved in the game. In addition, insofar as the advance of KR&R must allow formalisms and processes for representing and reasoning over *visual* propositional content, philosophy will be a key contributor into the future.

1 What is Knowledge Representation and Reasoning?

What is Knowledge Representation and Reasoning (KR&R)? Alas, a thorough account would require a book (e.g., see the excellent Brachman & Levesque 2004), or at least a dedicated, full-length paper (e.g., see Bringsjord & Yang 2003), so we shall have to make do with something simpler. Since most readers are likely to have an intuitive grasp of the essence of KR&R, our simple account should suffice. The interesting thing is that this simple account itself makes reference to some of the foundational distinctions in the field of philosophy. These distinctions also play a central role in artificial intelligence (AI) and computer science.

To begin with the first distinction, in KR&R we identify knowledge with knowledge *that* such-and-such holds (possibly to a degree), rather than knowing *how*. If you ask an expert tennis player how he manages to serve a ball at 130 miles-per-hour on his first serve, and then serve a safer, more-topspin serve on his second should the first be out, you may well receive a confession that, if truth be told, this athlete can't really tell you. He just does it; he does something he has been doing since his youth. Yet, there is no denying that he knows how to serve. In contrast, the knowledge in KR&R must be expressible in declarative statements. For example, our tennis player knows that if his first serve lands outside the service box, it's not in play. He thus knows a *proposition*, conditional in form. It is this brand of declarative statement that KR&R is concerned with.

At some point earlier, our tennis player did not not know the rules of tennis. Suppose that for his first lesson, this person walked out onto a tennis court for the first time in his life, but that he had previously glimpsed some tennis being played on television. We can thus imagine that before the first lesson began, our student *believed* that a serving player in tennis is allowed three chances to serve the ball legally. This belief would have of course been incorrect, as only two chances are permitted. Nonetheless, this would be a case of a second attitude directed toward a proposition. The first attitude was knowledge, the second mere belief.

Knowledge-based systems (KBSs), then, can be viewed as computational systems whose actions through time are a function of what they know and believe. Knowing that his first serve has landed outside the service box on the other side of the net from him, our (educated) tennis player performs the action of serving for a second time, and as such performs as a KBS. A fully general and formal account of KBSs can be found elsewhere (e.g., where the first is brief, and the second more extensive, see Sun & Bringsjord 2009, Bringsjord 2008). There are numerous algorithms designed to compute the functions in question, but in the present paper we shall be able to rest content with reference to but a few of them.

2 The Nature of Philosophy-less KR&R

In this section, after introducing the basic machinery of elementary extensional and intensional logic for purposes of KR&R carried out in the service of building KBSs, we present our characterization of the dividing line between philosophy-less KR&R versus philosophy-infused KR&R.

2.1 Overview of Elementary Extensional and Intensional Logic for KR&R

Propositions can be represented as formulas in formal languages. For example, in the present case, we might use a simple formula from the formal language $\mathcal{L}_{PC}$ of the *propositional calculus*, which allows propositions to be expressed as either specific propositional variables such as p_1, p_2, ... (or mnemonic replacements thereof, e.g. h for p_2 when we want to represent the proposition that John

is happy), or as formulas built from the p_i and the familiar Boolean connectives: $\neg\phi$ ("not ϕ"), $\phi \vee \psi$ ("ϕ or ψ"), $\phi \wedge \psi$ ("ϕ and ψ"), $\phi \rightarrow \psi$ ("if ϕ then ψ"), $\phi \leftrightarrow \psi$ ("ϕ if and only if ψ").[1] For example, letting *out* represent the proposition that the ball lands outside the service box, and *play* that the ball is in play, $out \rightarrow \neg play$ represents the above conditional. If a knowledge-based system knows *out* and this conditional, it would of course be able to infer $\neg play$, that the ball is not in play. Its reasoning would be deductive in nature, using the well-known rule of inference *modus ponens*. To use the standard provability relation $\vdash_X$ in knowledge-based/logic-based AI and cognitive science, where the subscript X is a placeholder for some particular proof calculus, we would write

$$\{out, out \rightarrow \neg play\} \vdash_X \neg play$$

to express the fact that the ball's being out of play can be proved from the the formulas to the left of $\vdash$. So here we have a (painfully!) simple case of a KBS, powered by KR&R, in action.

Some discussion concerning candidate proof calculi for X is necessary. Due to lack of space, we must leave aside specification of each of the myriad possibilities, in favor of a streamlined approach. This approach is based upon the incontestable fact that there clearly is a *generic* conception of fairly careful deductive reasoning according to which some lines of linear, step-by-step inference can be accepted as establishing their conclusions with the force of proof, even though detailed definition of particular calculus X, and use thereof, are absent. This streamlined approach works because the step-by-step sequence is such that each step from some set $\{\phi_1, \ldots, \phi_n\}$ to some inferred-to formula ψ can be quickly seen, with a small amount of mental energy, to be such that *it is impossible that each ϕ_i hold, while ψ does not.*[2] What follows is an example of such a sequence, couched in natural language; the sequence establishes with the force of proof that from "Everyone likes anyone who likes someone" and "Albert likes Brian" it can be inferred that "Everyone likes Brian." (Most people see that it can be inferred from this pair of statements that "Everyone likes Albert," but are initially surprised that "Everyone likes Brian" can be derived in a bit of a recursive twist.)

1	Everyone likes anyone who loves someone.	assumption
2	Albert likes Brian.	assumption
3	Albert likes someone.	from 2
4	Everyone likes Albert.	from 1, 3
5	Brian likes Albert.	from 4
6	Brian likes someone.	from 5
7	Everyone likes Brian.	from 1, 6

Despite opting for what we have called a streamlined approach to provability in the present paper, we would be remiss if we failed to point out that the format that best coincides with how professionals in those fields based on deductive reasoning actually construct proofs and disproofs is clearly something quite like "Fitch-style" *natural deduction* (Fitch 1952), with which many readers will be acquainted. In this kind of proof calculus, which aligns with the deductions written in relevant professional books and papers (in computer science, mathematics, logic, etc.), each of the truth-functional connectives, and the quantifiers (see below), has a pair of corresponding inference rules, one for introducing the connective, and one for eliminating the connective. One concrete

[1] Nice coverage of the propositional calculus is provided in (Barwise & Etchemendy 1999).

[2] For a discussion of the relationship between this concept of intuitive formal validity, provability in a deductive calculus, and the corresponding semantic consequence of ψ following deductively from $\{\phi_1, \ldots, \phi_n\}$ provided that there is no model in which all the ϕ_i are true while ψ is false, see (Kreisel 1967).

possibility for a natural-deduction calculus is the "human-friendly" one known as $\mathcal{F}$, set out in (Barwise & Etchemendy 1999). Another possibility is the natural-deduction style proof calculus used in the Athena system (Arkoudas n.d.). We make use of the Athena system below, but don't use or specify its proof calculus.

The propositional calculus is rather inexpressive. Most of what we know cannot be represented in $\mathcal{L}_{PC}$ without an unacceptably large loss of information. For example, from the statement "Albert likes Brian," we can infer that "Albert likes someone." We might attempt to represent these two statements, respectively, in $\mathcal{L}_{PC}$ as, say, A and $A_{someone}$. Unfortunately, this representation is defective, for the simple reason that by no acceptable rule of deductive inference can $A_{someone}$ be deduced from A. The problem is that $\mathcal{L}_{PC}$ cannot express quantification in formulas such as "Albert likes someone," and so lacks inference rules such as *existential introduction* (which formally obtains the intuitive result above).[3] The machinery of quantification (in one simple form), and this particular rule of inference, are part of *first-order logic*, whose formal language is $\mathcal{L}_{FOL}$. The alphabet for this language reflects an increase in that for the propositional calculus, to include:

identity	$=$	the identity or equality symbol;
connectives	$\neg$, $\vee$, ...	now familiar to you, same as in $\mathcal{L}_{PC}$;
variables	x, y, ...	variables ranging over objects;
constants	c_1, c_2, ...	you can think of these as proper names for objects;
relations	R, G, ...	used to denote properties, e.g., W for *being a widow*;
functions	f_1, f_2, ...	used to refer to functions;
quantifiers	$\exists$, $\forall$	$\exists$ says "for some ...," $\forall$ says "for every ..."

