Worlds as a Unifying Element of Knowledge Representation

J.R. Scally, Nicholas L. Cassimatis, and Hiroyuki Uchida

Department of Cognitive Science, Rensselaer Polytechnic Institute
scallj@rpi.edu, cassin@rpi.edu, uchidh@rpi.edu

Abstract

Cognitive systems with human-level intelligence must display a wide range of abilities, including reasoning about the beliefs of others, hypothetical and future situations, quantifiers, probabilities, and counterfactuals. While each of these deals in some way with reasoning about alternative states of reality, no single knowledge representation framework deals with them in a unified and scalable manner. As a consequence it is difficult to build cognitive systems for domains that require each of these abilities to be used together. To enable this integration we propose a representational framework based on synchronizing beliefs between worlds. Using this framework, each of these tasks can be reformulated into a reasoning problem involving worlds. This demonstrates that the notions of worlds and inheritance can bring significant parsimony and broad new abilities to knowledge representation.

Introduction

A cognitive system with human-level intelligence must be able to reason about the beliefs of other agents, perform inference about counterfactual and hypothetical situations, reason about uncertainty, consider possible future situations, and make decisions using quantified statements such as "all", "some" or "none". Each of these capabilities involves reasoning about an alternative state of reality. In general, researchers have pursued solutions to these challenges independently, with incompatible assumptions and formalisms. However, in many real-life scenarios, agents require all of these capabilities in a single situation: for example, using the beliefs of another agent, and probabilities for their behavior in a hypothetical situation to plan future actions. A human-level cognitive system needs to integrate reasoning with all of these abilities.

These capabilities all seem to deal in some way with reasoning over alternative versions of reality. Many approaches within cognitive science have dealt with this in some way. "Possible worlds" were introduced to define the semantics of modal logic and to describe the truth conditions of counterfactuals (Kripke 1963; Stalnaker 1968; Lewis 1973). Situation calculus (McCarthy and Hayes 1969; McCarthy 2002) uses situations to represent aspects of the environment that can be used by an agent for reasoning. Situational semantics and information flow theory (Barwise and Perry 1983; Devlin 1991) use the relationship between partial abstract situations to analyze how information is transmitted between agents. The theory of mental spaces (Fauconnier 1985) uses linked mental representations to explain the complexities of identity and reference in language about images, beliefs, conditionals, and counterfactuals. Mental model theory (Johnson-Laird 1983; 2006) treats reasoning as a process of constructing and evaluating possible models of the world. Simulation theories of human cognition (Barsalou 1999; 2009; Goldman 2002; Gordon 1995; Gordon and Cruz 2002) have attempted to explain many forms of reasoning in terms of mental simulations. Although each of these approaches include significant insights, few of them are sufficiently well-defined for implementation in a general-purpose reasoning system.

An alternative approach is to create modules for each form of high-level reasoning. In this case, however, the inferential results from each module must still be integrated into a common representation to share between modules. As stated above, this integration would need to take place at a low enough level to facilitate reasoning within a single scenario that requires all the abilities mentioned above. Assuming that a separate module exists for each of these capabilities is also problematic for developmental and evolutionary accounts of human cognition.

The modular approach is also at odds with the view that human intelligence can be explained in terms of a small number of cognitive mechanisms. Many have presented psychological, neuroscientific, evolutionary, and computational arguments and evidence that a small set of common elements underlie the whole range of human cognition (e.g., Lakoff and Johnson 1980; 1999; Cassimatis 2006). A

framework which shows how all of these cognitive abilities emerge from a single set of principles would provide a parsimonious account that could inform both cognitive system building, and an understanding of human cognition.

In this paper, we propose a unified framework for representing knowledge about the beliefs of others, hypothetical and future situations, quantifiers, probabilities, and counterfactuals. The framework is based on the notion of "worlds" that are alternate representations of reality. The framework requires only a small set of specific, formally-defined principles governing the synchronization of truth values among worlds to provide a unified treatment of these various forms of knowledge.

Principles of a World-Based Framework

We will begin by defining the notation for the framework used in the rest of the paper. An atom expresses a relation over one or more entities[1] and takes a truth value in world.[2]

(1) Predicate(a_1, ..., a_n, world)

An atom written by itself asserts that the atom is true; the negation operator (-) indicates an atom is false[3].

"John is holding the ball"
(2) Holds(john, ball, w)

"Meg is not holding the block"
(3) -Holds(meg, block, w)

Constraints are used to expresses contingencies between atoms, using the operators for conjunction (^), and implication (→). Entities (including predicates) can be given universal force by designating them with variables, prepending a "?".

