Advances in Cognitive Systems 2 (2012) 43-58 Submitted 09/2012; published 12/2012

Pretense and Cognitive Architecture

Paul Bello PAUL.BELLO@NAVY.MIL
Human and Bioengineered Systems Code 341, Office of Naval Research, Arlington, VA 22203 USA

Abstract

Herein I sketch out the foundations of a computational theory of pretense, where pretense is broadly defined as pretend-play. In keeping with the sizable literature on the topic, I assume that pretending involves a number of distinct features that include the intentional decoupling of pretense from reality. A scenario is presented that involves all of what have been claimed to be necessary conditions for engaging in pretense and are mapped to various representational and process-level commitments within the *Polyscheme* computational cognitive architecture. A model of the scenario is developed, and a summary trace through Polyscheme's computation is provided to demonstrate the architecture at work during an episode of play. Finally, the relationship of pretense to other mental states is discussed, especially with respect to how different varieties of mental state reasoning might be realized by the same underlying representations and architectural mechanisms.

1. Introduction

Much of the literature on the nature of mental states in philosophy and artificial intelligence is typically focused on analyses of beliefs, desires and intentions. However, a few researchers in cognitive development have spent a considerable amount of effort on the study of *pretense* as a type of mental state (Leslie, 1987; Harris, 1995; Lillard, 1993). For the purpose of this discussion, I take pretense to be the mental state that is centrally concerned with the human capacity for pretend-play. Quoting from Garvey (1990, p. 82), pretense is ‘‘the voluntary transformation of the here and now, the you and me, and the this or that, along with any potential action that these components of a situation might have.’’

One of the shared features of the extant literature on the topic is its general lack of a computational story that squares well with the empirical data. This paper is an attempt to provide the beginnings of such a story, and to explore connections between pretense and other mental states. Since the paper is primarily computational, I rely heavily on extant summaries of the empirical literature on pretense and its distinguishing characteristics without spending time describing individual studies or their implications. With this in mind, the paper begins with a summary of the ideas in Lillard (2001), which describes the basic features of pretend-play and motivates the inclusion of each via empirical work conducted by a variety of researchers. It will be argued that the cognitive mechanisms that support engagement in pretense are substantially involved in *counterfactual* reasoning.

Turning to the computational, I review the basics of representation and inference in Polyscheme (Cassimatis et al., 2010) with a special emphasis on its ability to entertain mental simulations of situations that differ from reality as the architecture knows it to be. A quasi-formal

presentation of Polyscheme's representation language and inference procedure is given, culminating with a focused example of counterfactual reasoning. With the desiderata from Lillard (2001) in mind, I elaborate a scenario involving the construction of a mud-pie; mapping aspects of the task to each of the distinguishing features of pretense. A task model is developed in Polyscheme, and highlights from the model file and execution trace are presented to illustrate key inferences and results. I conclude with a general discussion that links core features of pretense to architectural features in Polyscheme, with some suggestions regarding the relationship between participating in pretense and reasoning about the (potentially false) beliefs of others.

2. Some Key Features of Pretense

What set of requirements does participating in pretense impose on a cognitive system? Much of what follows in this section is roughly drawn from Lillard's overview paper that lays out a fairly uncontroversial set of desiderata associated with pretend-play. I round out Lillard's requirements with insights from a paper by Nichols and Stich (2000).

The most obvious feature of pretense is that it appears to require the pretender to mentally entertain a non-actual state of affairs, such as pretending that the banana in front of him is a telephone. Furthermore, the pretender needs to manage the relationships between pretense and reality. While a pretender could arguably entertain a state of affairs that radically differs from reality, it seems to be the case that what we know about reality is often mostly taken for granted in the pretense. If the pretender is entertaining the banana-as-telephone, he takes for granted that the banana-as-telephone is still *on* the table, and not *beneath* the chair. Reality is certainly suspended for the purposes of pretending, but not in its entirety. In this sense, pretense is circumscribed to some degree, with much of what we know about reality being assumed to stay the same over the course of an episode of play.

Along with the former characteristics, pretense has intentional structure and its own set of mental representations. This means that when an agent engages in pretense, it is with some measure of purpose. In this way, pretense is its own special kind of activity, and not just purely behavioral in nature. Agents entertain alternate states of affairs populated with transmuted or non-existent objects and relations, all while being conscious of purposely doing so. If they did not intentionally engage in pretense, their representations of bananas-as-telephones would merely be mistaken. Children as young as two appreciate the difference between *trying* to perform an action in the real world, and *pretending* to perform the same action (Rakoczy & Tomasello, 2006). If they were incapable of doing so, it would be next to impossible to discriminate pretense from any other form of action. While the behavioral indicators of pretense are under debate, the aforementioned study (among others) suggests that the general capability for discrimination is present, even in young children. Following the latter set of points, children recognize that pretense is a capability possessed by agents, and have the capacity to not only participate in pretense themselves, but also recognize that others are pretending. Finally, pretend-play often has an action-oriented component. At tea parties, young children act out specific roles, and use pretended-about objects in a manner consistent with the pretense (Nichols & Stich, 2000). But what sorts of cognitive capabilities that we know about could support such a complex set of requirements?