Predictable *formation rules* are introduced to allow one to represent propositions like the "Everyone likes anyone who likes someone" one above. In the interests of space, the grammar in question is omitted, and we simply show "in action" the kind of formulas that can be produced by this grammar, by referring back to the Albert-Brian example above. We do so by presenting here the English-based sequence from above in which natural language is replacement by suitable formulas from $\mathcal{L}_{FOL}$. Recall that this sequence, in keeping with the streamlined approach to presenting provability herein, is something that qualifies as an outright proof. In addition, the reader should rest assured that automated theorem proving technology of today can instantly find a proof of line 7 from lines 1 and 2.[4]

[3]The rule for *existential introduction* is expressed below, wherein $\phi(a)$ is a formula in which a appears as a constant, and $\phi(a/x)$ is that formula changed only by the replacement of a with the variable x.

$$\frac{\phi(a)}{\exists x\ \phi(a/x)}$$

[4]E.g., Otter, given these two lines as input, produces the following proof:

```
---------------- PROOF ----------------
1 [] -Likes(x,y)|Likes(z,x).
2 [] -Likes($c1,b).
3 [] Likes(a,b).
4 [hyper,3,1] Likes(x,a).
5 [hyper,4,1] Likes(x,y).
6 [binary,5.1,2.1] $F.
------------ end of proof -------------
```

1	$\forall x\ \forall y\ [(\exists z\ \mathit{Likes}(x, z)) \rightarrow \mathit{Likes}(y, x)]$	assumption
2	$\mathit{Likes}(a, b)$	assumption
3	$\exists x\ \mathit{Likes}(a, x)$	from 2
4	$\forall x\ \mathit{Likes}(x, a)$	from 1, 3
5	$\mathit{Likes}(b, a)$	from 4
6	$\exists x\ \mathit{Likes}(b, x)$	from 5
7	$\forall x\ \mathit{Likes}(x, b)$	from 1, 6

Recall that we referred above to natural-deduction proof calculi, in which each connective and quantifier is associated with a pair of inference rules, one for introducing and one for eliminating. Were this calculus to be applied to the sequence immediately above, the rule of inference for eliminating the universal quantifier would sanction moving from

$$\forall x\ \forall y\ [(\exists z\ \mathit{Likes}(x, z)) \rightarrow \mathit{Likes}(y, x)]$$

to — where a is substituted for x — the following:

$$\forall y\ [(\exists z\ \mathit{Likes}(a, z)) \rightarrow \mathit{Likes}(y, a)].$$

The reader is invited to see, with help from the works cite above (e.g., Barwise & Etchemendy 1999), how other such rules can be used to construct a fully formal proof out of the sequence.

There are languages for knowledge representation that fall between $\mathcal{L}_{PC}$ and $\mathcal{L}_{FOL}$, and for most part these are the languages that anchor the brand of KR&R supporting the Semantic Web.[5] These languages are more expressive than $\mathcal{L}_{PC}$, the language of the propositional calculus, but less expressive than the language $\mathcal{L}_{FOL}$ of first-order logic. And these languages are associated with their own proof calculi. These languages are generally those associated with *description logics* (Baader, Calvanese & McGuinness 2003). We don't have the space needed for a full exposition of such logics, but fortunately they can in general be quickly characterized with reference to the ingredients that compose the propositional calculus and first-order logic. We shall refer to these logics as *point-k* logical systems. A particular system in the class will be named later when k is set to some natural number. The ins and outs of how the natural numbers work as indices is based on an idiosyncratic but straightforward table invented for ease of reference by Bringsjord to keep straight decidability theorems for the main logical systems standardly discussed in such contexts.[6] Using this table, here is what pins down point-two (i.e., point-k where $k = 2$) logic:

	Monadic Relations	Dyadic Relations	Triadic Relations
None		•	•
One			
Unlimited	•		

The characterization of such systems is simple. To produce such logics, we simply begin by restricting the alphabet of $\mathcal{L}_{FOL}$ in various ways. As an example, we might insist that no triadic relation be allowed, that no dyadic relations be allowed, but that any number of monadic relations be allowed (as in the permutation shown in the table immediately above). The logical system with such a language (the language $\mathcal{L}_{P2}$) is *point-two logic*, or *monadic first-order logic*. Point-two logic will be otherwise just like FOL. A triadic relation is one that allows a relationship between

[5]For an overview of Semantic Web, see (Berners-Lee, Hendler & Lassila 2001).

[6]The core theorems can be found in (Boolos, Burgess & Jeffrey 2003, ch. 21).

three objects to be expressed. For example, the relation (B, let's say) of a natural number n being between two distinct natural numbers m and j is a triadic one; and here would be a truth regarding the natural numbers that involves this triadic relation:

$$\forall x\ \forall y\ \forall z\ (B(x, y, z) \rightarrow x \neq z).$$

This truth cannot be expressed in monadic FOL. Naturally, dyadic relations would range over two objects, and monadic relations over but one object.

Why would those logical systems between the propositional calculus and full first-order logic be so central to KR&R? The reason point-k logical systems are interesting and useful pertains to an aspect of logical systems that we have yet to discuss: namely, meta-properties of such systems. Important meta-properties of such systems include the meta-property of *decidability*. A logical system is decidable just in case there is an algorithm for determining whether or not a well-formed formula in the language in question is a theorem. Of course, assuming the the Church-Turing Thesis is true, the existence of such an algorithm guarantees that there is a computer program that can determine, given as input a $\phi \in \mathcal{L}_{P2}$, whether ϕ is a theorem.[7] While the propositional calculus is decidable, the predicate calculus is not. However, point-two logic is decidable. From the standpoint of KR&R, this is thought by many to be quite desirable. The reason is clear. It is that queries against knowledge-bases populated by formulas expressed in $\mathcal{L}_{P2}$ can always (eventually) be answered, that is, where Φ is such a knowledge-base, queries of the form

$$\Phi \vdash \phi\ ?$$

can, given enough time and working memory, always be answered by a standard computing machine.

A fact that beginning students of KR&R and logic often find quite surprising is that the moment even one dyadic relation is allowed into a logic otherwise like FOL, that logic becomes undecidable; the proofs are actually quite simple. However, if one allows another dimension of parameterization into the picture, namely, the number of quantifiers allowed in formulas, one can allow an expansion on the relation side and yet still preserve decidability, as long as k is quite small. We must leave details aside.

The final point that must be made in this section is that there are many, many (actually, an infinite number of) logical systems more expressive that first-order logic. We mention just two examples. The first is in the space of extensional logics, the second in the space of intensional logics.