"All large dogs bark."
(4) IsA(?dog, Dog, E, w) ^ Size(?dog, large, E, w)
→ Bark(?dog, E, w)

The above constraint is a *hard constraint*: when atoms the match its antecedent are true, then the implied consequent must also be true. Some constraints can be broken, but at a cost. Such *soft constraints* are indicated by prepending the implication symbol with the cost for breaking the constraint.

"All large dogs usually bark."
(5) IsA(?dog, Dog, E, w) ^ Size(?dog, large, E, w)
(.75)→Bark(?dog, E, w)

[1] In this framework, the definition of entities is very general and can include objects, states, sets, and relations

[2] The meaning of the world argument will be described below in detail, but for now it will be left unspecified with the generic constant "w".

[3] In order to increase clarity, this account assumes Boolean truth values. Nothing, however, in the following analysis precludes the use of more complex truth values or probabilities.

While a variable in a constraint that appears in the antecedent has universal force, a variable that appears in the consequent of a constraint has existential force.

"For every large dog there exists a collar that it wears."
(6) IsA(?dog, Dog, E, w) ^ Size(?dog, large, E, w)
→ Wears(?dog, ?c, E, w) ^ IsA(?c, Collar, E, w)

This notation straightforwardly maps onto logic. Example (6) can be represented with quantifiers as follows:

(7) $\forall x \exists y(\text{Dog}(x) \wedge \text{Large}(x) \rightarrow \text{Collar}(y) \wedge \text{Wears}(x, y))$

This notation was chosen for representational parsimony and compatibility with existing inference systems such as production rule systems, and constraint satisfaction algorithms. This is neither a claim that human reasoning uses logic, nor a claim that logic is the best way to implement a world-based framework. Previous work has shown (Cassimatis, et al 2004; Cassimatis, et al 2010) that world-based reasoning can be utilized in a system that includes non-logical inference mechanisms.

A Representation of Reality

Consider an agent that is interacting with the environment via sensors and manipulators. For reasoning and planning it will need to represent facts about the environment (e.g., "The target is behind the block"). The agent should be able to assume that it can accurately represent a model of reality. Therefore, we will add *R* to our framework. Atoms take a world argument of *R*, the real world, when they specify a truth value in reality. For example:

(8) Predicate(a_1, ..., a_n, R)

"The target is behind the block"
(9) Behind(target, block, R)

These atoms are designated as being in *R* because they describe relationships that hold in reality and are not part of an imagined or counterfactual situation. An agent's perceptions, inferences, memories and knowledge about reality can all be represented as atoms holding in *R*.

Worlds Defined By Assumptions

Suppose an agent knows that if a ball is behind a screen, a red light will go on. If the agent does not know whether a ball is behind the screen, it can consider both possibilities. However, these possibilities cannot be simultaneously asserted as being as true in *R*, as this will lead to a contradiction. Rather, the world where the ball is behind the screen and the world where the ball is not behind the screen must be considered separately. In the world where the ball is behind the screen, the red light will be on. We say that the assumptions a world is based on (e.g., a ball is behind the screen) constitute the *basis* of that world.

Each world has a basis
(10) Basis(*world*) = { $a_1, \ldots, a_n$ }

The worlds from the example above can now be represented as follows:

Hypothetical world where the ball is behind the screen
(11) Basis(w1) = { Location(ball, behindScreen, R) }

Hypothetical world where the ball is not behind the screen
(12) Basis(w2) = { -Location(ball, behindScreen, R) }

Note that in (11) and (12), the world argument of the basis atom is set to *R*. The assumption is made *for* this world – the atom does not receive a truth value in *R*. The basis atoms are true and false in *w1* and *w2* respectively.

The definition of *R* can now be refined: it is the world based on no assumptions:

(13) Basis(R) = {}

The basis assumptions give the unique and minimal definition of a world. Atoms which are consequences of the basis atoms are treated as conclusions of the world and are marked with the corresponding world argument. To continue the example, we can now assert the following:

"In the world where the ball is behind the screen, the light is on."
(14) State(light, on, w1)

Any atom with a world argument such as *w1* has the following meaning: "In the world where *basis_atoms* are true (or false) then *conclusion_atom* is true (or false)".

Inheritance and Overrides Between Worlds

Consider an agent in a scenario where it is holding a ball, with an additional goal of picking up a box from a shelf. If the agent wants to reason about alternative state of affairs (a hypothetical world where it puts the ball down, or where the grasping of the ball failed), then it must construct an alternative world. However, when the agent makes a hypothetical assumption, many facts do not change, for example: the location of the ball, the goals of the agent, the location of the walls, the position of obstacles, the color of the sky, etc. Thus whenever an assumption is made, while some relations change, many (often most) others do not.