3. Pretense and Counterfactual Reasoning

It should be clear from the previous section that engaging in pretense is not just a matter of exercising a single cognitive capability, such as the ability to categorize, or to recall an item from memory. Instead, pretense behavior is supported by a variety of mental representations and processes. In this section, I pick out counterfactual reasoning as being essential to the cognitive architecture supporting pretense; however this is not to say that pretense does not involve anything but counterfactual reasoning. In the last section, I discussed the teleological aspect of pretense, along with various correspondences between considered actions in the pretense and real-world actions to potentially be executed. These are not essential components of counterfactual reasoning *simpliciter*. An agent could entertain a counterfactual without anything like an accompanying conscious intention to do so, and the counterfactual need not have anything at all to do with real-world actions. Simply put, counterfactuals are statements of the form "if it were that φ, then it would be that ψ." One can see how this easily maps on to the case of pretend-play. The pretender assumes (for the sake of play) that some set of features about reality are suspended, altered or replaced whole-cloth within the context of the pretense. Usually, the consequences are worked out through the course of the pretense episode, often being accompanied by associated actions on corresponding real-world objects.

Perhaps the most well-known story about counterfactuals derives from the philosophical work of Lewis (1973). To summarize the essence of Lewis' account, and without diving into the details of possible worlds semantics, let us consider a *world* to be a complete description of how things could be. By complete, I mean that every proposition that describes some or other aspect of such a world has an assignment of either true or false. Now, let us consider one of these worlds to be privileged in the sense that it describes reality as we best know it to be. A counterfactual statement such as: "If it were the case that φ then it would be the case that ψ" is true of the privileged world if and only if either (1) there are no worlds where φ or (2) there is at least one world where φ holds and ψ holds that is *closer* to the privileged world than any other world where φ holds and ψ does not hold. Intuitively this delivers the closest version of reality to our own that is consistent with the antecedent of the counterfactual conditional. In the case of participation in play, counterfactual reasoning of the type I describe would result in the consideration of a pretense-world defined by one or more counterfactual assumptions (e.g., the banana is a telephone) and the revisions to real-world knowledge structures that follow from them. Lewis' account of counterfactuals remains a hotly debated issue, and is not without its problems, the most obvious of which consists in how to define "closeness" between worlds. Many of these theoretical lacunae may be side-effects of the assumptions built into possible worlds semantics and could ultimately be resolved by relaxing them to varying degrees. One way to do so might be to develop a plausible account of counterfactual reasoning in terms of resource-bounded cognitive computation. I focus on Lewis' interpretation in this section because it has been widely discussed, and because the idea of possible worlds, however cognitively implausible it is, has some kind of restricted analogue to the idea of a *mentally simulated world.* In the next sections I present the *Polyscheme* cognitive architecture, which is especially well-suited to modeling pretense and related phenomena since the idea of a *world* is built deeply into its theoretical commitments.

4. The Polyscheme Cognitive Architecture

My account of pretense is developed with the *Polyscheme* computational cognitive architecture (Cassimatis et al., 2010). While many cognitive architectures are specified at the level of constructions derived from cognitive psychology and have only been marginally concerned with high-level reasoning, Polyscheme's commitments are primarily to richer representation and inference mechanisms. In particular, Polyscheme was originally designed to simulate infants as they performed physical reasoning tasks from the literature on cognitive development (Cassimatis, 2002). The architecture consists of a number of reasoning modules called *specialists* that are coordinated by a focus of attention mechanism on every cognitive cycle (for details, see Cassimatis et al., 2010). Polyscheme has dedicated specialists for reasoning about time, space, identity, constraints, alternate worlds, categories, and partonomies, all of which were employed in the modeling of infant physical reasoning. These basic domains are part of our first guess at a *cognitive substrate* (Cassimatis, 2006) that supports the majority of higher-order cognition.

The Polyscheme research program relies crucially on the idea that higher-order cognitive phenomena can be reduced to reasoning about time, space, worlds, *et cetera*. There have been a number of demonstrations of successful reductions, including the development of a unification-style grammar allowing for the simultaneous processing of syntactic and semantic constraints in language understanding (Cassimatis, 2004; Muruguesan & Cassimatis, 2006). I have also been involved in the business of substrate reductions, with an account of early competence in infants to reason about the (false) beliefs of other agents (Bello et al., 2007; Bello, 2012). Given the relative success of the strategy, I have no reason to assume special-purpose representational or processing resources for pretense. If we are right about the basics of the substrate, much of what goes on with pretense ought to fall out of the operation of Polyscheme's existing commitments.

4.1 Knowledge Representation

An *atom* is a relation over one or more entities that is assigned a truth value at a specific time in a *world*. In general, atoms are of the form *RelName*(*e1*, *e2*, …, *ei*, *t*, *w*). The second-to-last argument represents a temporal interval. We use the letter "E" to designate the temporal interval representing "at all times." The last argument defines the world in which the relation holds. We use the letter "R" to represent the agent's beliefs about reality (rather than about imagined or counterfactual worlds). We might therefore represent "Paul is hungry at noon." as IsHungry(paul, noon, R). To represent the converse, we use standard negation: ¬IsHungry(paul, noon, R). Arguments of the form ?*x* as in IsHungry(?agent, ?t, ?w) are unbound variables. Relation names can also be prepended with ?, allowing for implicit quantification over relations. Constraints express contingencies between atoms. The standard logical operators ∧ and → are used to construct constraints. All constraints are implicitly universally quantified. For example, IsHungry(?agent, ?time1, ?w) ∧ LineOfSight(?agent, ?food, ?time1, ?w) → ReachFor(?agent, ?food, ?time2, ?w) expresses that if an agent is hungry at time1 and has line of sight on some food, then the agent will reach for the food at time2. Finally and importantly, we are able to represent *soft constraints* that generate costs on the worlds in which they are broken. To write "All professors are usually nutty," we say Professor(?x, E, ?w) $\rightarrow_{(0.75)}$ Nutty(?x, E, ?w). What this constraint essentially means is that for any professor in any worlds at any time, they are very likely to be nutty. If we find a world in which there is a professor who turns out to not be nutty,

that world incurs a cost of 0.75. Constraints written using the → conditional without an associated cost incur an infinite penalty if broken, and are called *hard constraints*.