The first example is *second-order logic* (SOL), which allows quantification over functions and relations, a phenomenon that routinely occurs in natural language. The formal language in question, $\mathcal{L}_{SOL}$, includes variables for functions and relations. For instance, it seems quite plausible not only that if John is the very same thing as the father of Bill, John and the father of Bill either both have or both lack the property of being obese, but more generally that these two entities are such that every relation is one they share or lack. The general principle operative would be that two things are one and the same just in case every attribute is one they either share or lack; this principle is known as *Leibniz' Law*. In SOL we can formalize this law as:

$$\text{(LL)} \quad \forall x\ \forall y\ (x = y \leftrightarrow \forall X(Xx \rightarrow Xy)).$$

[7] Church-Turing Thesis states that a function f is effectively computable if and only if f is Turing-computable. For a discussion, see (Bringsjord & Arkoudas 2006).

Note that in (LL) the variable X ranges over relations, whereas x and y range over individual objects in the domain. Humans find it easy enough to discuss scenarios in which attributes themselves have properties, but we leave aside *third-order logic* and beyond.

Our second example is a simple *epistemic logic*, in which, to the propositional calculus, we add an operator $\mathbf{K}$ for "knows," which allows us to represent such propositions as that Albert knows that Brian knows that p is the case, as in,

$$\mathbf{K}_a\mathbf{K}_b\ p$$

from which we can deduce, using an axiom that is standard in such logics (viz., that if an agent knows ϕ, ϕ is true), that in fact Brian knows that p. A recent exploration of the applicability KR&R based on advanced epistemic logic can be found in (Arkoudas & Bringsjord 2005).

We have now reached the point at which we can discuss the dividing line between philosophy-less and philosophy-infused KR&R.

2.2 A Proposed Dividing Line Between Philosophy-less KR&R and Philosophy-powered KR&R

The idea for a dividing line is really quite straightforward: KR&R can be productively pursued, and KBSs built, in the complete absence of philosophy — but only as long as the information represented and reasoned over is not in the realm of the formal sciences, nor in the realm of everyday sophisticated human socio-cognition. On the other hand, philosophy will need to be part and parcel of KR&R when that which is to be represented and reasoned over involves these realms. We can put this position in the form of a claim that makes reference to the expressiveness of formal languages of the sort canvassed above:

Claim $\mathcal{C}$ (Regarding the Relationship Between Philosophy and KR&R)

> KR&R that represents propositional content in formulas of a formal language less expressive than that used in full first-order logic ($\mathcal{L}_{FOL}$) is unable to represent and reason over propositions containing concepts routinely used in the formal sciences, and in everyday human socio-cognition, and as such, such KR&R will have no need for the field of philosophy. Moreover, to engineer KBSs able to represent and reason over the more demanding phenomena in these domains will require a contribution from philosophy, and will specifically require:
>
> 1. formulas in $\mathcal{L}_{FOL}$ that, once rewritten so that all quantifiers appear in a leftmost sequence in such formulas (i.e., once rewritten in *prenex normal form*[8]), are irreducibly populated by at least five non-vacuous quantifiers, must be allowed; and
> 2. formulas in formal languages that are more expressive than $\mathcal{L}_{FOL}$ must be allowed.

We turn now to concretizing this claim by discussing some challenges KR&R can meet only if both philosophy and the associated languages are employed.

[8]There are straightforward algorithms for adapting arbitrary formulas so as to produce such "front-loaded" versions of them. E.g., see (Boolos et al. 2003).

3 The Need for Philosophy

If $\mathcal{C}$ is true, then it should be easy enough to see the need for philosophy and the associated representation and reasoning schemes by considering some examples from the relevant domains, and we turn to such consideration now. We first look at an example from the formal sciences, and then one from socio-cognition. In both cases, the KR&R in question has been pursued, and is in fact currently still underway, in our own laboratory.

3.1 KR&R and the Formal Sciences

KR&R allows for, indeed in large measure exists to enable, the issuing of queries against knowledge-bases. As such, there is clearly a vantage point from which to see the applicability of KR&R within the formal sciences, that is, within such fields as formal logic, game theory, probability theory, the various sub-fields of mathematics (e.g., number theory and topology), and so on. In fact, the part of formal logic known as *mathematical logic*, aptly and (given our present purposes) conveniently sometimes also called 'meta-mathematics' (e.g., see Kleene 1952), explicitly provides this vantage point, as we now explain, in brief. After this presentation, we explain why the above conjecture's claim about the formal sciences seems to be quite plausible.

The first step is to view activity in the formal sciences from the standpoint of *theories*. By 'theory' here is meant something purely formal, not anything like, say, the "theory" of evolution, which is usually disturbingly informal.[9] In the formal sense, a theory $\mathcal{T}_\Phi$ is a set of formulas deducible from a set of axioms Φ; more precisely, given Φ, the corresponding theory $\mathcal{T}_\Phi$ is

$$\{\phi \in \mathcal{L} : \Phi \vdash \phi\}.$$

We say in this case that Φ are the axioms *of* the theory. Note that there is a background formal language $\mathcal{L}$ from which the relevant formulas are drawn.

But why might it be reasonable to regard *all* research in the formal sciences to revolve around theories? We don't have the space to fully articulate and defend this view, and hence rest content to convey the basic idea, which is straightforward. That idea is this: work in a formal science S can be idealized as the attempt to ascertain whether or not propositions of interest follow deductively from a core set of axioms for S; that is, whether these propositions of interest are indeed part of the theory that arises from the axioms for S. In this scheme, probability theory, game theory, mathematics, and so on — each consists in the attempt to increasingly pin down the theory determined by the core axioms in question.

For an example simple enough to present here, let's consider a fragment of mathematics, namely, elementary arithmetic. Specifically, we consider the theory of arithmetic known as "Peano Arithmetic," or simply as **PA**.[10] The axioms of **PA** are the sentences 1–6 (where s is the successor function, and the other symbols are interpreted in keeping with grade-school arithmetic; e.g., $\times$ is ordinary multiplication), and any sentence in the universal closure of the Induction Schema, 7.[11]

[9] E.g., even book-length presentations of mutation and natural selection routinely fail to provide any formal, rigorous statement of the theory; see, for instance, (Dennett 1995).

[10] Nice coverage is provided in (Smith 2007).

[11] Where $\phi(x)$ is a first-order formula with x, and possibly others, as free variables, the universal closure of $\phi(x)$ binds the variables other than x to universal quantifiers. E.g., the universal closure of $B(x, y, z)$ would be $\forall y \forall z B(x, y, z)$.

1. $\forall x\ (0 \neq s(x))$
2. $\forall x\ \forall y\ (s(x) = s(y) \rightarrow x = y)$
3. $\forall x\ (x + 0 = x)$
4. $\forall x\ \forall y\ (x + s(y) = s(x + y))$
5. $\forall x\ (x \times 0 = 0)$
6. $\forall x\ \forall y\ (x \times s(y) = (x \times y) + x)$
7. $[\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(s(x))] \rightarrow \forall x \phi(x)$

Given this machinery, the part of mathematics known as arithmetic can be viewed as an attempt to increasingly pin down $\mathcal{T}_{\mathbf{PA}}$. And now it should be clear, in turn, why KR&R can be regarded to have direct applicability. One reason is that **PA** can be thought of as a knowledge-base, and the attempt to make more and more progress figuring out what is and isn't in the theory $\mathcal{T}_{\mathbf{PA}}$ can be viewed as the attempt to ascertain whether or not, for various formulae generable from the language of arithmetic, say ϕ,

$$\mathbf{PA} \vdash \phi.$$

We could thus view "progress in the field of arithmetic" to be the answering of such questions as whether or not it's true that 29 plus 0 equals 29, and whether or not it's true that 3,000 times 0 equals 0, and so on.