There is also a clear hierarchical relationship between assumptions that an agent can exploit. It might consider a world based on the assumption that it dropped the ball. In this world, the agent's hands become free. The agent can then make a further assumption that it picks up the box from the shelf. This second assumption begins from the world of the first assumption and conclusions from the first world are used in the world of the second assumption.

Inheritance Between Worlds. These points can be illustrated by showing a set of possible atoms that such an agent might use to describe its immediate environment:

(15) InFrontOf(shelf, agent, R)
Location(box, shelf, R)
Location(agent, pos34, R)

A hypothetical world where the ball is dropped
(16) Basis(w3) : { PerformAction(drop, ball, R) }

The assumption that creates *w3* is made in light of world *R*, therefore atoms true in *R* should also be true in *w3*.

(17) InFrontOf(shelf, agent, w3)
Location(box, shelf, w3)
Location(agent, pos34, w3)

With this example in mind, we can formally define these principles. A world inherits the truth values of atoms from all other worlds whose basis is fully contained in the basis atoms which define that world. We refer to this property as *relevance*.

(18) Basis(?w1) ⊆ Basis(?w2) → RelevantTo(?w1, ?w2)

Because *R*'s basis has no elements, its basis is a subset of the basis of all other worlds and *R* is therefore relevant to all worlds. Since set containment is transitive, it follows directly that relevance is transitive:

(19) RelevantTo(?w1, ?w2) ^ RelevantTo(?w2, ?w3)
→ RelevantTo(?w1, ?w3)

Now that we have defined the notion of relevance we can write a constraint to inherit atoms from relevant worlds:

(20) ?Predicate(?a_1, ..., ?a_n, ?w1) ^RelevantTo(?w1,?w2)
→ ?Predicate(?a_1, ..., ?a_n, ?w2)

Overriding Inheritance. In domains where the agent seeks only to reason about situations based in reality, (20) is sufficient. Any contradictions in these situations would imply invalid reasoning. However, in counterfactual situations, where the agent begins with a contradictory assumption, this constraint will create a contradiction. The agent, for example, might want to consider how it might have recovered had a "pick up" action failed:

"The agent's hand is not empty"
(21) -State(hand, empty, R)

Hypothetical world where the agent's hand is empty
(22) Basis(w4) : { State(hand, empty, R) }

With these atoms, (20) will cause a contradiction in world *w4*. In counterfactual situations, while we want inheritance, we want it to be blocked when contradictory conclusions are reached. We still want to minimize the differences between relevant worlds, while allowing for contradictions.

This can be accomplished by making inheritance a soft constraint, adding a cost whenever inheritance is blocked[4]:

(23) ?Predicate(?a_1, ..., ?a_n, ?w1) ^RelevantTo(?w1,?w2)
(*cost*)→?Predicate(?a_1, ..., ?a_n, ?w2)

This analysis indicates two ways that an alternative world can augment representations of reality. In one, the assumptions of the new world *extend* what is known, in order to explore a future situation or uncertain situation. If the inference that follows causes a contradiction, then the assumption can be taken to be invalid. In the other, the contradiction is expected; the purpose of the world is to *adjust* what is known to accommodate reasoning in a different situation, bringing to bear as much as possible from reality. Starting from a set of counterfactual assumptions, maximal information is carried over from reality, and overridden as needed during inference.

Summary

We have motivated and introduced the concept of worlds and a small number of principles governing the relationships between them. These are summarized as follows:

- An agent maintains a representation of reality. In our notation, this corresponds to atoms defined in *R*.

(8) Predicate(a_1, ..., a_n, R)

- Reasoning involves representing alternative states of the world based on assumptions. In our notation, this is represented through worlds defined by basis atoms.

(10) Basis(*world*) = { a_1, ... , a_n }

- By default, the beliefs of an agent are consistent between worlds. We represent this through a scheme of default inheritance where overrides will incur a cost.

(23) ?Predicate(?a_1,..., ?a_n, ?w1) ^ RelevantTo(?w1,?w2)
(*cost*)→?Predicate(?a_1, ..., ?a_n, ?w2)

Applications of a World-Based Framework

With these principles, we can analyze a set of problems involving human-level intelligence and show how they can be re-expressed in our framework.

Reasoning about Beliefs

The beliefs of other agents can be thought of as alternate states of the world. Bello, Bignoli, and Cassimatis (2007) have shown that reasoning about another agent can take place by imagining similarities with the other agent, modifying elements that are known to be different. There is considerable evidence that even very young children use this form of "like me" reasoning (Meltzoff 2005).