4.2 Evidence Combination During Inference

Rather than atoms being restricted to either true or false, as is typically the case in symbolic reasoning approaches, Polyscheme employs the slightly more complicated notion of an *evidence tuple*. Every atom has an associated evidence tuple of the form <*E*+, *E*->, where *E*+ signifies the positive evidence for the atom, and *E*- represents the negative evidence against the atom. Evidence for or against an atom takes one of the following values: (C)ertain, (L)ikely, (l)ikely, (m)aybe, or (n)eutral. For example, the atom Red(apple1, E, R) <C, n> means that it's certainly true that apple1 is red at all times in R, with no negative evidence associated. Conversely, we could say Red(apple1, E, R) <n, C>, which would mean that it's certainly false that apple1 is red at all times in R, with no positive evidence to weigh against. Similarly, we could say it is very likely that apple1 is red at all times in R by assigning it <L, n>, and so on.

When the antecedent of a soft constraint is matched by an atom having a certain truth-value, the derived atom on the right-hand-side acquires a *diluted* truth-value. For example, if we have the constraint: Red(?x, E, ?w) $\rightarrow_{(0.75)}$ Ripe(?x, E, ?w), and we know that Red(apple1, E, R) <C, n>, then the resulting atom will be Ripe(apple1, E, R) <L, n>. For the sake of simplicity, we do not consider truth-values other than C, n, and L throughout the rest of the paper. Evidence is combined qualitatively in a commonsense fashion, such that when Polyscheme is originally uncertain about an atom and acquires information via perception or inference that assigns certainty to that atom, the uncertain truth-value is replaced by the new certain judgment. Space precludes a detailed exposition of evidence combination rules given all possible states of an evidence tuple, but for the purpose of this paper, we only need to be concerned with replacing L's in the evidence tuple with C's, especially in the context of counterfactual reasoning, described later in this section.

4.3 Dynamically Created Objects and Relations

Polyscheme has the ability to generate and reason about new objects and relations in mid-inference, offering it an advantage over many popular inference strategies that involve propositionalizing a set of constraints by fully instantiating them using a pre-defined propositional domain. Polyscheme uses *lifted inference*, instantiating constraints incrementally as it computes (Cassimatis et al., 2009). While a commitment to lifted inference does not provide the same guarantees on decidability that come along with propositionalized approaches, there are relatively few cases where lifted inference becomes problematic, and those can be treated as special cases. To illustrate the capability, let us assume Polyscheme knows about the following constraint, stating that whenever an email comes in, a notification is generated on the desktop:

NewMail(?x, E, ?w) → Notification(?y, E , ?w)

Now, it is sometimes the case that we do not know that we have gotten an email until the notification appears. For the sake of the example, let us assume that Polyscheme gets a notification via perception: Notification(not1, E, R). Since ?x and ?y are assumed to be different

variables, Polyscheme will generate a new object (e.g., mail1) on the left-hand-side, resulting in the following two atoms: Exists(mail1, E, R), and NewMail(mail1, E, R). This is an essential pattern of inference in causal explanation, where one typically observes an outcome, and reasons backward to potential causes. Positing new objects can also happen on the right-hand-side as well. Cassimatis et al. (2009) provide a detailed discussion of object generation during the course of inference, including discussion of the limitations of reasoning over potentially infinite domains.

4.4 Utilizing Worlds

Worlds in Polyscheme are defined uniquely by their *basis*, or the set of assumptions (e.g. atoms) on which they are based. The real world "R" is defined as having an empty basis:

Basis(R) = {}

Every subsequent example will consist in two parts: the definition of worlds by their basis, and a set of *constraints* that will often range over entities in the different worlds under consideration. Let us begin by building up a simple example. Suppose we are uncertain about the location of the red apple in the real world. We start with a simple constraint that states that if a fruit is in a certain location, then the location has a fruity smell. It should be noted that this constraint is implicitly universally quantified over all worlds, all times, for all locations ?loc, and for all fruits ?y. Recall that R represents the real world as Polyscheme knows it to be. R always has an empty basis. To begin, we define a number of atoms that are certainly true in R, including knowledge that the sky is always blue, apples are red fruits, and that there isn't a fruity smell at loc1. Our worlds-based framework allows us to consider *hypothetical* worlds that are related to the real world in the following way:

Constraints:

C_1: IsA(?y, Fruit, E, ?w) $\wedge$ Location(?y, ?loc, E, ?w) $\rightarrow$ FruitySmell(?loc, E, ?w)

Basis(R) = {}

- Blue(sky, E, R)
- IsA(apple, Fruit, E, R)
- Red(apple, E, R)
- $\neg$FruitySmell(loc1, E, R)

Basis(w1) = {Location(apple, loc1, E, R)}
Basis(w2) = {$\neg$Location(apple, loc1, E, R)}

Notice that the atoms in the bases of hypothetical worlds w1 and w2 are marked with R for their world argument. This is to denote that the assumptions made in these worlds are assumptions *about* R. Since these are assumptions on which the hypothetical worlds rest, the Location(apple, loc1, E, R) atom in the basis of w1 is true, and it is false in w2 due to the negation operator prepending it.