Why, in light of the foregoing and other material, is the claim $\mathcal{C}$ plausible? The answer, put non-technically, is quite straightforward; and comes in three parts, to wit:

- Courtesy of Gödel's first incompleteness theorem, we know that there are truths about arithmetic that cannot be proved from **PA**.[12]
- Thanks to additional formal work, we know that some of these truths can nonetheless be proved.[13] Let's call this set $\widehat{G}$.
- The general nature of the representation and reasoning needed to establish the truths in $\widehat{G}$ is an open question in KR&R and philosophy, but it is clear that full first-order logic is required (for the simple reason that formulas in $\widehat{G}$ require $\mathcal{L}_{FOL}$ to be expressed).

So here we have an ongoing investigation (chronicled to this point in Smith 2007) in the intersection of the formal sciences and KR&R that both intersects with philosophy, and is consistent with claim $\mathcal{C}$.

3.2 KR&R and Socio-Cognition

We now give our second example of $\mathcal{C}$ "in action" by considering a specific concept in the sphere of socio-cognition. The concept is mendacity. We shall show that careful KR&R in this area necessitates both philosophy, and very expressive formal languages for knowledge representation.

3.2.1 Mendacity

We introduce the topic of mendacity in connection with KR&R by asking you to consider a confessedly idealized scenario.[14] The scenario involves a superficial and implausible concept of lying,

[12] A succinct proof of this theorem can be found in (Ebbinghaus, Flum & Thomas 1984).

[13] One example is Goodstein's Theorem. This theorem, and others in the relevant class, are too complex to cover in the present paper. See (Smith 2007) for full nice coverage.

[14] This being a paper on philosophy-empowered KR&R, as opposed to deception and counterdeception, an idealized example is appropriate. We have not the space to exhibit the power of the brand of KR&R we are championing in connection with real-world deception and counterdeception. Motivated readers can confirm that our techniques are applicable in real-world scenarios by consulting (Bennett & Waltz 2007).

but as a warm-up to the genuine article, we indicate how the machinery of unsophisticated KR&R can be brought to bear to provide a solution to the scenario. Afterward, we present a philosophically inspired, plausible definition of lying, and demonstrate how a sophisticated, philosophically informed, KR&R system can be used to distinguish lies and liars from honesty and the honest. Without further prelude, we ask you to consider the following scenario.

> You have been sent to the war-torn and faction-plagued planet of Raq. Your mission is to broker peace between the warring Larpal and Tarsal factions. In a pre-trip briefing, you were informed that the Larpals are sending one delegate to the negotiations, and the Tarsals are sending a pair. You were also warned that Larpals are liars, i.e., whatever they say is false, while Tarsals are not, i.e., whatever they say is true. Upon arrival, you are met by the three alien delegates. Suddenly, you realize that though the aliens know whom among them are Larpals, and whom are Tarsals, you do not. So, you ask the first alien, "To which faction do you belong?" In response, the first alien murmurs something you can't decipher. Seeing your look of puzzlement, the second alien says to you, "It said that it was a Larpal." Then, with a cautionary wave of an appendage and an accusatory glance at the second alien, the third alien says to you "That was a lie!" Whom among the three aliens can you trust?

Resolution of the Larpals & Tarsals scenario, at least in its present form, requires no more sophistication than $\mathcal{L}_{PC}$ and reasoning therewith. The scenario is recast into $\mathcal{L}_{PC}$ by, say, representing the three aliens with three constants, their factional membership (Larpal or Tarsal) as mutually exclusive properties, and their assertions as conditional formulas. In Figure 1, we show the scenario thus represented, and automatically solved, in Athena (Arkoudas n.d.), a KR&R system based on multi-sorted, first-order logic, and integrated with both the Vampire theorem prover and the Paradox model finder. The solution to this scenario, expressed in English, is as follows: The second alien is either a Larpal or a Tarsal. If it is a Tarsal, then truly the first alien said that it was a Larpal. Yet, if the first alien said that it was a Larpal, then it told the truth because a Tarsal would not lie and say it was a Larpal, but in so telling the truth, the first alien has distinguished itself as a Tarsal — a contradiction! Ergo, the second alien cannot be a Tarsal; it is a Larpal. Therefore, the first and third aliens are Tarsals, and thus trustworthy.

Though the Larpals & Tarsals scenario nicely illustrates unsophisticated KR&R in action, the fact of the matter is that the concept of lying used in the scenario is, as we have already indicated, simple-minded. In real life, the idea that liars' propositional claims are always materially false is, well, silly. We might reserve such phrases as *habitual liar*, or *pathological liar* for such beings, but in the real-world, even pathological liars sometimes assert true propositions, if only by accident. Likewise, it is utterly unrealistic to expect honest agents to be infallible, i.e., to expect that their assertions are always materially true, because honest agents, nevertheless, may state false propositions out of ignorance, or error in belief.

To say that an agent is a *liar* presupposes that one has at hand an account of what it is to lie — yet no such account was set out, let alone included in the knowledge-base constructed above for the Larpals & Tarsals scenario. Any reasonable account of lying must include not just what an agent does — the *actus reus* of lying — but also what the agent believes and intends. Mendacity and less egregious forms of deception are consummate only when an agent acts with the *mens rea* to deceive, i.e., when an agent acts intending others to hold beliefs that are contrary to what the agent believes to be true. To illustrate, assume that Amy is asked in geography class to name the state capital of

larpals-and-tarsals.ath:

```
(domain Alien)             # there is a domain of Aliens.
(declare (A1 A2 A3) Alien) # A1, A2, and A3 are Aliens.

# Larpal and Tarsal are properties of Aliens.
(declare (larpal tarsal) (-> (Alien) Boolean))

# each Alien is either a Larpal or a Tarsal, but not both.
(assert (and (or (larpal A1) (tarsal A1))
             (not (and (larpal A1) (tarsal A1)))))
(assert (and (or (larpal A2) (tarsal A2))
             (not (and (larpal A2) (tarsal A2)))))
(assert (and (or (larpal A3) (tarsal A3))
             (not (and (larpal A3) (tarsal A3)))))

# among A1, A2 & A3 are one Larpal and two Tarsals.
(assert (iff (larpal A1) (and (tarsal A2) (tarsal A3))))
(assert (iff (larpal A2) (and (tarsal A1) (tarsal A3))))
(assert (iff (larpal A3) (and (tarsal A1) (tarsal A2))))

# if A3 is a Larpal, then A2 is a Tarsal...
(assert (if (larpal A3) (tarsal A2)))

# and if A2 is a Tarsal, then A1 said that it is a Larpal.
(assert (if (tarsal A2) (larpal A1)))

# but if A1 said that it is a Larpal, then it is a Tarsal!
(assert (if (larpal A1) (tarsal A1)))
```

Athena transcript:

```
>(load-file "larpals-and-tarsals.ath")
...
>(!prove (and (tarsal A1) (larpal A2) (tarsal A3)))

Theorem: (and (tarsal A1) (larpal A2) (tarsal A3))
```

Figure 1: Larpals & Tarsals scenario resolved in Athena.

California. Amy, erroneously believing that Los Angeles is the capital of California, answers with Los Angeles. Though Amy's answer is materially false, we would not ordinarily accuse Amy of lying because she has answered faithfully according to her belief — her statement was truthfully made, though it was not factually true. However, had Amy known that Sacramento is the capital of California, but answered Los Angeles intending to give a false impression of at least her own mind, then, indeed, she would have been lying. Now assume that Bob is helping Carl, a fugitive, flee from the police. The two agree that Carl should begin a new life in Canada, and then part ways. Later, when the police question Bob about Carl's whereabouts, Bob, intending to misdirect the police, tells them that Carl has gone to Mexico. Yet unbeknownst to Bob, Carl has changed his mind (and destination), moving to Mexico instead of Canada. Thus, Bob's statement to the police is materially true, though we would normally say that Bob lied because he believed that what he said was false, and said it intending to deceive — though factually true, his statement was falsely made. As these examples illustrate, the *mens rea* for lying and deception depends on the

relationship between an agent's beliefs and the beliefs the agent intends for others.