To illustrate this approach, consider a situation based on the Wimmer and Perner (1983) false belief task. Sally is in the room with an agent, and both observe a cookie placed into a jar. Sally leaves, and the agent watches as the cookie is moved to a second jar. Now, the agent is asked where Sally believes the cookie is located. The agent knows it is in the second jar, but needs to maintain a representation of Sally's mind where the cookie is still in the first jar.

This account can be formalized using worlds. For the sake of brevity, not all constraints will be listed. Models of this example have been built for the Polyscheme Cognitive Architecture[5].

The agent (self) observes a placement in jar A, and then a movement to jar B; Sally only observes the placement.

(24) Placement(placeEvent, cookie, jarA, t1, R)
Moves(moveEvent, cookie, jarB, t2, R)
Perceives(self, placeEvent, R)
Perceives(sally, placeEvent, R)
Perceives(self, moveEvent, R)
-Perceives(sally, moveEvent, R)

Reasoning about Sally's mind is reasoning in the counterfactual world where Sally is identical to the self:

(25) Basis(w5) = { Same(self, sally, R) }

According to *Leibnitz's Law*, if two objects are identical, all of the properties that the objects share should also be identical. This is achieved through (23) which will cause (24) to be inherited into *w5*, and (25) which will replace the instance of "*sally*" in *w5* with "*self*"[6]. This, however, causes a contradiction in *w5*:

(26) Perceives(self, moveEvent, w5)
-Perceives(self, moveEvent, w5)

When reasoning about Sally's mind, conflicts between what the agent knows to be true, and what Sally knows to be true need to be resolved by privileging Sally's perspective over the agent's. This can be formalized by adding another soft constraint:

(27) ?Predicate(self, ?a_1, ..., ?a_n, ?w1)^
-?Predicate(?other, ?a_1, ..., ?a_n, ?w1)^
RelevantTo(?w1, ?w2) ^ ?Same(self, ?other, ?w2)
(cost)→ -?Predicate(self, ?a_1, ..., ?a_n, ?w2)

[4] The cost in this constraint is a configurable parameter that can vary based on the agent's context and situation. Certain terms might be more likely to be overridden than others. Psychological research has shown that human counterfactual reasoning does seem to follow a set of weighted principles (Byrne 2005)

[5] A sample model is available at http://www.rpi.edu/~scallj/models/

[6] There are several possible ways of dealing with identities in alternate worlds; in this example all instances of the second identity term are replaced with an instance of the first identity term.

Once elaboration of *w5* is complete, the self in *w5* observes no movement, and the inference mechanisms for reasoning about reality can be used to infer that the position of the cookie has not changed in *w5*. The contents of *w5* thus correctly reflect Sally's beliefs about the state of the world.

A formal statement can now be given of what it means in this framework to say that "x (in world w) believes p"

(28) Basis(w_n) = { Same(self, x, w) }
p(w_n)

The mechanism defined above is not restricted to world *R*, and thus it can be extended to second order beliefs (the beliefs another agent has about the beliefs of yet another) by nesting worlds. For the agent to reason about Sally's beliefs about Tom, it creates two counterfactual worlds:

(29) Basis(w6) = { Same(self, sally, R) }
(30) Basis(w7) = { Same(self, sally, R)^
Same(sally, tom, E, w6) }

The basis of *w7* includes the basis of *w6*. Everything that the self includes in its list of Sally's beliefs is inherited by default into the self's list of Sally's list of Tom's beliefs.

Counterfactual Reasoning and Mental Spaces

The literature on counterfactual reasoning is vast, and we make no attempt to provide a comprehensive treatment of counterfactuals. We can, however, briefly demonstrate how the framework that we have defined accommodates two accounts of counterfactual reasoning. Mental space theory (Fauconnier 1985) uses "world builders" to connect models of reality with alternate models based on assumptions. Counterfactual situations also emerge in conceptual blending theory (Fauconnier and Turner 1999) through the blending of different input spaces by combining them and making selective overrides. These operations are straightforwardly supported by a world-based framework.

We will demonstrate with an example from Fauconnier and Turner: It has been said that "In France, Watergate would not have hurt Nixon", from which one is to conclude that French politicians are held to different ethical standards. This example illustrates how assertions and reasoning over counterfactual situations can be relevant to reasoning about actual situations. Below are atoms[7] in R, constraints, and counterfactual basis assumptions that capture this example. Running models of this example have been built for the Polyscheme Cognitive Architecture[8].