4.5 Inheritance Between Worlds

When we construct a hypothetical world, we would like as many atoms as possible about the real world to stay the same in the hypothetical world. If I assume that the apple is at loc1 in w1, I would also like to have the fact that the apple is red available to me in w1. Given the definitions above, I know the apple is red in R. What we need is a way to connect R to w1 in such a way that atoms from R become available for use in w1. This process is called *inheritance* and will be crucial to understanding the mechanisms driving pretense. Two worlds are relevant to one another when the basis of one world is fully contained in the basis of the other. Additionally, it is noted that the relevance relation is transitive:

$\boldsymbol{Rel_1}$: Basis(?w1) $\subseteq$ Basis(?w2) $\rightarrow$ RelevantTo(?w1, ?w2)
$\boldsymbol{Rel_2}$: RelevantTo(?w1, ?w2) $\wedge$ RelevantTo(?w2, ?w3) $\rightarrow$ RelevantTo(?w1, ?w3)

The basic form of a constraint that enables inheritance (3) can now be written:

$\boldsymbol{I_1}$: ?Relation(?e1, ..., ?ei, ?t, ?w1) $\wedge$ RelevantTo(?w1, ?w2) $\rightarrow$?Relation(?e1, ..., ?ei, ?t, ?w2)

We can now reason about the example set of worlds given above. With the inheritance rule in place, and both w1 and w2 being relevant to R by way of $\boldsymbol{Rel_1}$, we have the configuration shown in Table 1. Here we consider two hypothetical worlds as they relate to the real world R. In w1, the first of these worlds, it is assumed that the apple is located in loc1. By inheritance rule $\boldsymbol{I_1}$, all atoms true in R become true in w1. IsA(apple, Fruit, E, R) combines with Location(apple, loc1, E, R) in the basis of w1 to produce FruitySmell(loc1, E, w1) by an application of constraint $\boldsymbol{C_1}$. Since $\neg$FruitySmell(loc1, E, R) inherits into w1, the two produce a contradiction, and w1 is *clobbered*, and removed from further consideration. Since the basis of w2 is consistent with every atom inherited into w2 from R, no contradiction is derived, and w2 survives.

4.6 Counterfactual Reasoning Using Worlds

As previously discussed, pretense relies centrally on the notion of entertaining counterfactuals. In the framework we've been developing, counterfactual worlds are no different than the hypothetical worlds discussed in the previous examples. Counterfactuals present an immediate difficulty for the inheritance rule $\boldsymbol{I_1}$. I will illustrate with an example, but since identity will play a critical role in the forthcoming account of pretense, let us define identity as follows:

$\boldsymbol{Id_1}$: Same(?x,?y, E, ?w) $\wedge$?Relation(?x, ..., ?xi, ?t, ?w1) $\wedge$ $\neg$?Relation(?y, ..., ?yi, ?t, ?w1)
$\rightarrow$ False(E, ?w)
$\boldsymbol{Id_2}$: Same(?x,?y, E, ?w) $\wedge$?Relation(?x, ..., ?xi, ?t, ?w1) $\rightarrow$?Relation(?y, ..., ?yi, ?t, ?w1)

$\boldsymbol{Id_1}$ written above states that it is not possible for Same(?x, ?y, E, ?w) to be true in any world where ?x and ?y have different relational properties. $\boldsymbol{Id_2}$ states that if two entities are the same, and the first is in the extension of some relation, then the second must also be in the extension of that relation. To show how inheritance rule $\boldsymbol{I_1}$ doesn't work here, we consider the real world where we know mud is not edible and pie filling is. We also consider the atom that expresses the trivially true fact that mud is the same as mud. We then wish to reason counterfactually by simulating a world w1 where mud is the same as pie filling, as shown in Table 2.

Table 1. Resolution of uncertainty using inference over worlds.

Parent World R: Basis(R) = {}	
Constraints:	
C_1: IsA(?y, Fruit, E, ?w) ∧ Location(?y, ?loc, E, ?w) → FruitySmell(?loc, E, ?w)	
Blue(sky, E, R) <C, n>	
IsA(apple, Fruit, E, R) <C, n>	
Red(apple, E, R) <C, n>	
¬FruitySmell(loc1, E, R) <C, n>	
Child World w1: Basis(w1) = {Location(apple, loc1, E, R)}	**Child World w2:** Basis(w2) = {¬Location(apple, loc1, E, R)}
Blue(sky, E, w1) (<C, n> by I_1)	Blue(sky, E, w2) (<C, n> by I_1)
IsA(apple, Fruit, E, w1) (<C, n> by I_1)	IsA(apple, Fruit, E, w2) (<C, n> by I_1)
Red(apple, E, w1) (<C, n> by I_1)	Red(apple, E, w2) (<C, n> by I_1)
¬FruitySmell(loc1, E, w1) (<C, n> by I_1)	¬FruitySmell(loc1, E, w2) (<C, n> by I_1)
FruitySmell(loc1, E, w1) (<C, n> by C_1)	
⊥ (contradiction)	

By its nature, any instance of counterfactual reasoning will result in a contradiction between atoms inherited from base worlds into their counterfactual children. In order to allow for such reasoning to take place, we need to relax the hard constraint encoded by inheritance rule I_1 by taking advantage of Polyscheme's evidence tuples. We write the inheritance constraint for counterfactual worlds as:

I_2: ?Relation(?e1, …, ?ei, ?t, ?w1) ∧ RelevantTo(?w1, ?w2) ∧
IsCounterfactualWorld(?w2, E, ?w1) $\rightarrow_{(.9)}$?Relation(?e1, …, ?ei, ?t, ?w2)

where the numeric tag on the conditional takes a value in the range (0,1), and ?w2 is declared to be counterfactual to ?w1 by means of the relation: IsCounterfactualWorld(?w2, E, ?w1).

Table 2. Counterfactual reasoning cannot be supported by simple inheritance.