Now, drawing upon philosophy, we set out a plausible definition of lying. We present this definition first informally, and then formally, using the language of a logical system, viz., the socio-cognitive calculus ($\mathcal{SCC}$). Once lying is thus defined, we explain, by revisiting the Larpals & Tarsals scenario, how a highly sophisticated KR&R systems can prove, say, that an agent is a liar, or that, one agent has lied to another.

Philosophy has a long tradition of contemplating the nature of mendacity and positing definitions thereof (a tradition going back at least to St. Augustine). For exposition, we adopt Chisholm & Feehan's (1977) account of lying — a seminal work in the study of mendacity and deception. Using L and D to represent correspondingly the speaker (i.e., the *liar*) and the hearer (i.e., the would-be *deceived*), we paraphrase below Chisholm & Feehan's (ibid., p. 152 D3, D2) definitions of *lying* and the supporting act of *asserting*.

> *L lies to D* $=_{\text{df}}$ There is a proposition p such that (i) either L believes that p is not true or L believes that p is false and (ii) L asserts p to D.[15]

> *L asserts p to D* $=_{\text{df}}$ L states p to D and does so under conditions which, he believes, justify D in believing that he, L, accepts p.[16]

Chisholm & Feehan's (ibid.) conception of lying is that of promise breaking. Assertions, unlike non-solemn (e.g., ironic, humorous, or playful) statements, proffer an implicit social concord: one that offers to reveal to the hearer the mind of the speaker. In truthful, forthright communication, the speaker fulfills the promise and obligation of this concord. In lying, the speaker proffers the concord in bad faith: the speaker intends not, and does not, fulfill the obligation to reveal his/her true mind, but instead reveals a pretense of belief. In this way, lying "is essentially a breach of faith" (ibid., p. 153).

The above is, of course, a highly condensed presentation of Chisholm & Feehan's (ibid.) work, and there are various nuanced philosophical facets to it (for full-fledged analysis of these definitions, see Clark 2009).[17] Yet, even in condensed form, it is evident that the concepts of *lying* and *asserting* depend on agents' temporally coupled beliefs and actions. Thus, formal definition of these concepts requires the use of highly expressive languages for KR&R: ones that can represent, and allow reasoning over, the beliefs and actions of agents through time.

To formally define lying and asserting, we employ the *socio-cognitive calculus* ($\mathcal{SCC}$). The $\mathcal{SCC}$ (Arkoudas & Bringsjord 2008) is a KR&R system for representing, and reasoning over, events and causation, and perceptual, doxastic, and epistemic states (it integrates ideas from the event calculus

[15] Whether the disjunction, "L believes that p is not true or L believes that p is false," is redundant depends on how one formally represents beliefs about propositions. In the $\mathcal{SCC}$, the formal system we use to define lying precisely, there is no representational difference between believing a proposition to be not true and believing the proposition to be false. However, in other formal systems there may be a representational and logical distinction between the two.

[16] Linguistic convention dictates that statements are assertions by default, i.e., when cues to the contrary, such as irony and humor, are absent (ibid., p. 151). The conditions mentioned in the definition of *asserting* are meant to exclude situations where the speaker believes that he/she will be understood as making a non-solemn statement — for example, when the speaker makes a joke, uses a metaphor, or conveys by other indicator (e.g., a wink or a nod) that he/she is not intending to be taken seriously (ibid., p. 152).

[17] E.g.: (i) "L believes that p is false" is an expression of a higher-order belief — this belief cannot be attained unless L has the concept of something *being false* (ibid., p. 146); (ii) L's beliefs, and L's beliefs about D's beliefs, are both occurrent and defeasible (ibid., p. 151) — the latter, defeasibility, indicates that *justifications* ought to be treated as first-class entities within a formal system.

and multi-agent epistemic logic). The $\mathcal{SCC}$ is an extension to the Athena system (Arkoudas n.d.), providing, among other things, operators for perception, belief, knowledge, and common knowledge. The signature and grammar of the $\mathcal{SCC}$ is shown following. Since some readers may not be familiar with the concept of a signature, we note that it is simply a set of announcements about the categories of objects that will be involved, and about the functions that will be used to talk about these objects. Thus it will be noted that immediately below, the signature in question includes the specific announcements that one category includes agents, and that *happens* is a function that maps a pair composed of an *event* and a *moment*, and returns `true` or `false` (depending upon whether the event does or doesn't occur at the moment in question).

Sorts	$S ::=$	$\texttt{Object} \mid \texttt{Agent} \mid \texttt{ActionType} \mid \texttt{Action} \sqsubseteq \texttt{Event} \mid \texttt{Fluent} \mid \texttt{Moment} \mid \texttt{Boolean}$
Functions	$f ::=$	$\mathit{action} : \texttt{Agent} \times \texttt{ActionType} \longrightarrow \texttt{Action}$ $\mathit{initially} : \texttt{Fluent} \longrightarrow \texttt{Boolean}$ $\mathit{holds} : \texttt{Fluent} \times \texttt{Moment} \longrightarrow \texttt{Boolean}$ $\mathit{happens} : \texttt{Event} \times \texttt{Moment} \longrightarrow \texttt{Boolean}$ $\mathit{clipped} : \texttt{Moment} \times \texttt{Fluent} \times \texttt{Moment} \longrightarrow \texttt{Boolean}$ $\mathit{initiates} : \texttt{Event} \times \texttt{Fluent} \times \texttt{Moment} \longrightarrow \texttt{Boolean}$ $\mathit{terminates} : \texttt{Event} \times \texttt{Fluent} \times \texttt{Moment} \longrightarrow \texttt{Boolean}$ $\mathit{prior} : \texttt{Moment} \times \texttt{Moment} \longrightarrow \texttt{Boolean}$
Terms	$t ::=$	$x : S \mid c : S \mid f(t_1, \ldots, t_n)$
Propositions	$P ::=$	$t : \texttt{Boolean} \mid \neg P \mid P \wedge Q \mid P \rightarrow Q \mid P \leftrightarrow Q \mid \forall_{x:S}\ P \mid \exists_{x:S}\ P \mid \mathbf{S}(a, P) \mid \mathbf{K}(a, P) \mid \mathbf{B}(a, P) \mid \mathbf{C}(P)$

Reasoning in the $\mathcal{SCC}$ is realized via natural-deduction style inference rules. For instance, R_2 shows that knowledge entails belief; R_3 infers from "P is common knowledge" that, for any agents a_1, a_2, and a_3, "a_1 knows that a_2 knows that a_3 knows that P." And R_4 guarantees the veracity of knowledge; that is, if an agent "knows that P," then P is, in fact, the case.

$$\frac{}{\mathbf{C}(\mathbf{K}(a, P) \rightarrow \mathbf{B}(a, P))}\ [R_2] \qquad \frac{\mathbf{C}(P)}{\mathbf{K}(a_1, \mathbf{K}(a_2, \mathbf{K}(a_3, P)))}\ [R_3] \qquad \frac{\mathbf{K}(a, P)}{P}\ [R_4]$$

In the $\mathcal{SCC}$, agent actions are modeled as types of events. We model lying, asserting, and stating propositions as types of actions that an agent may perform. These action types are denoted by the functions *lies*, *asserts*, and *states*. The argument to such action types are conceived of as reified propositions, specifically fluents. Thus, the formula $\mathit{happens}(\mathit{action}(l, \mathit{states}(p, d)), m)$ is read, "it happens at moment m that agent l states (reified) proposition p to agent d." For convenience, we model that an agent is a liar by using the property *liar*. The signature for these additions to the $\mathcal{SCC}$ is:

Functions	$f ::=$	$\mathit{states} : \texttt{Fluent} \times \texttt{Agent} \longrightarrow \texttt{ActionType}$ $\mathit{asserts} : \texttt{Fluent} \times \texttt{Agent} \longrightarrow \texttt{ActionType}$ $\mathit{lies} : \texttt{Fluent} \times \texttt{Agent} \longrightarrow \texttt{ActionType}$ $\mathit{liar} : \texttt{Agent} \longrightarrow \texttt{Boolean}$

The definitions of *liar*, *lies*, and *asserts* are stipulated as common knowledge by Axioms (1)–(3).

$$\mathbf{C}\left(\forall_{l}\ liar(l) \leftrightarrow \exists_{d,p,m}\ happens(action(l, lies(p,d)), m)\right) \tag{1}$$

$$\mathbf{C}\left(\forall_{l,d,p,m}\ happens(action(l, lies(p,d)), m) \leftrightarrow \left(\begin{array}{c}\mathbf{B}(l, \neg holds(p,m)) \wedge \\ happens(action(l, asserts(p,d)), m)\end{array}\right)\right) \tag{2}$$

$$\mathbf{C}\left(\begin{array}{c}\forall_{l,d,p,m}\ happens(action(l, asserts(p,d)), m) \leftrightarrow \\ \left(\begin{array}{c}happens(action(l, states(p,d)), m) \wedge \\ \mathbf{B}(l, \mathbf{B}(d, happens(action(l, states(p,d)), m) \rightarrow \mathbf{B}(l, holds(p,m))))\end{array}\right)\end{array}\right) \tag{3}$$

Now that we have in hand a formal account for lying, we can reexamine the scenario posed earlier. Assuming that Larpals conform to the plausible definition of lying — not that every statement they make is false, but rather that their assertions, at times, misrepresent their beliefs — and that Tarsals conform to a counterpart notion of honesty — that their assertions faithfully reflect their beliefs, which are, however, still fallible — can one determine which aliens are trustworthy Tarsals?

In order to represent the Larpals & Tarsals scenario, we further extend the $\mathcal{SCC}$ signature with the functions *alien*, *larpal*, and *tarsal*, and a number of constants:

		$alien$: `Agent` $\longrightarrow$ `Boolean`
Functions	$f ::=$	$larpal$: `Agent` $\longrightarrow$ `Boolean`
		$tarsal$: `Agent` $\longrightarrow$ `Boolean`
		H, A_1, A_2, A_3 : `Agent`
Constants	$c ::=$	p_1, p_2, p_3 : `Fluent`
		m_1, m_2, m_3 : `Moment`

The constants A_1, A_2, and A_3 denote the three distinct aliens, and the constant H denotes the human. From the scenario we extract several pieces of common knowledge: (i) A_1, A_2, and A_3 are aliens, while H is not; (ii) Larpals are liars, and Tarsals are not; (iii) every alien is a Larpal or a Tarsal; and (iv) aliens recognize whether other aliens are Larpals or Tarsals. This common knowledge is represented by Axioms (4)–(7).

$$\mathbf{C}(alien(A_1) \wedge alien(A_2) \wedge alien(A_3) \wedge \neg alien(H)) \tag{4}$$

$$\mathbf{C}\left(\forall_{a}\ (larpal(a) \rightarrow liar(a)) \wedge (tarsal(a) \rightarrow \neg liar(a))\right) \tag{5}$$

$$\mathbf{C}(\forall_{a}\ alien(a) \rightarrow (larpal(a) \vee tarsal(a))) \tag{6}$$

$$\mathbf{C}\left(\forall_{a_1,a_2}\ (alien(a_1) \wedge alien(a_2)) \rightarrow \left(\begin{array}{c}(tarsal(a_2) \rightarrow \mathbf{K}(a_1, tarsal(a_2))) \wedge \\ (larpal(a_2) \rightarrow \mathbf{K}(a_1, larpal(a_2)))\end{array}\right)\right) \tag{7}$$

In addition to the above common knowledge, H knows that there are exactly two Tarsals and one Larpal in the delegation. This knowledge is represented by Axiom (8).

$$\mathbf{K}\left(H,\ \begin{array}{c}(larpal(A_1) \leftrightarrow (tarsal(A_2) \wedge tarsal(A_3))) \wedge \\ (larpal(A_2) \leftrightarrow (tarsal(A_1) \wedge tarsal(A_3))) \wedge \\ (larpal(A_3) \leftrightarrow (tarsal(A_1) \wedge tarsal(A_2)))\end{array}\right) \tag{8}$$

The specific interactions in the scenario at hand brings about several more axioms. The first alien's utterance was unclear, but it is common knowledge that, at moment m_1, it asserted something to H; that something is denoted by the reified proposition p_1. Similarly, it is common

knowledge that, at moment m_2, the second alien asserted p_2 to H. Furthermore, it is common knowledge that p_2 is materially true if and only if the first alien declared itself a Larpal. Finally, it is common knowledge that, at moment m_3, the third alien asserted p_3 to H, and that p_3 is materially true if and only if the second alien's assertion to H was a lie. These actions, and the truth conditions of the various assertions, are represented by Axioms (9)–(13).[18]

$$\mathbf{C}(happens(action(A_1, asserts(p_1, H)), m_1)) \tag{9}$$

$$\mathbf{C}(happens(action(A_2, asserts(p_2, H)), m_2)) \tag{10}$$

$$\mathbf{C}(holds(p_2, m_2) \leftrightarrow (holds(p_1, m_1) \leftrightarrow larpal(A_1))) \tag{11}$$

$$\mathbf{C}(happens(action(A_3, asserts(p_3, H)), m_3)) \tag{12}$$

$$\mathbf{C}(holds(p_3, m_3) \leftrightarrow happens(action(A_2, lies(p_2, H)), m_2)) \tag{13}$$

With the Larpals & Tarsals scenario now formalized in the $\mathcal{SCC}$, we proceed to sketch how H, given sufficient contemplation, can know that A_1 and A_3 are Tarsals, and that A_2 is a Larpal. Our sketch consists of three parts: (i) we indicate how H can know that if A_2 is a Tarsal, then A_3 is a Larpal; (ii) we indicate how H can know that if A_1 is a Tarsal, then A_2 is a Larpal; (iii) we indicate how H can know, based on these two conditionals, that A_1 and A_3 are Tarsals, and that A_2 is a Larpal. In the prose elaboration of the three parts, the reasoning is described from H's perspective.