(31) CommitsCrime(nixon, R)
President(nixon, US, R)
ImpeachedBy(nixon,US, R)
PunishConvictedLeader(US, R)

[7] For brevity the dynamics of the inference are demonstrated with a simplified model containing somewhat *ad hoc* predicates; a full treatment would contain a broader framework of atoms and constraints which specify the dynamics of political systems.

[8] A sample model is available at http://www.rpi.edu/~scallj/models/

"If a country punishes leaders who commit crimes, and if the President commits a crime they will be impeached"

(32) PunishConvictedLeader(?country, ?w) ^
CommitsCrime(?pres, ?w) ^
President(?pres, ?country, ?w)
→ ImpeachedBy(?pres, ?country, ?w)

Hypothetical world where Nixon is President of France and is not impeached.

(33) Basis(w8) = { President(nixon, France, R)^
-ImpeachedBy(nixon, France, R) }

By applying (23) to (31), information will be inherited into *w8*, which will trigger falsification through (32).

(34) -PunishConvictedLeader (France, w8)

This conclusion by itself is not our goal. A speaker using the above counterfactual wants their audience to make an assumption about France in *R*, not about *w8*. This final step is accomplished again by (23) (reproduced below).

(23) ?Predicate(?a_1, ..., a_n, ?w1) ^ RelevantTo(?w1,?w2)
(*cost*)→?Predicate(?a_1, ..., a_n, ?w2)

Previously, this constraint enabled *downward inheritance*, but it also entails *upward inheritance.* New information appearing in *w8* is a candidate for consideration in *R*. There are two possibilities to consider, each of which can be used as the basis for a hypothetical world:

(35) Basis(w9) = { PunishConvictedLeader(France,R) }
(36) Basis(w10) = {-PunishConvictedLeader(France,R)}

It is important to note that (36) does not contradict (34). Thus (23) will add no cost to *w10*. Because there is a contradiction between (35) and (34), *w9* will accrue cost and be considered less likely. It thus emerges naturally that information inferred in a counterfactual world is assumed to be true in *R*, unless there is prior reason to contradict it.

Reasoning with Mental Models, Planning, Logic and Probability

An extremely wide variety of reasoning algorithms can be characterized in terms of exploring alternative states of the world. For example, according to mental model theory (Johnson-Laird, 1983; 2006) humans perform inference by building and evaluating possible models of reality. We can characterize these models in terms of worlds. To illustrate, consider an example of mental model-style inference:

- All golden retrievers bark
- Fido does not bark

An agent can reason about whether or not Fido is a golden retriever by considering the two possibilities as worlds:

(37) Basis(w11) = { IsA(fido, GoldenRetriever, R) }
(38) Basis(w12) = { -IsA(fido, GoldenRetriever, R) }

World *w11* will yield a contradiction, leaving *w12* as the only possible world. Therefore it follows that Fido is not a golden retriever. Recursively applying this kind of reasoning is essentially a form of depth-first search. Depth first search, in the form of the DPLL (Davis and Putnam 1960; Davis, Logemann, and Loveland 1962) algorithm and its modern variants (e.g., Sorensson and Een 2005; Heras, Larrosa, and Oliveras 2008; Hoos and Stutzle 2002; Marques-Silva and Sakallah 1996) represent some of the most efficient algorithms for general purpose reasoning in artificial intelligence. These algorithms all solve satisfiability (SAT) problems or their weighted variants (MaxSAT).

Two important examples are planning and probabilistic inference. For example, it has been demonstrated that planning problems, including formulations expressed in event calculus, can be reduced to SAT solving (Shanahan and Witkowski 2004; Mueller 2004). Probabilistic reasoning works by considering various possibilities and weighting them according to their likelihood. These have also been successfully reduced to weighted SAT problems (Kautz and Selman 1999; Sang and Beame 2005; Singla and Domingos 2006).

Cassimatis, Murugesan, and Bignoli (2009) have demonstrated that many AI algorithms, such DPLL, WalkSAT and Gibbs sampling, can be straightforwardly expressed in terms of simulating alternate states of the world. To illustrate, a version of the classic DPLL algorithm is shown in Figure 1, as compared to a version rewritten in terms of the world framework[9] in Figure 2.

```
DPLL(clauses, symbols, model) returns bool
1. If all clauses are true in model, return true
2. If any clause is false in model, return false
3. Propagate all pure symbols
4. Assign all unit clauses
5. p ← symbols.next()
6. return DPLL(clauses,symbols-p,EXTEND(model, p, true)) ||
          DPLL(clauses,symbols-p,EXTEND(model,p, false))
```

Figure 1: DPLL (adapted from Russell and Norvig 2003). Operates over sentences in conjunctive normal form.