Basis(R) = {}	Basis(w1) = {Same(mud, pieFilling, E, R)}
Same(mud, mud, E, R) <C, n>	Same(mud, mud, E, w1) (<C, n> by I_1)
¬Edible(mud, E, R) <C, n>	¬Edible(mud, E, w1) (<C, n> by I_1)
Edible(pieFilling, E, R) <C, n>	Edible(pieFilling, E, w1) (<C, n> by I_1)
	⊥ (contradiction)

Any atom that is the consequent in a soft constraint of this form will not have a certain truth value, but will instead be considered uncertain. We mark these atoms as either <L, n> or <n, L>, meaning very likely true or very likely false, respectively. However, Polyscheme does everything it can to resolve every uncertain atom in every world to being certainly true or certainly false. When encountering an uncertain atom in a world w, Polyscheme simulates two worlds relevant to w: one where the atom is certainly true and one where the atom is certainly false. Inference proceeds in these worlds in an attempt to rule one of the two worlds out by way of contradiction. The process as described looks just like the example of hypothetical reasoning we discussed in Section 2.2. To illustrate these processes working together, let's return to the previous example of drawing an identity between mud and pie filling during an episode of counterfactual reasoning. After the inheritance process is complete and evidence is combined in the manner discussed in Section 3.2, we are left with the lower right-hand corner of Table 3. In essence, w1 becomes a version of the real world under the assumption that mud is actually pie filling. Relations that apply to pie filling in the real world apply to mud in the counterfactual world, and relations that previously applied to mud in the real world no longer apply in w1. The resulting world w1 is the closest world to R in which the counterfactual assumption holds, consistent with the popular account of counterfactuals as possible worlds given by Lewis (1973).

Table 3. Inheritance as a soft constraint to support counterfactual reasoning in w1.

<table>
<tr><td colspan="2">Parent World R:
Basis(R) = {}

Same(mud, mud, E, R) <C, n>
¬Edible(mud, E, R) <C, n>
Edible(pieFilling, E, R) <C, n>
IsCounterfactualWorld(w1, E, R) <C, n></td></tr>
<tr><td>Child World w1 (pre-revision):
Basis(w1) = {Same(mud, pieFilling, E, R)}

Same(mud, mud, E, w1) (<L, n> by I₂)
¬Edible(mud, E, w1) (<L, n> by I₂)
Edible(pieFilling, E, w1) (<L, n> by I₂)
Edible(mud, E, w1) (<C, n> by Basis and Id₂)</td><td>Child World w1 (post-revision):
Basis(w1) = {Same(mud, pieFilling, E, R)}

Same(mud, mud, E, w1) (<n, C>)
¬Edible(mud, E, w1) (<n, C>)
Edible(pieFilling, E, w1) (<C, n>)
Edible(mud, E, w1) (<C, n>)</td></tr>
</table>

4.7 Upward and Downward Inheritance

Recall that one of the more commonplace features of pretense involved an action-oriented component. On the earlier analysis of pretense, we differentiated imagination from pretense by virtue of the fact that imagining oneself doing something is fundamentally different than acting in the real world consistent with the contents of such an episode of imagining. Pretense seems to

Table 4. Domain-specific constraints for making a mudpie.

$\mathbf{C_0}$	Goal(makeMudPie, E, ?w) ∧ IsA(mudPie, Pie, E, ?w) ∧ IsA(makeMudPie, Pretense, E, ?w) ∧ IsA(?m, Mud, E, ?w) ∧ IsA(?f, Filling, E, ?w) → IsCounterfactualWorld(?pw, E, ?w) ∧ RelevantTo(?w, ?pw) ∧ Same(?m, ?f, E, ?pw) ∧ Holds(pieMade, t_made, ?pw) ∧ Goal(makePie, E, ?w)
$\mathbf{C_1}$	Goal(makePie, E, ?w) → Do(?putInOven, ?t_put, ?w) ∧ Do(?mouthOpen, ?t_mopen, ?w) ∧ Do(?putInMouth, ?t_inMouth, ?w) ∧ Do(?chewSwallow, ?t_cs, ?w)
$\mathbf{C_2}$	Do(?putInOven, ?t_put, ?w) → Holds(pieInOven, ?t_baked, ?w) ∧ Meets(?t_put, ?t_baked, E, ?w) ∧ IsA(?p, Pie, E, ?w)
$\mathbf{C_3}$	Do(?mouthOpen, ?t_mopen, ?w) → Holds(mouthOpen, ?t_open, ?w) ∧ Meets(?t_mopen, ?t_open, E, ?w)
$\mathbf{C_4}$	Do(?putInMouth, ?t_inMouth, ?w) ∧ Edible(?p, E, ?w) → Holds(itemInMouth, ?t_in, ?w) ∧ Meets(?t_inMouth, ?t_in, E, ?w)
$\mathbf{C_5}$	Do(?chewSwallow, ?t_cs, ?w) → Holds(chewedAndSwallowed, ?t_chewedSwallowed, ?w) ∧ Meets(?t_cs, ?t_chewedSwallowed, E, ?w)
$\mathbf{C_6}$	IsA(?p, Pie, E, ?w) ∧ Holds(pieInOven, ?t_baked, ?w) → SmellsGood(?p, ?t_baked, ?w) ∧ Edible(?p, E, ?w)
$\mathbf{C_7}$	IsA(?p, Pie, E, ?w) ∧ IsA(?f, Filling, E, ?w) → Edible(?f, E, ?w)
$\mathbf{C_8}$	¬Edible(?f, E, ?w) ∧ Holds(itemInMouth, ?t_in, ?w) ∧ Meets(?t, ?t_in, E, ?w) → ¬Do(?action, ?t, ?w)

involve the latter. What does this mean for the account I have given so far? It means that actions generated within the counterfactual worlds supporting the pretense need to be sent down to the real world to be executed. Atoms in the counterfactual world need to be made available to Polyscheme's action-selection routines in R. So far, I have explored cases in which atoms in a parent world are available to their child-worlds. Call this upward inheritance. Downward inheritance is what you might expect it to be: items from a child world inheriting downward into the parent world. Setting aside the fact that downward inheritance seems to be counterintuitive, the rules for inheritance are expressed in a similar fashion:

$\boldsymbol{I_3}$: ?Relation(?e1, …, ?ei, ?t, ?w2) ∧ RelevantTo(?w1, ?w2) ∧
IsCounterfactualWorld(?w2, E, ?w1) $\rightarrow_{(.9)}$?Relation(?e1, …, ?ei, ?t, ?w1)

Counterfactual conclusions generated in a child world can migrate downward to the parent world as uncertain, with a search process resolving the uncertainty, where possible. Most all of the time, there will be atoms in the parent world that will suppress the counterfactual consequences inherited downward into the parent world, but on some occasions, this might not be the case.

Table 5. Initial conditions for the mudpie scenario.

Parent World R: Basis(R) = {}	
Same(mud, mud, E, R)	<C, n>
Edible(mud, E, R)	<C, n>
Edible(pieFilling, E, R)	<C, n>
IsA(pieFilling, Filling, E, R)	<C, n>
Goal(makeMudPie, E, R)	<C, n>
IsA(mudPie, Pie, E, R)	<C, n>
IsA(makeMudPie, Pretense, E, R)	<C, n>
IsA(mud, Mud, E, R)	<C, n>

5. Polyscheme Makes a Mudpie

Putting the machinery developed in the last section to work, I continue with the example of making a mudpie in Polyscheme. While space precludes a full presentation of the fully working computational model, the computations mapping to the core features of pretense episodes will be elaborated to demonstrate how it works. Let us begin by sketching out, in Table 4, the basic action-related knowledge involved in making pie, and eating. All of what follows is a model sketch and not executable in the form reported in the table.

Table 5 gives the initial conditions in the real world prior to initiating the pretense. In this case, there are a few simple assertions about the categories and properties of objects involved in pie-making. Upon determining makeMudPie as a kind of pretense activity, Polyscheme simulates a counterfactual world w1 in which the pretense unfolds. Applying the domain-specific constraints described in Table 1 along with the inheritance rule $\boldsymbol{I_2}$ leaves us with an elaborated and revised version of w1 described in Table 6.

As inference proceeds in the counterfactual world w1, the *downward inheritance* rule $\boldsymbol{I_3}$ is active, sending inferential results back to the parent world R, albeit with likely truth-values in the evidence tuples, as shown in Table 7. Post-revision, we are left with three actions executed in the real-world:

- Do(putInOven001, t001, R) (<C, n> by $\boldsymbol{I_3}$, post-revision)
- Do(mouthOpen001, t002, R) (<C, n> by $\boldsymbol{I_3}$, post-revision)
- Do(chewSwallow001, t004, R) (<C, n> by $\boldsymbol{I_3}$, post-revision)

Polyscheme executed every action associated with the pretense in the real-world except for the one action that was forbidden. Constraint C_8 forbade the execution of eating inedible items. C_8 was effectively suspended in the counterfactual world w1 because one of the entailments of the counterfactual assumptions was that mud was indeed edible (as pie filling). One inherited back into R, where the mud/filling identity fails to hold, Polyscheme's real-world knowledge of mud being inedible blocks the execution of the putInMouth action.

Table 6. Pretense world w1 after inheritance and elaboration.

Child World w1: Basis(w1) = {Same(mud, pieFilling, E, R), Holds(pieMade, E, R) , Goal(makePie, E, R)}	
Same(mud, mud, E, w1)	(<n, C> by $\boldsymbol{I_2}$, post-revision)
¬Edible(mud, E, w1)	(<n, C> by $\boldsymbol{I_2}$, post-revision)
IsA(pieFilling, Filling, E, w1)	(<C, n> by $\boldsymbol{I_2}$, post-revision)
IsA(mud, Filling, E, w1)	(<C, n> by Basis and $\boldsymbol{Id_2}$)
Do(putInOven001, t001, w1)	(<C, n> by $\mathbf{C_1}$)
Do(mouthOpen001, t002, w1)	(<C, n> by $\mathbf{C_1}$)
Do(putInMouth001, t003, w1)	(<C, n> by $\mathbf{C_1}$)
Do(chewSwallow001, t004, w1)	(<C, n> by $\mathbf{C_1}$)
IsA(pie001, Pie, E, w1)	(<C, n> by $\mathbf{C_2}$)
Holds(pieInOven, t_baked, w1)	(<C, n> by $\mathbf{C_2}$)
Holds(mouthOpen, t_open, w1)	(<C, n> by $\mathbf{C_3}$)
Holds(itemInMouth, t_open, w1)	(<C, n> by $\mathbf{C_4}$)
Holds(chewedAndSwallowed, t_chewedSwallowed, w1)	(<C, n> by $\mathbf{C_5}$)
SmellsGood(pie001, t_baked, w1)	(<C, n> by $\mathbf{C_6}$)
Edible(pie001, E, w1)	(<C, n> by $\mathbf{C_6}$)
Edible(mud, E, w1)	(<C, n> by $\mathbf{C_7}$)

6. General Discussion

I have shown how a paradigmatic case of pretense can be accounted for in a cognitive architecture equipped with a capacity for sophisticated counterfactual reasoning. While this investigation was conducted via Polyscheme, I've endeavored to provide an exposition of the representations and processes supporting pretense that are largely architecture-independent. The model produced in the previous section was selected to illustrate the six core features of pretense:

1. An agent who is doing the pretending;
2. A reality that is being pretended *about*;
3. Explicit mental representation(s) that guide the pretense;
4. Projection of the pretense onto reality;
5. Intentional initiation and maintenance of the pretense; and
6. External manifestation of the pretense via action.