First, suppose (H reasons to itself) that A_2 is a Tarsal. Faction membership is apparent to aliens, and so A_3 also knows that A_2 is a Tarsal. A_3 also knows that Tarsals are not liars, and, more specifically, that it does not happen that A_2 lies to H. Therefore, A_3 knows that the proposition that it asserted, p_3, does not hold, i.e., it is materially false. Since A_3 knows this, it also believes this. Hence, A_3 was lying when it made its assertion, so it must be a liar, and so not a Tarsal, and thus a Larpal. In this way, H reasons to itself that if A_2 is a Tarsal, then A_3 is a Larpal. Here, expressed in the aforementioned "streamlined" format for describing proof, is an abbreviated proof that mirrors this description of H's reasoning:

1	$\mathbf{K}(H, tarsal(A_2))$	assumption
2	$\mathbf{K}(H, \mathbf{K}(A_3, alien(A_2)))$	by Axiom (4)
3	$\mathbf{K}(H, \mathbf{K}(A_3, tarsal(A_2)))$	by Axiom (7) and step 1
4	$\mathbf{K}(H, \mathbf{K}(A_3, \neg liar(A_2)))$	by Axiom (5) and step 3
5	$\mathbf{K}(H, \mathbf{K}(A_3, \neg happens(action(A_2, lies(p_2, H)), m_2)))$	by Axiom (1) and step 4
6	$\mathbf{K}(H, \mathbf{K}(A_3, \neg holds(p_3, m_3)))$	by Axiom (13) and step 5
7	$\mathbf{K}(H, \mathbf{B}(A_3, \neg holds(p_3, m_3)))$	by step 6 and R_2
8	$\mathbf{K}(H, happens(action(A_3, lies(p_3, H)), m_3))$	by Axioms (2) and (12) and step 7
9	$\mathbf{K}(H, liar(A_3))$	by Axiom (1) and step 8
10	$\mathbf{K}(H, \neg tarsal(A_3))$	by Axiom (5) and step 9
11	$\mathbf{K}(H, larpal(A_3))$	by Axiom (6) and step 10

Next, suppose (H reasons to itself) that A_1 is a Tarsal. Faction membership is apparent to aliens, and so A_2 knows that A_1 is a Tarsal. A_2 also knows that Tarsals are not liars, and, more specifically, that it does not happen that A_1 lies to H. Since A_2 knows that A_1 asserted p_1 to H and A_1 does not lie, A_2 knows that it is not the case that A_1 believes that p_1 does not hold. Yet, A_2 also knows A_1 knows (and thus believes) that A_1 is not a Larpal. Therefore, A_2 knows that it

[18] We elide the axiom stipulating the temporal order of m_1, m_2, and m_3. Informally, m_1 is prior to m_2 and m_2 is prior to m_3.

is impossible for A_1 to have asserted that it is a Larpal, for if it did, then it would be a liar. That is to say, A_2 knows that its own assertion p_2 does not hold, i.e., is materially false. Thus, A_2 lies in asserting p_2. In this way, H reasons to itself that if A_1 is a Tarsal, then A_2 is a Larpal.

Last, were it the case (H reasons to itself) that A_1 is a Larpal, then A_2 and A_3 would be Tarsals, because there is only one Larpal among the three. Yet, if A_2 is a Tarsal, then A_3 is a Larpal, which contradicts A_3 being a Tarsal. Hence, A_1 is not a Larpal, thus A_1 is a Tarsal. Since A_1 is a Tarsal, A_2 is a Larpal, and then A_3 is, like A_1, a Tarsal. Finally, in this way, H reasons to itself that A_1 and A_3 are trustworthy Tarsals, and A_2 is the dishonest Larpal.

Note that in the final part, where H definitively determines which aliens are trustworthy, H's reasoning depends on knowing that there are two Tarsals and one Larpal among the three aliens. Without such knowledge, it is impossible for H to decide who to trust. Alas, in the real-world, such knowledge is not likely. Anyone, at least any human, may lie. Furthermore, nefarious plots (e.g., fraud, pyramid schemes, espionage, guerrilla tactics, and terrorism) depend on lying and lesser deceptions. Machines may play a role in guarding, e.g., free-market consumers, private citizens, and sovereign states against such plots, but only if machines are able to comprehend philosophical concepts like mendacity and deception. In turn, KR&R systems cannot begin to grasp such concepts unless they embrace philosophy, and the formal sophistication that philosophy demands.

3.2.2 Brief Remark on Evil KBSs

In general, there seems to be no reason in principle why KR&R cannot be applied not only to socio-cognitive concepts like mendacity and deception, but also to even richer and more nuanced concepts that incontestably require philosophical analysis in order to be couched in terms precise enough to allow knowledge-bases to hold queryable information about them. For instance, one could consider the possibility of engineering a KBS that is capable of betraying someone, or capable, in general, of being evil. It seems quite undeniably that no KR&R expert could engage in such engineering without both engaging philosophy and making use of highly expressive logics.

Engineering for the former case, which did indeed explicitly involve both philosophy and highly expressive formal languages, has already been carried out (see Bringsjord & Ferrucci 2000). In this work, philosophical analysis was used to gradually craft a definition of the concept of one agent betraying another. This definition was consistent with $\mathcal{C}$.

What about evil? Here the investigation is still in its early stages.[19] The basic process, though, is the same as what we showed in action in connection with mendacity: philosophy is used to build the definition of evil; the definition is formalized in some logical system; knowledge-bases describing evil agents are populated; and queries against such knowledge-bases are issued and answered, which gives rise to the relevant knowledge-based systems. The interesting thing about this KR&R work is that if, as some have claimed (e.g., M. Scott Peck; see Peck 1983), a truly evil agent is one that harbor outright contradictions in what he or she believes, logical systems able to allow the representation of contradictory information, and the unproblematic reasoning over that information, would be necessary. Such logical systems are highly expressive and would be $\mathcal{C}$-confirming. These systems are known as *paraconsistent logics*; a nice introduction for the more industrious of our readers can be obtained via (Priest, Routley & Norman 1989).

[19] Some preliminary work is described in (Bringsjord, Khemlani, Arkoudas, McEvoy, Destefano & Daigle 2005).

3.3 "Visual" KR&R and the Future

Heretofore, when representing propositional content, the field of KR&R has been exclusively linguistic in nature.[20] This is consistent with the fact that, to this point in the present paper, all formal languages used for the representation of propositional content have been exclusively linguistic: Well-formed formulas generable by the alphabets and grammars of these languages are invariably strings of characters, and these strings in no way "directly resemble" that which they are intended to denote. For example, when we spoke earlier of liars and truth-tellers, and used names to refer to them in our case studies of mendacity, we specifically used the constant 'A_1' to refer to one of the aliens. Had we felt like doing so, we could just as easily have used instead the constant 'A99,' or 'a1,' or 'A-23,' and so on, *ad indefinitum*. In contrast, a diagrammatic representation of the alien in question would bear a resemblance relation to him, and even slight changes in the diagram could prevent it from denoting the alien. As the philosopher Peirce put it, "a diagram is naturally analogous to the thing represented" (Peirce 1998).

Despite the fact that KR&R has traditionally left aside pictorial representation schemes, there can be no disputing the fact that human reasoning is powerful in no small part because it is often *diagrammatic* in nature. Ironically, while KR&R, as elegantly explained by Glymour (1992), has been purely linguistic since the first formal language for KR&R was introduced by Aristotle (viz., the theory of the syllogism), Aristotle, along with his linguistic-oriented successors in the modern era (e.g., Boole and Frege), sought to explain how the highly visual activity of the great Euclid could be formalized via some logical system. It is plausible to hold, as we do, that substantive parts of this long-sought explanation began to arrive on the scene courtesy of seminal work carried out by a pair of oft-collaborating logician/philosophers: Jon Barwise and John Etchemendy. Since the space we have to discuss diagrammatic KR&R is quite limited, we shall briefly explain, by way of a problem posed by this pair, how a hybrid diagrammatic/linguistic formal language in the so-called Vivid family of such languages can be used to solve this problem. The problem is a seating puzzle given in (Barwise & Etchemendy 2008). The Vivid family is presented in (Arkoudas & Bringsjord forthcoming).

Here is the seating puzzle, which has become something of a classic: Five people — A, B, C, D, E — are to seated in a row of seats, under the following three constraints.

C1 A and C must flank E.

C2 C must be closer to the middle seat than B.

C3 B and D must be seated next to each other.

Now, three problems are to be solved:

P1 Prove that E cannot be in the middle, or on either end.

P2 Can it be determined who must be sitting in the middle seat?

P3 Can it be determined who is to be seated on the two ends?