The world-based version begins with an empty set of world basis elements, a list of uncertain atoms, and a set of constraints. If the current set of basis elements is not consistent with the constraints then the world is not valid. Every assumption creates a world where the assumption is assumed to be true, and one where it is assumed to be false. The algorithm will test each world for contradictions until no more uncertain atoms remain. A version of this algorithm is used for reasoning in the Polyscheme Cognitive Architecture (Cassimatis, et al 2010).

[9] This version is not guaranteed to return the "best" world, but simply returns whether the given constraints can be satisfied. Weighted DPLL will return the combination of atoms which produce the "best" world, but for clarity we have only demonstrated the classic non weighted version.

```
DPLL(world, uncertain_atoms, constraints) returns bool
1. If world has contradictions return false
2. If uncertain_atoms.isEmpty() return true
3. p ← uncertain_atoms.next()
4. return DPLL(world + p uncertain_atoms-p,constraints) ||
          DPLL(world+ -p,uncertain_atoms-p constraints)
```

Figure 2: DPLL as manipulation of worlds. (*world + p* denotes the the new world whose basis is the basis atoms of *world* plus the new atom *p*).

Reasoning with Quantifiers

A system with human-level intelligence must have knowledge that holds over large classes of objects. Variables and quantifiers in representational formalisms such first-order logic and production rules enable such knowledge to be expressed.

To illustrate, if an agent needs to make inferences such as "All dogs are mammals" or "Every dog wears a collar", it must be able to distinguish constants that refer to a single object from constants with universal or existential force. Quantified objects are not things that the agent can readily observe; "all dogs" includes all the dogs which have ever existed and ever will exist. Yet, quantified reasoning is foundational to human inference. Shapiro (2004) has argued for a common representation for quantified reasoning, both for uniformity in knowledge representation and ease in natural language translation.

Many methods which exhibit desired abilities of cognitive systems (e.g., neural networks or Bayesian networks) do not include variables or any other mechanisms for representing quantificational knowledge. A method of integrating quantified reasoning remains a significant outstanding problem in using these systems for high-level inference.

To address this problem, we show that it is in fact possible to represent quantificational facts in a formalism that does not have quantifiers, but does have worlds. To illustrate, suppose Mary is asked to imagine a dog that she has never seen before, and suppose she volunteers that the dog she is imagining is a mammal. If Mary's reasoning represents the state of the world, we are justified in saying that Mary believes all dogs are mammals, as there would be no other reason for her to make this assertion. This approach treats reasoning over quantifiers as reasoning about alternative states of the world containing novel or arbitrary objects. Because these objects have just been introduced, anything that is concluded about them must necessarily be true of all corresponding objects in reality.

"All dogs are mammals"

(39) $\forall x(\text{Dog}(x) \rightarrow \text{Mammal}(x))$

(40) $\text{Basis}(w12) = \{ \text{IsA}(d, \text{Dog}, E, R) \}$
$\text{IsA}(d, \text{Mammal}, E, w12)$

Since *d* represents an arbitrary object, we can conclude that anything true of *d* is also true of other dogs and use it as a constraint during inference.

A similar treatment can be used for existential quantification. Suppose that when Mary imagines a dog that she has never seen or heard of, that dog wears a collar. Based on the above reasoning, we can infer that Mary believes every dog wears a collar.

"All dogs have collars"
(41) $\forall x \exists y(Dog(x) \rightarrow HasA(x,y) \wedge Collar(y))$

(42) Basis(w13) = { IsA(d, Dog, E, R) }
HasA(d, c, E, w13)^ IsA(c, Collar, E, w13)

When *c*, a collar, appears for the first time in the consequents of an alternate world, it can be assumed to have existential force.

Uchida, Cassimatis, and Scally (in preparation) have shown that the same approach can be applied to other elements of logical formalisms including negation, conjunction, conditionals, and disjunction.

An Example of Worlds Enabling Integration

We can now demonstrate how treating worlds as fundamental objects supports strong integration of high-level reasoning. Consider the following chain of reasoning.

> "Mary probably checked the fridge for milk this morning, and if she did, she probably believes that we are out, not knowing that I already picked some up on my way home. That means she'll likely stop at the grocery on her way home tonight and be a little late."

This statement makes use of probabilistic inference, constraint satisfaction, reasoning about the beliefs of others, and planning about the future. When cast into the framework we have just laid out, the entire chain of reasoning becomes a set of worlds with basis elements and costs.