Each of the six features mentioned in the list have analogues in the account of given in this paper. The first feature states that agents are required for pretense. While seemingly a trivial point, it is worth remembering that young children not only participate in pretense from a fairly early age, but are also capable of *recognizing* pretense in other agents, as opposed to non-animates such as rocks, tables, and the like. The agency condition is not explicitly addressed in my example, but is easily accommodated in a more complete model where actions have agent arguments.

Table 7. Inheritance from counterfactual world w1 into parent world R.

Same(mud, mud, E, R)	<C, n>
¬Edible(mud, E, R)	<C, n>
Edible(pieFilling, E, R)	<C, n>
IsA(pieFilling, Filling, E, R)	<C, n>
IsCounterfactualWorld(w1, E, R)	<C, n>
Do(putInOven001, t001, R)	(<L, n> by $\boldsymbol{I_3}$)
Do(mouthOpen001, t002, R)	(<L, n> by $\boldsymbol{I_3}$)
Do(putInMouth001, t003, R)	(<L, n> by $\boldsymbol{I_3}$)
Do(chewSwallow001, t004, R)	(<L, n> by $\boldsymbol{I_3}$)
¬Do(putInMouth001, t003, R)	(<C, n> by $\mathbf{C_8}$)

The second feature demands that pretense involve a reality which serves as part of the content of pretense. Polyscheme pretends *about* certain aspects of reality by virtue of the atoms in the basis of the counterfactual world in which the pretense is elaborated. At least one of the basis atoms in the pretense-world reference an atom in the parent world within which the pretense was initiated. The upshot of generalizing this relationship is that Polyscheme is not only capable of pretending about reality as it knows it to be, but is also capable of thinking about an agent (and/or itself) pretending in the context of a hypothetical, counterfactual, past, or future world.

The third and fourth features of pretense involve explicit mental representations that guide the pretense and relate it to reality in specific ways. These conditions enforces a requirement that pretense involves *bona fide* mental representation of the objects and relations involved in the pretense. If part of the pretense involves a particular pile of mud (in reality) being represented as pie filling (in the pretense), the third feature on our list requires having an actual object in the pretense corresponding to an instance of pie filling. One could simply act as if the pile of mud was pie filling without ever entertaining the mud *as* pie filling (say for the purpose of demonstrating how to make a pie). Such demonstrations are not constitutive of pretense, since the goal of demonstration usually is teaching, and not play. This is taken care of in the model by having identity statements (e.g., Same(mud, pieFilling, E, R)) in the basis of the pretense world. Variable-free statements of this type presuppose two distinct objects, one labeled mud and the other labeled pieFilling.

Downward inheritance allows for the sixth feature on our list to be addressed. Actions generated as part of the pretense must be sent down to the real world for concomitant execution. Downward inheritance makes actions generated as part of a pretense available for real-world execution. As we saw in the case of mudpie-eating, having hard constraints against performing certain kinds of actions blocks the execution of certain (normally) undesirable aspects of the pretense. In the case of mudpie-eating, Polyscheme had a hard constraint against eating mud, so even when mud-eating was sent to R from the pretense world by way of downward inheritance, it was only sent as <L, n>, and immediately overridden by the hard constraint against eating mud. This resulted in Polyscheme executing *almost* every step of the pie-making-and-eating script, with only the consumption of mud being left out. Finally, the fifth feature of pretense involves the intentional component of pretense: there is an effort on the part of the pretender to engage in

pretense for the sake of engaging in pretense. Participation in the pretense must be an ostensible goal for the agent doing the pretending. The goal-directed nature of pretense was captured by having an active goal in the real world of participating in the pretense, which initiated the construction and population of the supporting counterfactual world. Although not explored here, having pretense as an active goal in the real world also serves to structure and terminate actions in the real world that are conditioned on happenings in the counterfactual world supporting the pretense.

Along with these six features, the mudpie example contains the three fundamental forms of pretense defined by Leslie (1987): object substitution, attribution of pretend properties, and entertaining imaginary objects. Polyscheme substitutes mud for pie-filling, attributes a pretend property (e.g., smelling nice) to the "cooked" mud, and fills out the scenario by inferring a non-existent pie, covering all three types in one example.

One of the most striking set of correspondences in the developmental literature is between performance on tasks involving counterfactual reasoning, engaging in pretense, and success on tasks requiring an agent to reason about the false beliefs of another agent (Riggs et al., 1998; Harris, 2005; Buchsbaum et al., 2012). On the account of pretense developed here in accordance with Polyscheme's *Cognitive Substrate Hypothesis*, this ought to be somewhat unsurprising. While having distinguishing features, reasoning about false beliefs, participating in pretense, and thinking about counterfactual alternatives can all be largely characterized by the consideration of worlds that diverge from one's current best model of reality.

I suspect that reasoning about the mental states of others almost always has something like a counterfactual character, for a variety of reasons. If one subscribes to the idea that mental state attribution is largely underwritten by a process of *mental simulation* in which the reasoner steps into the mental shoes of the target agent about whom he reasons, then the reasoner seemingly needs to entertain a counterfactual of the form: "If I were him, then…" Conversely, if one subscribes to the idea that mental state reasoning is a process of learning and refining a body of generalizations (i.e. a theory) about mental states and how they connect to potential behavior, then counterfactual reasoning becomes important in theory revision. I happen to be more sympathetic to the former than the latter, and find the connection between cognitive structures supporting counterfactual reasoning and mental state attribution to be extremely tight.