The class of relevant diagrams in this case can conveniently be viewed as a quintuple, each member of which is either one of the five people, or the question-mark. For example, here is a diagram:

$$AECBD$$

[20] Sometimes the adjective 'symbolic' is used instead of 'linguistic.'

Note that this diagram satisfies all three constraints. As another example, the diagram

$$??A??$$

is one in which A is seated in the middle chair.

We are now in position to consider a proof that constitutes a solution to the puzzle. This proof will be *heterogeneous*: it will make use of propositional content expressed in traditional form (i.e., in the form of formulas from the kind of formal languages presented and employed above), *and* it will make use of such content expressed in the form of diagrams. Since none of the formal languages seen above (e.g., $\mathcal{L}_{PC}$, $\mathcal{L}_{FOL}$, etc.) allow for diagrams as well-formed expressions, this proof cannot be based in the logical systems visited above. Here is the proof:

Proof. Given that E must be between A and C, there are six diagrams to consider, viz.:

$AEC??$	(1)	$?AEC?$	(2)	$??AEC$	(3)
$CEA??$	(4)	$?CEA?$	(5)	$??CEA$	(6)

However, only (1) and (6) are consistent with the other two constraints. P1 is therefore accomplished. Since in both of these diagrams C is in the middle seat, P2 is answered in the affirmative. As to P3, the answer is 'No,' since any end could have one of A, B, or D. QED

The diagrammatic representations seen in this seating puzzle, and in the solution thereof, are frequently used by human reasoners, but have hitherto not been part and parcel of KR&R. Yet, in philosophy, there is a very strong tradition of not only recognizing that such representations are often used, but also of making their use precise in systems that go beyond the purely linguistic.[21] We do not have the space to present and discuss such systems here. We direct the reader to the Vivid system (Arkoudas & Bringsjord forthcoming) for details on how seating puzzle, as well as much more complicated representation and reasoning of a visual sort, can be made precise and mechanized.

We conclude this section with a brief remark about the historical context: Logic grew out of philosophy; computer science, and specifically AI, in turn, grew out of logic. This progression is nicely chronicled by Glymour (1992). But we are now into a *new* progression, one driven by philosophy and logic as midwives, and is gradually expanding KR&R into the visual realm.

4 Conclusion

We have set out and defended the view that if KR&R is to reach into the realms of mathematics and socio-cognition, then philosophy must become a genuine partner in the enterprise. While we have mentioned a number of phenomena in these realms, we have shown this view in action through a particular emphasis on a concept — mendacity — that by its very nature involves social cognition. We predict that as KR&R expands and matures into the future, if for no other reason than that it should allow humans to work collaboratively with intelligent machines having direct and immediate access to electronic propositional content about social, cognitive, mathematical, and visual matters,

[21] In addition to the work of Barwise and Etchemendy, there is for example also the seminal work on visual logic carried out by the philosopher Peirce; see, e.g., (Shin 2002, Peirce 1998).

philosophy and philosophers will be consulted to an increasingly high degree. If such consultation does not come to pass, and the conjecture above is correct, it follows that KR&R will be limited to propositional content that is but a tiny fragment of what is known by human beings; and it follows in turn from that that intelligent machines, relative to human minds, will, knowledge-wise, remain exceedingly primitive by comparison.

References

Arkoudas, K. (n.d.), Athena.
URL: *http://www.cag.csail.mit.edu/∼ostas/dpls/athena*

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. (2008), Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task, *in* T.-B. Ho & Z.-H. Zhou, eds, 'PRICAI 2008: Trends in Artificial Intelligence, 10th Pacific Rim International Conference on Artificial Intelligence, Hanoi, Vietnam, December 15–19, 2008, Proceedings', number 5351 *in* 'Lecture Notes in Artificial Intelligence (LNAI)', Springer-Verlag, pp. 17–29.

Arkoudas, K. & Bringsjord, S. (forthcoming), 'Vivid: An AI Framework for Heterogeneous Problem Solving', *Artificial Intelligence* . Penultimate draft available electronically.
URL: *http://kryten.mm.rpi.edu/vivid/vivid.pdf*

Baader, F., Calvanese, D. & McGuinness, D., eds (2003), *The Description Logic Handbook: Theory, Implementation and Applications*, Cambridge University Press, Cambridge, UK.

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

Barwise, J. & Etchemendy, J. (2008), Information, Infons, and Inference, *in* 'Situation Theory and Its Applications', CSLI, Stanford, CA, pp. 33–78.

Bennett, M. & Waltz, E. (2007), *Counterdeception Principles and Applications for National Security*, Artech House, Norwood, MA.

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.

Boolos, G. S., Burgess, J. P. & Jeffrey, R. C. (2003), *Computability and Logic*, 4th edn, Cambridge University Press, Cambridge, UK.

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

Bringsjord, S. (2008), 'The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself', *Journal of Applied Logic* **6**(4), 502–525.

Bringsjord, S. & Arkoudas, K. (2006), On the Provability, Veracity, and AI-Relevance of the Church-Turing Thesis, *in* A. Olszewski, J. Wolenski & R. Janusz, eds, 'Church's Thesis After 70 Years', Ontos Verlag, Frankfurt, Germany, pp. 66–118.

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.

Bringsjord, S. & Yang, Y. (2003), Representations Using Formal Logics, *in* L. Nadel, ed., 'Encyclopedia of Cognitive Science', Vol. 3, Nature Publishing Group, London, UK, pp. 940–950.

Chisholm, R. M. & Feehan, T. D. (1977), 'The Intent to Deceive', *Journal of Philosophy* **74**(3), 143–159.

Clark, M. H. (2009), Cognitive Illusions and the Lying Machine: A Blueprint for Sophistic Mendacity, PhD thesis, Rensselaer Polytechnic Institute, Department of Cognitive Science, Troy, NY.

Dennett, D. C. (1995), *Darwin's Dangerous Idea: Evolution and the Meanings of Life*, Simon & Shuster, New York, NY.

Ebbinghaus, H.-D., Flum, J. & Thomas, W. (1984), *Mathematical Logic*, Springer-Verlag, New York, NY.

Fitch, F. B. (1952), *Symbolic Logic: An Introduction*, Ronald Press, New York, NY.

Glymour, C. (1992), *Thinking Things Through: An Introduction to Philosophical Issues and Achievements*, MIT Press, Cambridge, MA.

Kleene, S. C. (1952), *Introduction to Metamathematics*, Van Nostrand, New York, NY.

Kreisel, G. (1967), Informal Rigor and Completeness Proofs, *in* I. Lakatos, ed., 'Problems in the Philosophy of Mathematics', North-Holland, Amsterdam, The Netherlands, pp. 138–186.

Peck, M. S. (1983), *People of the Lie*, Simon & Shuster, New York, NY.

Peirce, C. S. (1998), *The Collected Papers of Charles Sanders Peirce*, Thoemmes Press, Bristol, UK. This set is edited by C. Hartshorne and P. Weiss.

Priest, G., Routley, R. & Norman, J., eds (1989), *Paraconsistent Logic: Essays on the Inconsistent*, Philosophia Verlag, München, Germany.

Shin, S.-J. (2002), *The Iconic Logic of Peirce's Graphs*, MIT Press, Cambridge, MA.

Smith, P. (2007), *An Introduction to Gödel's Theorems*, Cambridge University Press, Cambridge, UK.

Sun, R. & Bringsjord, S. (2009), Cognitive Systems and Cognitive Architectures, *in* B. W. Wah, ed., 'The Wiley Encyclopedia of Computer Science and Engineering', Vol. 1, Wiley, New York, NY, pp. 420–428.