"Mary probably checked the fridge for milk this morning"
(43) Basis(w14) = { LookedInFridge(mary, R) }

"If she did, she probably believes that we are out, not knowing that I already picked some up on my way home"
(44) FridgeContainsMilk(self, R)
Basis(w15) = { LookedInFridge(mary, R) ^
Same(self, mary, w14}
-FridgeContainsMilk(self, w15)

"That means she'll likely stop at the grocery on her way home tonight and be a little late."
(45) Basis(w16) = { LookedInFridge(mary, R) ^
Same(self, mary, w14} ^
StopAtGrocery(mary, w15) }
StopAtGrocery(mary, R) → Late(mary, R)

At each step, costs are added to the worlds based on probabilities and overrides. As described in the preceding sections, novel inferences in these worlds can cause information to flow back into *R*. The least costly world will then be used by the agent for future planning.

Related Work

Several general purpose systems have added reasoning about worlds into their existing frameworks. Scone (Fahlman 2006) facilitates fast parallel retrieval from a knowledge base that can include multiple contexts for representing counterfactual situations and beliefs. Contradictions are avoided by building in "cancel-links" that block certain information from propagating between contexts. SNePS (Shapiro 2000) utilizes contexts for counterfactual worlds and belief representation (Shapiro and Rapaport 1987), using a belief revision system to remove incompatibilities. Situation calculus has also been extended to perform counterfactual reasoning (Costello and McCarthy 1999). While these systems all allow for the explicit formulation of given counterfactual world or belief state, the principles for creating the world themselves are underdetermined. In general, their worlds are custom built in the knowledge representation stage, through an explicit selection of which atoms from reality to carry over to the counterfactual world. A full account of human reasoning in hypothetical, counterfactual, and belief situations needs a set of principles that define how the appropriate alternate world is elaborated from the assumptions that create it.

Conclusions

We have shown how a representational framework based on worlds unifies the treatment of several difficult problems in human-level intelligence. In addition to the obvious benefits of parsimony, and the reduction of the number of different entities that a cognitive system must support, we believe that such a framework removes a significant obstacle to realizing human-level intelligence in cognitive systems. With a common framework, efforts to improve the performance and increase the capabilities of cognitive systems can focus on fewer mechanisms, with fewer obstacles involving representational tradeoffs, or integration overhead.

Acknowledgements

The authors would like to thank Paul Bello, Perrin Bignoli, Selmer Bringsjord, and members of the Human-Level Intelligence Lab at RPI for discussions on this work. This work was supported in part by grants from the Office of Naval Research (N000140910094), the Air Force Office of Scientific Research (FA9550-10-1-0389), and MURI award (N000140911029).

References

Barsalou, L. 1999. Perceptual Symbol Systems. *The Behavioral and Brain Sciences* 22(4):577-609.

Barsalou, L. 2009. Simulation, Situated Conceptualization, and Prediction. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* 364:1281-9.

Barwise, J., and Perry, J. 1983. *Situations and Attitudes*. Cambridge, Mass.: MIT Press.

Bello, P., Bignoli, P., and Cassimatis, N. 2007. Attention and Association Explain the Emergence of Reasoning about False Beliefs in Young Children. In *Proceedings of ICCM-2007*, 169-174. Oxford, UK: Taylor and Francis/Psychology Press.

Byrne, R. M. J. 2005. *The Rational Imagination*. Cambridge, Mass: The MIT Press.

Cassimatis, N. 2006. A Cognitive Substrate for Achieving Human-Level Intelligence. *AI Magazine* 27(2): 45–56.

Cassimatis, N., Bignoli, P., Bugajska, M., Dugas, S., Kurup, U., Murugesan, A., and Bello, P. 2010. An Architecture for Adaptive Algorithmic Hybrids. *IEEE Transactions on Systems, Man, and Cybernetics. Part B*, 4(3): 903-914.

Cassimatis, N., Murugesan, A., and Bignoli, P. 2009. Reasoning as Simulation. *Cognitive Processing* 10(4): 343-353.

Cassimatis, N., Trafton, J., Bugajska, M., and Schultz, A. 2004. Integrating Cognition, Perception and Action Through Mental Simulation in Robots. *Robotics and Autonomous Systems*, 49(1-2): 13-23.

Costello, T., and McCarthy, J. 1999. Useful Counterfactuals. *Electronic Transactions on Artificial Intelligence* 3(2): 1-28.

Davis M, and Putnam, H. 1960. A Computing Procedure for Quantification Theory. *Journal of the ACM* 7(1): 201–215.

Davis, M. Logemann, G. and Loveland, D. 1962. A Machine Program for Theorem Proving. *Communications of the ACM* 5(7): 394–397.

Devlin, K. 1991. *Logic and Information*. New York: Cambridge University Press.

Fahlman, S. 2006. Marker-Passing Inference in the Scone Knowledge-Base System. In *Proceedings of KSEM'06*. Springer-Verlag.