A particularly illustrative set of experimental results due to Kovacs, Teglas, and Endress (2010) suggests that even the unintentional consideration of the false belief of another agent seemingly intrudes into the mindreader's own perspective on the world. They illustrate the phenomena by way of a task where subjects are asked to press a button as soon as they know whether or not a certain ball is located behind an occluder. Incidentally, the stimuli are presented as having another agent (a smurf) in the scene, but the subjects are not asked to reason about the smurf's perspective at any point during the task – they are only to push the button. The experiment proceeds in various conditions with the ball exiting the scene, rolling behind occluders, and with the smurf being present or absent during the ball's movements. At least one sequence is produced where the smurf has the false belief that the ball is behind the occluder, while the subject is ignorant of the ball's location (e.g., it has exited the scene). When the button is pushed by the subject in this condition, the reaction-time profile seems to reflect the subject having spontaneously adopted the false belief of the smurf. It seems as if the content of the mental simulation of the subject-as-smurf becomes available to the subject in the real-world and subsequently modulates action-selection.

Similar findings have begun to appear in studies of simple perspective-taking tasks (Samson et al., 2010). In both cases, it seems that when perspectives of others are spontaneously (and often unconsciously) simulated by a human subject, information can sometimes "leak" backward into the perspective of that subject. These observations are consistent with the *downward inheritance* process that underwrites the coordination of actions generated in pretense worlds with the corresponding actions taken in the real world. Relationships between mechanisms in pretense and perspectival errors in belief ascription seem to be too coincidental to write off. The correspondence suggests to me that much of the process-level computation supporting pretense is also deeply-involved with belief ascription. I suspect similar phenomena will crop up when considering divergent desires, hopes, and other mental states as well.

7. Concluding Remarks

Much work remains to be done in order to develop these initial explorations into robust models. Most of the work to be done involves delimiting the function of the inheritance mechanism to support scalable episodes of pretense replete with many more agents, objects, relations and timepoints. Other aspects of pretense are ripe for investigation: role-taking in multi-agent joint play, embedding other kinds of mental-state inferences into play, and examining the role of affect in pretense.

To conclude, the study of pretense from a computational perspective is an important and worthwhile pursuit. Insofar as pretense is underwritten by more general capacities for counterfactual thinking, its study may lead us to a more detailed process-level understanding of planning, theory-revision, learning and other cognitive activities thought to involve entertaining non-actual worlds and their entailments. For cognitive systems researchers, it might be the architectural addition that opens up the door to building models of unprecedented complexity.

Acknowledgements

Special thanks are due to Nick Cassimatis for early discussion of the ideas in this paper.

References

Bello, P. (2012). Cognitive foundations for a computational theory of mindreading. *Advances in Cognitive Systems, 1*, 59–72.

Bello, P. (2011). Shared representations of belief and their effects on action selection: A preliminary computational cognitive model. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 2997–3002). Boston, MA.

Bello, P., Bignoli, P., & Cassimatis, N. (2007). Attention and association explain the emergence of reasoning about false beliefs in young children. *Proceedings of the Eighth International Conference on Cognitive Modeling* (pp. 169–174). Ann Arbor, MI.

Buchsbaum, D., Bridgers S., Weisberg D. S., & Gopnik A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences 367*, 2202–2212.

Cassimatis, N. L., Bignoli, P. G., Bugajska, M. D., Dugas, S., Kurup, U., Murugesan, A., & Bello, P. (2010). An architecture for adaptive algorithmic hybrids. *IEEE Transactions on Systems, Man, and Cybernetics: Part B*, *40*, 903–914.

Cassimatis, N. L., Murugesan, A., & Bignoli, P. G. (2009). Inference with relational theories over infinite domains. *Proceedings of the 2009 Meeting of the Florida Artificial Intelligence Research Society* (pp. 21–26). Sanibel Island, FL.

Cassimatis, N. L. (2006). A cognitive substrate for achieving human-level intelligence. *AI Magazine*, *27*(2), 45–56.

Cassimatis, N. L. (2004). Grammatical processing using the mechanisms of physical inferences. *Proceedings of the Twentieth-Sixth Annual Conference of the Cognitive Science Society* (pp. 192–197). Chicago, IL.

Cassimatis, N. L. (2002). *Polyscheme: A cognitive architecture for integrating multiple representation and inference schemes*. Doctoral Dissertation, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Garvey, C. (1990). *Play*. (2nd ed.). Cambridge, MA: Harvard University Press.

Harris, P. (1995). Imagining and pretending. In M. Davies & T. Stone (Eds.), *Mental simulation*. Cambridge: Blackwell.

Harris, P. (2005). *The work of the imagination*. Oxford, UK: Wiley-Blackwell.

Kovacs, A., Teglas, E., & Endress, A. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*, 1830–1834.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind". *Psychological Review*, *94*, 412–426.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard Univ. Press.

Lillard, A. (2001). Pretend play as twin earth: A social-cognitive analysis. *Developmental Review*, *21*, 495–531.

Lillard, A. (1993). Young children's conceptualization of pretense: Action or mental representational state? *Child Development*, *64*, 372–386.

Murugesan, A., & Cassimatis, N. L. (2006). A model of syntactic parsing based on domain-general cognitive mechanisms. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 1850–1855). Vancouver, BC, Canada.

Nichols, S., & Stich, S. (2000). A cognitive theory of pretense. *Cognition*, *74*, 115–147.

Rakoczy, H., & Tomasello, M. (2006). Two-year-olds grasp the intentional structure of pretense acts. *Developmental Science*, *9*, 557–564.

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, *13*, 73–90.

Samson, D., Apperly, I. A., Braithwaite, J., & Andrews, B. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 1255–1266.