Fauconnier, G. 1985. *Mental Spaces*. Cambridge, Mass: MIT Press.

Fauconnier, G., and Turner, M. 2002. *The Way We Think*. New York: Basic Books.

Goldman, A. 2002. Simulation Theory and Mental Concepts. In Dokic, J. and Proust, J. eds. *Simulation and Knowledge of Action*, 1-20. Oxford: Blackwell.

Gordon, R. M. 1995. The Simulation Theory: Objections and Misconceptions. In Davis, M. and Stone, T. eds. *Folk Psychology*, 100-122. Oxford: Blackwell.

Gordon, R. M., and Cruz, J. 2002. Simulation Theory. In Nadel, L. ed. *Encyclopedia of Cognitive Science*, 9-14. Macmillan.

Heras, F., Larrosa, J., and Oliveras, A. 2008. MiniMaxSat: A New Weighted Max-SAT Solver. *Journal of Artificial Intelligence Research* 31:1-32.

Hoos, H., and Stutzle, T. 2002. SATLIB: An Online Resource for Research on SAT. In Gent, I., Maaren, H., and Walsh, T. eds. *SAT 2000*, 283-292. Amsterdam: IOS Press.

Johnson-Laird, P. 1983. *Mental Models*. Cambridge, Mass: Harvard University Press.

Johnson-Laird, P. 2006. *How We Reason*. New York: Oxford University Press.

Kautz, H., and Selman, B. 1999. Unifying SAT-based and Graph-based Planning. Paper Presented at the IJCAI-99.

Kripke, S. 1963. Semantical Considerations on Modal Logic. *Acta Philosophica Fennica* 16: 83-94.

Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. Chicago: The University of Chicago Press.

Lakoff, G., and Johnson, M. 1999. *Philosophy in the Flesh: the Embodied Mind and its Challenge to Western Thought*. New York: Basic Books.

Lewis, D. 1973. *Counterfactuals*. Cambridge, Mass: Harvard University Press.

Marques-Silva, J. P., and Sakallah, K. A. 1996. GRASP—A New Search Algorithm for Satisfiability. Paper presented at the international conference on computer-aided design.

McCarthy, J., and Hayes, P. 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence. In Meltzer, B. and Michie, D. eds. *Machine Intelligence 4.* 463-502. Edinburgh University Press.

McCarthy, J. 2002. Actions and Other Events in Situation Calculus. In *Proceedings of KR-2002*, 615–628.

Meltzoff, A. 2005. Imitation and Other Minds: The 'Like Me' Hypothesis. In Hurley, S. and Chater, N. eds. *Perspectives on Imitation: From Neuroscience to Social Science* 2: 55-77. Cambridge, Mass: The MIT Press.

Mueller, E. 2004. Event Calculus Reasoning Through Satisfiability. *Journal of Logic and Computation* 14(5): 703-730.

Russell, S. and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Sang, T., and Beame, P. 2005. Solving Bayesian Networks by Weighted Model Counting. In *Proceedings. of AAAI-05.*

Shanahan, M., and Witkowski, M. 2004. Event Calculus Planning Through Satisfiability. *Journal of Logic and Computation* 14(5): 731-745.

Shapiro, S. 2000. SNePS: A Logic for Natural Language Understanding and Commonsense Reasoning. In Iwanska, L. and Shapiro, S. eds., *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*, 175-195. Menlo Park, CA/Cambridge, Mass: AAAI Press/MIT Press.

Shapiro, S. 2004. A Logic of Arbitrary and Indefinite Objects. In *Proceedings of KR-2004*, 565-575.

Shapiro, S., and Rapaport, W. 1987. SNePS Considered as a Fully Intensional Propositional Semantic Network. In Cercone, N. and McCalla, G. eds. *The Knowledge Frontier: Essays in the Representation of Knowledg*, 262-315. New York: Springer-Verlag.

Singla, P., and Domingos, P. 2006. Memory-Efficient Inference in Relational Domains. In *Proceedings of AAAI-06*.

Sorensson, N., and Een, N. 2005. MiniSat v1.13 – A SAT Solver with Conflict-Clause Minimization. SAT 2005 Competition.

Stalnaker, R. 1968. A Theory of Conditionals. In N. Rescher (Ed.), *Studies in logical theory (American Philosophical Quarterly Monograph No. 2)*, 98–112.Oxford: Blackwell.

Uchida, H., Cassimatis, N., and Scally, J. (in preparation). Perceptual Simulations Can be as Expressive as First-order Logic.

Wimmer, H., and Perner, J. 1983. Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition* 13(1): 103–128.